

Andreas Ziegler and Inke R. König
**A Statistical Approach to
Genetic Epidemiology**

Related Titles

Elston, R. C., Johnson, W.

Basic Biostatistics for Geneticists and Epidemiologists **A Practical Approach**

2009

ISBN: 978-0-470-02489-8

Balding, D. J., Bishop, M., Cannings, C. (eds.)

Handbook of Statistical Genetics

2008

ISBN: 978-0-470-05830-5

Meyers, R. A. (ed.)

Genomics and Genetics **From Molecular Details to Analysis and Techniques**

2007

ISBN: 978-3-527-31609-0

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., Herwig, R.

Systems Biology **A Textbook**

2009

ISBN: 978-3-527-31874-2

Helms, V.

Principles of Computational Cell Biology **From Protein Complexes to Cellular Networks**

2008

ISBN: 978-3-527-31555-0

Emmert-Streib, F., Dehmer, M. (eds.)

Analysis of Microarray Data **A Network-Based Approach**

2008

ISBN: 978-3-527-31822-3

Andreas Ziegler and Inke R. König

A Statistical Approach to Genetic Epidemiology

With access to e-learning platform by Friedrich Pahlke

Second Edition



WILEY-VCH Verlag GmbH & Co. KGaA

The Authors

Prof. Dr. Andreas Ziegler
Dr. Inke R. König

Institut für Medizinische Biometrie und
Statistik
Universität zu Lübeck
Maria-Goeppert-Str. 1
23562 Lübeck
Germany

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.:
applied for

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <<http://dnb.d-nb.de>>.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Composition Uwe Krieg, Berlin

Printing and Binding betz-druck GmbH,
Darmstadt

Cover Design Formgeber, Eppelheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN: 978-3-527-32389-0

To Anke

To Marita and Nico Bøddeker

Contents

<i>Foreword to the First Edition</i>	vii
<i>Foreword to the Second Edition</i>	viii
<i>Preface</i>	xi
<i>Acknowledgments</i>	xv
1 <i>Molecular Genetics</i>	1
1.1 <i>Genetic information</i>	2
1.1.1 <i>Location of genetic information</i>	2
1.1.2 <i>Interpretation of genetic information</i>	5
1.1.3 <i>Translation of genetic information</i>	5
1.2 <i>Transmission of genetic information</i>	7
1.3 <i>Variations in genetic information</i>	10
1.3.1 <i>Individual differences in genetic information</i>	10
1.3.2 <i>Detection of variations</i>	12
1.3.3 <i>Probability for detection of variations</i>	16
1.4 <i>Problems</i>	18
2 <i>Formal Genetics</i>	21
2.1 <i>Mendel and his laws</i>	22
2.2 <i>Segregation patterns</i>	23
2.2.1 <i>Autosomal dominant inheritance</i>	24
2.2.2 <i>Autosomal recessive inheritance</i>	25
2.2.3 <i>X-chromosomal dominant inheritance</i>	26

2.2.4	<i>X-chromosomal recessive inheritance</i>	27
2.2.5	<i>Y-chromosomal inheritance</i>	28
2.3	<i>Complications of Mendelian segregation</i>	28
2.3.1	<i>Variable penetrance and expression</i>	29
2.3.2	<i>Age-dependent penetrance</i>	31
2.3.3	<i>Imprinting</i>	33
2.3.4	<i>Phenotypic and genotypic heterogeneity</i>	35
2.3.5	<i>Complex diseases</i>	36
2.4	<i>Hardy–Weinberg law</i>	38
2.5	<i>Problems</i>	43
3	<i>Genetic Markers</i>	47
3.1	<i>Properties of genetic markers</i>	47
3.2	<i>Types of genetic markers</i>	52
3.2.1	<i>Short tandem repeats (STRs)</i>	52
3.2.2	<i>Single nucleotide polymorphisms (SNPs)</i>	54
3.3	<i>Genotyping methods for SNPs</i>	57
3.3.1	<i>Restriction fragment length polymorphism analysis</i>	58
3.3.2	<i>Real-time polymerase chain reaction</i>	58
3.3.3	<i>Matrix assisted laser desorption/ionization time of flight genotyping</i>	61
3.3.4	<i>Chip-based genotyping</i>	61
3.3.5	<i>Choice of genotyping method</i>	63
3.4	<i>Problems</i>	65
4	<i>Data Quality</i>	67
4.1	<i>Pedigree errors</i>	68
4.2	<i>Genotyping errors in pedigrees</i>	70
4.2.1	<i>Frequency of genotyping errors</i>	70
4.2.2	<i>Reasons for genotyping errors</i>	71
4.2.3	<i>Mendel checks</i>	72
4.2.4	<i>Checks for double recombinants</i>	74
4.3	<i>Genotyping errors and Hardy–Weinberg equilibrium (HWE)</i>	76
4.3.1	<i>Causes of deviations from HWE</i>	77
4.3.2	<i>Tests for deviation from HWE for SNPs</i>	78
4.3.3	<i>Tests for deviation from HWE for STRs</i>	81
4.3.4	<i>Measures for deviation from HWE</i>	83
4.3.5	<i>Tests for compatibility with HWE for SNPs</i>	86

4.4	<i>Quality control in high-throughput studies</i>	91
4.4.1	<i>Sample quality control</i>	94
4.4.2	<i>SNP quality control</i>	97
4.5	<i>Cluster plot checks and internal validity</i>	98
4.5.1	<i>Cluster compactness measures</i>	101
4.5.2	<i>Cluster connectedness measures</i>	101
4.5.3	<i>Cluster separation measures</i>	101
4.5.4	<i>Genotype stability measures</i>	102
4.5.5	<i>Combinations of criteria</i>	102
4.6	<i>Problems</i>	109
5	<i>Genetic Map Distances</i>	113
5.1	<i>Physical distance</i>	113
5.2	<i>Map distance</i>	114
5.2.1	<i>Distance</i>	114
5.2.2	<i>Specific map functions</i>	115
5.2.3	<i>Correspondence between physical distance and map distance</i>	116
5.2.4	<i>Multilocus feasibility</i>	117
5.3	<i>Linkage disequilibrium distance</i>	118
5.4	<i>Problems</i>	123
6	<i>Family Studies</i>	125
6.1	<i>Family history method and family study method</i>	127
6.2	<i>Familial correlations and recurrence risks</i>	129
6.2.1	<i>Familial resemblance</i>	129
6.2.2	<i>Recurrence risk ratios</i>	131
6.3	<i>Heritability</i>	134
6.3.1	<i>The simple Falconer model</i>	135
6.3.2	<i>The general Falconer model</i>	137
6.3.3	<i>Kinship coefficient and Jacquard's Δ_7 coefficient</i>	138
6.4	<i>Twin and adoption studies</i>	141
6.4.1	<i>Twin studies</i>	141
6.4.2	<i>Adoption studies</i>	142
6.5	<i>Critique on investigating familial resemblance</i>	143
6.6	<i>Segregation analysis</i>	144
6.7	<i>Problems</i>	154

7	<i>Model-Based Linkage Analysis</i>	155
7.1	<i>Linkage analysis between two genetic markers</i>	156
7.1.1	<i>Linkage analysis in phase-known pedigrees</i>	156
7.1.2	<i>Linkage analysis in phase-unknown pedigrees</i>	160
7.1.3	<i>Linkage analysis in pedigrees with missing genotypes</i>	161
7.2	<i>Linkage analysis between a genetic marker and a disease</i>	167
7.2.1	<i>Linkage analysis between a genetic marker and a disease in phase-known pedigrees</i>	168
7.2.2	<i>Linkage analysis between a genetic marker and a disease in general cases</i>	172
7.2.3	<i>Gain in information by genotyping additional individuals; power calculations</i>	177
7.3	<i>Significance levels in linkage analysis</i>	180
7.4	<i>Problems</i>	184
8	<i>Model-Free Linkage Analysis</i>	189
8.1	<i>The principle of similarity</i>	190
8.2	<i>Mathematical foundation of affected sib-pair analysis</i>	192
8.3	<i>Common tests for affected sib-pair analysis</i>	193
8.3.1	<i>The maximum LOD score and the triangle test</i>	194
8.3.2	<i>Score- and Wald-type 1 degree of freedom tests</i>	201
8.3.3	<i>Affected sib-pair tests using alleles shared identical by state</i>	206
8.4	<i>Properties of affected sib-pair tests</i>	206
8.5	<i>Sample size and power calculations for affected sib-pair studies</i>	207
8.5.1	<i>Functional relation between identical by descent probabilities and recurrence risk ratios</i>	207
8.5.2	<i>Sample size and power calculations for the mean test using recurrence risk ratios</i>	209
8.6	<i>Extensions to multiple marker loci</i>	212
8.7	<i>Extension to large sibships</i>	213
8.8	<i>Extension to large pedigrees</i>	214
8.9	<i>Extensions of the affected sib-pair approach</i>	216
8.9.1	<i>Covariates in affected sib-pair analyses</i>	216
8.9.2	<i>Multiple disease loci in affected sib-pair analyses</i>	216

8.9.3	<i>Estimating the position of the disease locus in affected sib-pair analyses</i>	217
8.9.4	<i>Typing unaffected relatives in sib-pair analyses</i>	217
8.10	<i>Problems</i>	218
9	<i>Quantitative Traits</i>	221
9.1	<i>Quantitative versus qualitative traits</i>	222
9.2	<i>The Haseman–Elston method</i>	223
9.2.1	<i>The expected squared phenotypic difference at the trait locus</i>	225
9.2.2	<i>The expected squared phenotypic difference at the marker locus</i>	227
9.3	<i>Extensions of the Haseman–Elston method</i>	229
9.3.1	<i>Double squared trait difference</i>	230
9.3.2	<i>Extension to large sibships</i>	230
9.3.3	<i>Haseman–Elston revisited and the new Haseman–Elston method</i>	231
9.3.4	<i>Power and sample size calculations</i>	234
9.4	<i>Variance components models</i>	237
9.4.1	<i>The univariate variance components model</i>	237
9.4.2	<i>The multivariate variance components model</i>	238
9.5	<i>Random sib-pairs, extreme probands and extreme sib-pairs</i>	240
9.6	<i>Empirical determination of p-values</i>	243
9.7	<i>Problems</i>	245
10	<i>Fundamental Concepts of Association Analyses</i>	247
10.1	<i>Introduction to association</i>	247
10.1.1	<i>Principles of association</i>	247
10.1.2	<i>Study designs for association</i>	249
10.2	<i>Linkage disequilibrium</i>	250
10.2.1	<i>Allelic linkage disequilibrium</i>	250
10.2.2	<i>Genotypic linkage disequilibrium</i>	255
10.2.3	<i>Extent of linkage disequilibrium</i>	259
10.3	<i>Problems</i>	262
11	<i>Association Analysis in Unrelated Individuals</i>	265
11.1	<i>Selection of cases and controls</i>	266

11.2	<i>Tests, estimates, and a comparison</i>	266
11.2.1	<i>Association tests</i>	267
11.2.2	<i>Choice of a test in applications</i>	272
11.2.3	<i>Effect measures</i>	274
11.2.4	<i>Selection of the genetic model</i>	280
11.2.5	<i>Association tests for the X chromosome</i>	287
11.3	<i>Sample size calculation</i>	289
11.4	<i>Population stratification</i>	291
11.4.1	<i>Testing for population stratification</i>	293
11.4.2	<i>Structured association</i>	294
11.4.3	<i>Genomic control</i>	295
11.4.4	<i>Comparison of structured association and genomic control</i>	297
11.4.5	<i>Principal components analysis</i>	297
11.5	<i>Gene-gene and gene-environment interaction</i>	299
11.5.1	<i>Classical examples for gene-gene and gene-environment interaction</i>	299
11.5.2	<i>Coat color in the Labrador retriever</i>	301
11.5.3	<i>Concepts of interaction</i>	303
11.5.4	<i>Statistical testing of gene-environment interactions</i>	307
11.5.5	<i>Statistical testing of gene-gene interactions</i>	311
11.5.6	<i>Multifactor dimensionality reduction</i>	315
11.6	<i>Problems</i>	316
12	<i>Family-based Association Analysis</i>	319
12.1	<i>Haplotype relative risk</i>	320
12.2	<i>Transmission disequilibrium test (TDT)</i>	322
12.3	<i>Risk estimates for trio data</i>	325
12.4	<i>Sample size and power calculations for the TDT</i>	327
12.5	<i>Alternative test statistics</i>	329
12.6	<i>TDT for multiallelic markers</i>	330
12.6.1	<i>Test of single alleles</i>	330
12.6.2	<i>Global test statistics</i>	331
12.7	<i>TDT type tests for different family structures</i>	333
12.7.1	<i>TDT type tests for missing parental data</i>	334
12.7.2	<i>TDT type tests for sibship data</i>	336
12.7.3	<i>TDT type tests for extended pedigrees</i>	341

12.8	<i>Association analysis for quantitative traits</i>	344
12.9	<i>Problems</i>	346
13	<i>Haplotypes in Association Analyses</i>	349
13.1	<i>Reasons for studying haplotypes</i>	350
13.2	<i>Inference of haplotypes</i>	351
13.2.1	<i>Algorithms for haplotype assignment</i>	352
13.2.2	<i>Algorithms for estimating haplotype probabilities</i>	353
13.3	<i>Association tests using haplotypes</i>	356
13.4	<i>Haplotype blocks and tagging SNPs</i>	359
13.4.1	<i>Selection of markers by haplotypes or linkage disequilibrium</i>	360
13.4.2	<i>Evaluation of marker selection approaches</i>	363
13.5	<i>Problems</i>	364
14	<i>Genome-wide Association (GWA) Studies</i>	367
14.1	<i>Design options in GWA studies</i>	369
14.2	<i>Genotype imputation</i>	370
14.2.1	<i>Imputation algorithms</i>	370
14.2.2	<i>Quality of imputation</i>	371
14.3	<i>Statistical analysis of GWA studies</i>	372
14.4	<i>Multiple testing</i>	374
14.4.1	<i>Region-wide multiple testing adjustment by simulation</i>	375
14.4.2	<i>Genome-wide multiple testing adjustment by simulation</i>	376
14.4.3	<i>Multiple testing adjustment by effective number of tests</i>	377
14.5	<i>Analysis of accumulating GWA data</i>	378
14.5.1	<i>Multistage designs for GWA studies</i>	378
14.5.2	<i>Replication in GWA studies</i>	379
14.5.3	<i>Meta-analysis of GWA studies</i>	380
14.6	<i>Clinical impact of a GWA study</i>	383
14.6.1	<i>Evaluation of a genetic predictive test</i>	383
14.6.2	<i>Clinical validity of a single genetic marker</i>	385
14.6.3	<i>Clinical validity of multiple genetic markers</i>	386
14.7	<i>Outlook</i>	389
14.8	<i>Problems</i>	391

<i>Appendix</i>	
<i>Algorithms Used in Linkage Analyses</i>	393
A.1 <i>The Elston–Stewart algorithm</i>	394
A.1.1 <i>The fundamental ideas of the Elston–Stewart algorithm</i>	394
A.1.2 <i>The Elston–Stewart algorithm for a trait and a linked marker locus</i>	400
A.2 <i>The Lander–Green algorithm</i>	401
A.2.1 <i>The inheritance vector at a single genetic marker</i>	401
A.2.2 <i>The inheritance distribution given all genetic markers</i>	405
A.3 <i>The Cardon–Fulker algorithm</i>	412
A.4 <i>Problem</i>	414
<i>Solutions</i>	415
<i>References</i>	451
<i>Index</i>	489

Foreword to the First Edition

The field of genetic epidemiology is barely years old and, concomitant with the spectacular technological advances in molecular biology that have taken place over the past years, there have been many advances in statistical methodology developed for the analysis of data specific to this field. Although there is no dearth of textbooks on either statistical methods or epidemiology in general, very little is available for the person who already has a basic statistical background but is new to the special needs of genetic epidemiology. It is therefore with great pleasure that I write a foreword to this new text, written by two expert statisticians who have gathered together the important concepts and methods into a comprehensive introductory textbook for upper-class undergraduate or graduate students. The next decade will see an enormous need for persons to analyze the massive amounts of genetic information that will be produced as a result of the Human Genome Project and the HapMap Project and I see this new text, complete with examples, homework questions and copious references, as filling a need in the training of such persons.

ROBERT C. ELSTON

Cleveland, OH, USA, December 2005

Foreword to the Second Edition

The pace of discovery in human genetics over the past three years has been extraordinary. Before , family- and population-based studies had identified a handful of genetic markers associated with complex human traits and diseases. Now over markers associated with over traits are known, and it seems every issue of the top genetics journals contains at least one paper linking a new locus to a clinically-relevant trait.

The pace of technological change has also been exceptional. The Human Genome Project took years to produce billion basepairs of sequence; a single well equipped-lab can now produce that much sequence in a week.

While these are exciting times for researchers interested in human genetics, they are also challenging times—especially for students, teachers, and textbook authors. When a paper can go from cutting edge to passé in the time from submission to review, how much more difficult to keep a textbook current!

Drs. Ziegler and König have done a great service by thoroughly updating and revising their textbook. The second edition touches on the most important developments of the last few years, most notably the success of genome-wide association studies. And it keeps the strengths of the previous edition, which include its compatibility with the classroom (see, e.g., the companion web course and the exercises) and the list of internet resources at the end of each chapter. These references are important, because publicly available databases like dbSNP, the HapMap and Genomes Project are central to the practice of modern genetic epidemiology.

These (for the uninitiated) hard-to-find resources and specialized terminology can make the study of genetic epidemiology appear more difficult than it is. This book gives those looking for an introduction to this vibrant field the basic tools they need to find their own way around current research and to learn by doing.

PETER KRAFT

Boston, MA, USA, December 2009

Preface

This book presents the second edition of a statistical approach to genetic epidemiology. But what do we mean by genetic epidemiology? According to the definition of Khoury, Beaty, and Cohen [353], it is the discipline investigating genetic and environmental factors that influence the development and distribution of diseases. It differs from epidemiology in that explicitly genetic factors and similarities within families are taken into account. On the other hand, it can be distinguished from medical genetics by considering populations rather than single patients or families.

As the name implies, genetic epidemiology is an interdisciplinary subject, and it is a working field for scientists from different backgrounds. In contrast to many outstanding textbooks on molecular genetic techniques, such as *Human Molecular Genetics 3* by Strachan and Read [636], this book introduces statistical concepts to current approaches used in genetic epidemiology. It is written at a level that it should make it useful to undergraduate and graduate students as well as researchers. The necessary background in statistics is an introductory course to statistical testing and estimation. Excellent books allowing one to revive this knowledge include, for example, Ref. [498]; for a little knowledge about likelihood ratio, score, and Wald tests, see, for example, Hills [295], Section “Hypothesis Testing,” or Kleinbaum [358], pp. 128–136.

Just like Gaul at Cesar’s time [99], this volume is divided into three different parts. The material in this text assumes no background in biology, molecular biology, or genetics. The required fundamentals of these topics are described in the first part of this book, covering Chapters 1 through 4. Chapter 1 provides an introduction to the basic molecular genetic mechanisms that are required to serve as a background for understanding the statistical methods in the later chapters. However, for comprehensive overviews on human molecular genetics the reader may refer to standard textbooks

[254, 636]. Mendel's laws and their consequences for familial inheritance patterns as well as the important population genetic Hardy-Weinberg law are discussed in Chapter 2 on formal genetics. Genetic variability is the key to studying the genetic architecture of a disease, and it can be measured by genetic markers, which are the topic of Chapter 3. Two different types of molecular genetic markers are at present in use, and we discuss these together with standard nomenclature. Because different molecular biological technologies have various pros and cons, we have integrated a short description of the most important technologies, and this section has been written by our colleague Jeanette Erdmann, an experienced molecular biologist. Before we introduce the techniques for statistically analyzing genetic disorders, Chapter 4 considers data quality control in genetic epidemiological studies. Caused by the new chip technologies, data quality control has increased dramatically in its importance, and this chapter has therefore more than tripled in size compared to our first edition.

The second part of this book on segregation and linkage analysis starts with Chapter 5 on genetic distance. Many colleagues have suggested to include an additional chapter on family studies not involving genetic markers. We have therefore added a corresponding new Chapter 6 and restructured the subsequent chapters on linkage analysis. An in-depth understanding of segregation analysis relies on a specific standard algorithm. To thoroughly comprehend segregation analysis, a specific standard algorithm is required, which is described in the Appendix owing to its high technical level. Model-based linkage analysis is the topic of Chapter 7, while model-free linkage approaches to dichotomous and quantitative traits are considered in Chapters 8 and 9, respectively. An in-depth understanding of the linkage methods relies on some standard algorithms enabling one to deduce genetic marker information. Because these are quite technical, they have been placed in the Appendix.

While linkage analysis relies on segregation information in families, association studies focus on differences at the population level. They are the subject of the last part of this book, starting with Chapter 10, where the fundamental concepts of association analyses are discussed. The analysis of association itself using single markers is presented in Chapters 11 and 12. Chapter 11 deals with studies utilizing data from unrelated individuals, and it has undergone a substantial revision and extension. Specifically, tests and estimates are described separately, and population-based measures are described in greater detail. The different test statistics are derived, and equivalent formulations are presented. The section on controlling population stratification has been supplemented by new methods, and a section on studying interactions has been added. Chapter 12 is concerned with family-based association studies, and Chapter 13 is devoted to the topic of studying haplotypes. Genome-wide association studies have been a great success for unraveling the genetic basis of complex genetic diseases in the past three years, and they are described in the new Chapter 14.

At the end of each chapter, the reader will find a number of problems covering both theoretical derivations and practical calculations. They can be solved by paper and pencil and require only the help of a pocket calculator. However, we refer to the software that could be utilized for computation in the text and list the relevant URLs at the end of each chapter. The solutions to all problems are detailed at the end of the

book. Throughout the book, new, relevant terminology is indicated in italics where first introduced.

An important change to the first edition both in content and in presentation is the distinction between two different versions of this book. While one version is the standard printed one, the other version provides access to an online course developed and implemented by our colleague Friedrich Pahlke. The online course has been tested over several years, and we have found it a very useful resource for students, covering the first and the third part of the book, i.e., the molecular genetic background and association analysis. We teach the complete book in a two semester course with points according to the European Credit Transfer and Accumulation System (ECTS) and the online course with ECTS points on an annual basis. The present form of the distance learning course is an beginner's course. It starts with a distance learning phase for Chapters 1 through 4, followed by a -day phase of attendance for Chapters 9, 10, 11, 14 and either 12 or 13. This course has the value of ECTS. If the course is taught with a -day phase of attendance, we value it . ECTS. For the additional day, we recommended a more intensive discussion of Chapters 11 and 14.

ANDREAS ZIEGLER AND INKE R. KÖNIG

Lübeck, Germany

Acknowledgments

This book would not have been completed without the help of many colleagues from our Institute of Medical Biometry and Statistics. Friedrich Pahlke again wins the pole position for preparing the online course available with one version of this book and almost all of the figures. He was funded by the grant “Training in Genetic Epidemiology” from the German Ministry of Education and Research within the National Genome Research Network. We are extremely grateful to Jeanette Erdmann who has written the section on molecular genetic technologies. Many of the new results presented in the chapter on data quality emerged from joint work of A.Z. with our colleague Arne Schillert. We express our special thanks to our PhD student Christina Loley and our former PhD student Anika Großhennig who contributed to several parts of the book.

Innumerable suggestions for improvement were made by our Computational Life Science students from the University at Lübeck during our two-semester long courses. We are also grateful to all colleagues who have been trained remote and in Lübeck in our course “Training in Genetic Epidemiology” over the past four years. We sincerely apologize for all the errors of the first edition, but at the same time we are very grateful to all colleagues who have made us aware of these.

Wolfgang Lieb, Jeanette Erdmann, Katrina A. B. Goddard, Suzanne Margaret Leal, Lize van der Merwe, and Konstantin Strauch helped us substantially to improve the manuscript by commenting on specific chapters. Furthermore, we gratefully acknowledge Andrew Collins, Jochen Hampe, Tanja Zeller, and the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory for providing data for our independent analysis. We are very grateful to Yvonne Eckel-Fiedler, president of the Retriever Club Europa, for providing beautiful photos of her dogs.

Many thanks go to our collaborators who over the past years have supplied us ideas and problems as well as data and figures for this book. They have motivated us to further work in this area, and we are especially grateful to Jeanette Erdmann and Heribert Schunkert from the Medical Clinic II Lübeck, Tanja Zeller and Stefan Blankenberg from the Department of Medicine II Mainz, the Cardiogenics Project Group, Hans Konrad Schackert and Guido Fitze from the Arbeitsgruppe Chirurgische Forschung, Dresden, and Christine Klein from the Clinical Molecular Neurogenetics Group, Lübeck.

This work has its roots in many courses, and the one that can be seen as its foundation was held by A.Z. in 2000 as a 1-week course in Pelotás, Brazil. It was outstandingly organized by Hiram Larangeira de Almeida Jr., and A.Z. will never forget the hummingbirds and the fresh fruit on his hacienda. More fruitful discussions helped the first edition of this book on its way during the sabbatical of A.Z. in 2005 at the Division of Molecular and Genetic Epidemiology, headed by Robert C. Elston, in Cleveland, Ohio.

We are also indebted to Hugo Marth from the Graphical Office at the University at Lübeck for his expert design of the book cover and to Gregor Cicchetti and Andreas Sendtko from Wiley-VCH for their constant excellent support. With this edition, the copyright of the figures has been transferred to Wiley-VCH. However, we still have the permission for using all own material for other publications.

Finally, acknowledgments go to our families. From I.R.K. to my husband Peter and my son Timotheus for sacrificing weekends and bolstering me up with loving patience and trust. Also, to my father-in-law Karl for sharing helpful experience. From A.Z. to my daughters Rebecca Elisabeth and Sarah Johanna for pointing out that there is a life beyond writing a book and, finally, to my loving wife Anke who accepted the unpredictable hours and always strongly encouraged me to complete this work.

A. Z. and I. R. K.

1

Molecular Genetics

Working on genetic epidemiological questions, we have to deal with a variety of information. On the one hand, as in other epidemiological approaches, we use clinical and environmental information. For instance, we might be interested in the relationship between fairer or darker skin, the extent of sun exposure, and the development of malignant melanoma. On the other hand, our specialty is to incorporate genetic information. We might therefore want to look at whether these relationships are mediated by genetic factors. The result might suggest that because of their genetic background, some people are more susceptible to melanoma even though they have darker skin. But what exactly is this genetic background? The aim of this chapter is to familiarize ourselves with this kind of genetic data.

Specifically, we first need to understand what the biological substance of the genetic information is, where it is located in the human body, what it means, and how it is translated for further use. These questions will be answered in Section 1.1. What distinguishes genetic information from other information is that parts of it are transmitted from generation to generation. Understanding the biological mechanisms that underlie this concept forms the basis for later test statistics, and this will be the focus of Section 1.2. Finally, to be useful for statistical purposes, genetic information has to be subject to variation. Hence, the last section of this chapter is devoted to the questions of how individuals differ with regard to their genetic information, how variations can occur, how they can be detected, and how likely the detection is.

This chapter will not give a comprehensive overview on human molecular genetics. Instead, we will focus on issues that are important to understand the statistical methods introduced later. For more in-depth descriptions, the reader is referred to standard textbooks on this topic (e.g., Refs. [254, 636]). Readers who are already familiar

with molecular genetics could certainly skip this chapter, especially those who feel comfortable with the problems at the end of this chapter.

1.1 WHAT IS THE NATURE OF GENETIC INFORMATION?

1.1.1 Where is the genetic information located?

Every human cell except the red blood cells has a nucleus that carries an individual's genetic information in *chromosomes*. At the same time, chromosomes are found almost exclusively in the nucleus of the cell. Hence, almost every cell of the body carries the information that is required for the entire organism. The chromosomes are composed of *deoxyribonucleic acid (DNA)* and proteins. The DNA is the carrier of the genetic information, whereas the protein components provide different functions.

The DNA is a large molecule consisting of two strands. Each strand has a linear backbone of alternating sugar (deoxyribose) and phosphate residues. To facilitate the description of the structure, the five carbon atoms of the deoxyribose are consecutively numbered from 1' to 5' (see Figure 1.1, left-hand side). Covalently attached to the backbone is a sequence of bases. Here, four bases are found, with adenine (A) and guanine (G) being purines, and cytosine (C) and thymine (T) being pyrimidines. The structural unit of one sugar with an attached base is called a nucleoside, and one nucleoside with a phosphate group tied to the carbon atom 5' or 3' makes one *nucleotide*. In addition to this structure of a single strand, the two strands of the DNA molecule are connected by a hydrogen bond between two opposing bases of the two strands. Specifically, thymine always bonds with adenine via two hydrogen bonds, and cytosine with guanine via three hydrogen bonds. The resulting DNA resembles a ladder whose sides are connected by the bases. However, because of the chemical nature of its components, the ladder does not go straight up but is slightly twisted, which is why it has been described as a *double helix* (Figure 1.1, right-hand side).

Any two DNA fragments differ only with respect to the order of their bases. Therefore, the genetic information we are looking for is coded exactly in the linear sequence of the bases of the DNA fragment. This linear sequence of the bases is called the *primary structure* of the DNA. Because the bonding of the bases between the two strands is specific, the two strands can be said to be complementary, meaning that the sequence of one strand can be exactly inferred from the other. As a consequence, if one wants to describe the sequence, it suffices to write the sequence of one strand. Here, it has become customary to write the sequence in the 5' to 3' direction. The basic length unit of the DNA is one nucleotide, or one *basepair (bp)*, which refers to the two bases that connect the two strands. In total, the human DNA contains about 3.3 billion bp.

As a second carrier of genetic information in addition to the DNA in chromosomes, copies of parts of the DNA are found in smaller molecules called *ribonucleic acid (RNA)* in the nucleus and the surrounding plasma of the cell. The RNA is constructed in a way very similar to the DNA but shows four main differences. In addition to being much shorter, RNA consists of only a single strand instead of two. Also,

this single strand is slightly different from a single DNA strand in that the sugar component is made up of ribose instead of deoxyribose. Finally, although the other bases are the same, uracil (U) is found instead of thymine.

The total set of genetic information is distributed in a series of 23 chromosomes. Of these, 22 are autosomes that are consecutively numbered from 1, the longest chromosome, to 21 and 22, the shortest chromosomes. Chromosomes 1 and 2 encompass more than 240 million bp, whereas chromosomes 21 and 22 have no more than 50 million bp. The remaining chromosome is one of the sex chromosomes X and Y, with lengths of about 152 million bp and 50 million bp, respectively. A cell containing a single set of chromosomes with all 22 autosomes and one of the two sex chromosomes is termed to be *haploid*. A regular human cell, however, is *diploid*,

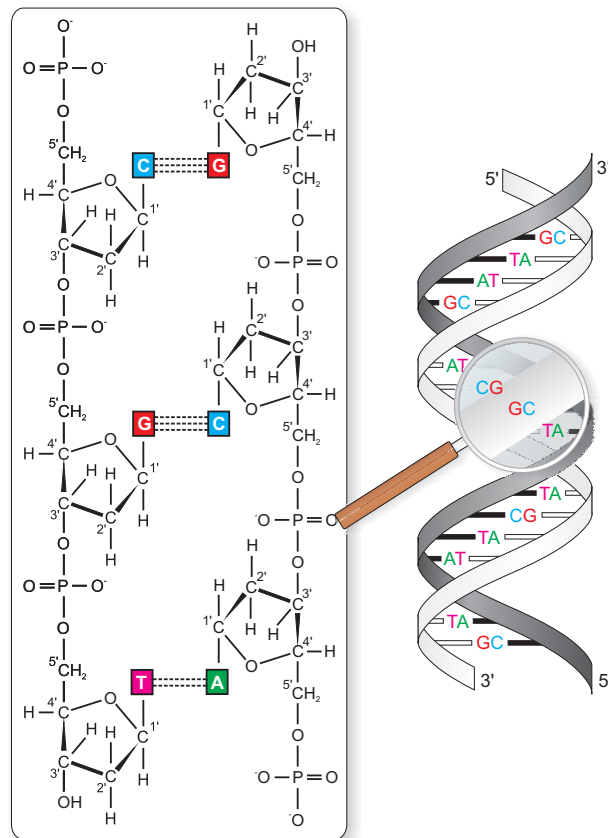


Fig. 1.1 Schematic structure of the DNA. The left-hand side shows the sugar phosphate backbone of the DNA with attached carbon atoms and bases. Two backbones are connected via the bases with two or three hydrogen bonds. The right-hand side is zoomed out to depict that the two strands are slightly twisted, forming the double helix.

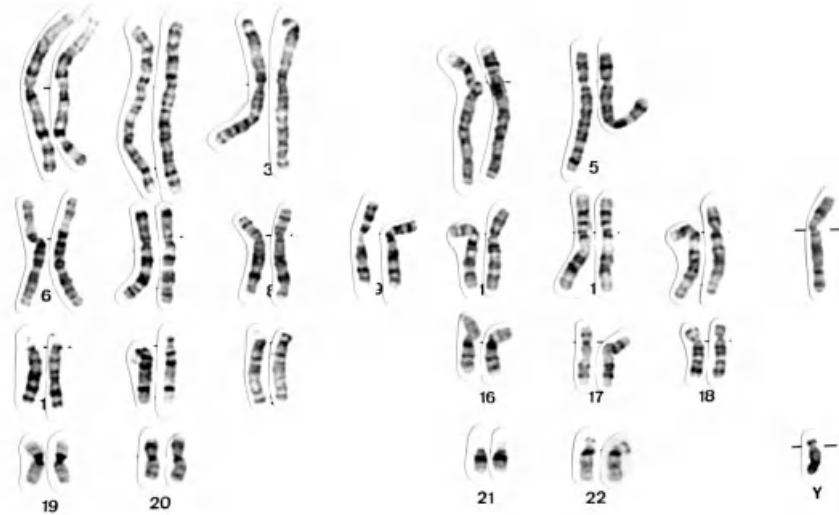


Fig. 1.2 Karyogram of a healthy male. This figure was kindly provided by the Institut für Humangenetik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany.

meaning that it contains a double set, with one coming from the father and the other coming from the mother. Hence, a regular cell has $2 \cdot 22 = 44$ autosomes and two sex chromosomes. Specifically, cells in a female contain two X chromosomes, whereas males carry one X chromosome and one Y chromosome in their cells. There are two regions on the X and Y chromosomes deserving special attention. They are the *pseudoautosomal regions* PAR1 and PAR2 which are homologous sequences of nucleotides on the X and Y chromosomes. The PARs behave like an autosome. Thus genes in this region are inherited in an autosomal rather than a strictly sex-linked fashion. PAR1 comprises 2.6 million bp of the short-arm tips of both X and Y chromosomes in humans. PAR2 is located at the tips of the long arms, spanning 320 kilo bp; the PARs are comprehensively reviewed in Ref. [437].

At a specific point in the division of a cell, the chromosomes can be made visible under the light microscope. When they are stained with certain dyes, they reveal a specific pattern of light and dark bands reflecting regional variations in the amounts of the bases. Differences with respect to length and the banding pattern allow the chromosomes to be easily distinguished from each other, as is visible in Figure 1.2. This

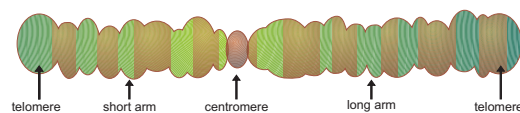


Fig. 1.3 Schematic structure of a chromosome with its characteristic elements.

figure shows an example of the *karyotype* of a healthy male, which is the constitution of the chromosomes of an individual according to standard classification systems. On a single chromosome, different structural elements are distinguished (see Figure 1.3). At the ends of the chromosome, the *telomeres* have special functions involved in the duplication of the chromosomal ends during cell division. A corresponding structure near close to the middle of the chromosome is the *centromere*. The short arm of the chromosome is usually termed *p* for *petit* (small) and the long arm, *q* for *queue* (tail). Accordingly, the telomeres are referred to as *pter* and *qter*, respectively.

1.1.2 What does the genetic information mean?

After explaining that the genetic information is stored in the linear sequence of the bases of DNA or RNA, we now need to read this information. For this, it is important to note that most functions in human organisms are carried out by proteins. Proteins consist of polypeptides, which are nothing but a linear sequence of repeating units that are called *amino acids*. In humans, 20 different amino acids occur. Hence, there needs to be a translation between the linear sequence of four different bases in the DNA or RNA into a linear sequence of 20 different amino acids for a protein.

How does a base sequence translate into protein structure? It has been found that three bases, a triplet, code for one amino acid. Accordingly, the sequence of three bases is called a *codon*. Using such a three-letter code, it is possible to form $4^3 = 64$ different codons from four possible bases. Because there are only 20 amino acids that need to be coded, the genetic code can be said to be degenerate, with the third position often being redundant. Depending on the starting point of reading, there are three possible variants to translate a given base sequence into an amino acid sequence. These variants are called *reading frames*. The beginning of the translation process from bases into amino acids is signaled by special functional start codons, mostly A(denine)-U(racil)-G(uanine). The opposite stop codons, for instance, U(racil)-A(denine)-G(uanine), will terminate the translation. It should be noted that in practice, the translation of bases into amino acids does not use DNA but RNA; the base uracil appears in the code that is displayed in Figure 1.4.

1.1.3 How is the genetic information translated?

As we have seen, genetic information is basically a construction plan for proteins. Hence, we now need to understand how, beginning with the DNA, proteins are actually synthesized. Because of the importance of this process, protein synthesis has been termed the *central dogma of molecular biology*. To anticipate the major pathway, this dogma has been expressed as:

DNA makes RNA, RNA makes proteins, proteins make us.

The overall process of protein synthesis can be partitioned into two steps. First, we have to remember that DNA is stationary and located in the nucleus of the cell. In contrast, protein synthesis takes place in the surrounding plasma of the cell. Therefore, the first step involves the *transcription* of DNA into messenger RNA

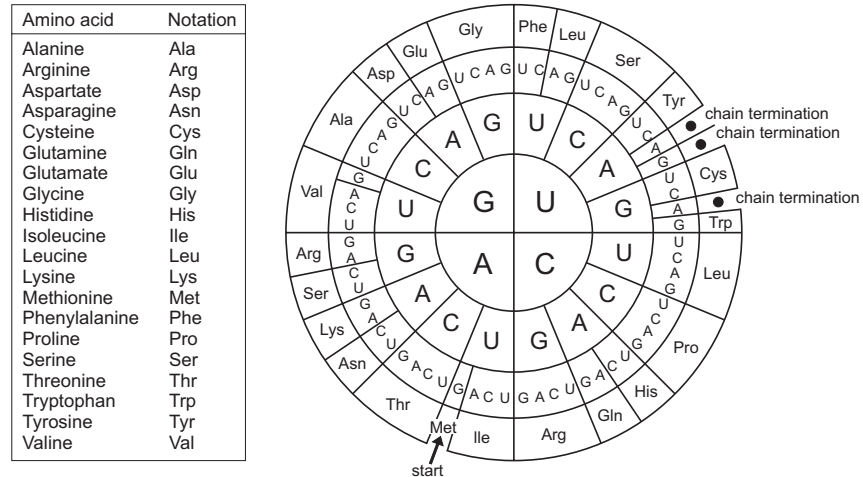


Fig. 1.4 Codon table to translate RNA into amino acid. The codon table is read from center to outside, so that, for example, the sequence A (center) U (middle) G (outside) codes for the amino acid Met (methionine) and at the same time serves as an initiation site for translation (see Section 1.1.3).

(mRNA) that then carries the information from the nucleus into the plasma. More specifically, the DNA double helix is unwound and unzipped into single strands. Catalyzed by the RNA polymerase enzyme, the strand in the 5' to 3' direction is used as a template. Led by the polymerase that migrates from one nucleotide to the next, free RNA nucleotides anneal to the DNA and are tied together to a strand. After coming across a stop signal, the RNA strand uncouples from the DNA that is then again zipped together with its complementary strand. The resulting RNA subsequently leaves the nucleus of the cell. Because the RNA now has the same direction and base sequence as the strand of the DNA that was not used as a template, this non-template strand is called the *sense strand*. In contrast, the template DNA strand is termed the *antisense strand*.

After the transcription, the RNA molecule is further edited. Specifically, the so-called *introns* are cut out, whereas *exons* are spliced together. For a large number of segments, multiple alternative splicing variants exist. Finally, unique features are added to each end of the transcript to make a mature mRNA.

The transcription process is regulated by a number of factors. For example, several kinds of the enzyme RNA polymerase and associated protein transcription factors regulate the specificity and rate of transcription. Specific regions of the DNA called *promoters* that include binding sites for the RNA polymerases control the initiation of the transcription. In addition, transcription can also be regulated by variations in the DNA structure or by chemical changes in the bases. One such important chemical modification is methylation. The degree of methylation tends to be negatively correlated with the functional activity of DNA and plays an important role, for example, in genomic imprinting (see Section 2.3.3).

The second step of protein synthesis involves the *translation* of the genetic information that is now contained in the mRNA from genetic code into proteins. The process of translation takes place in the cell plasma at cell organelles called *ribosomes*. After the actual synthesis of amino acid sequences, the proteins are further modified. The advantages of this procedure are that only small segments of DNA are used at a time and that many transcriptions and translations can be rapidly made.

After describing the nature of our genetic information in some depth, we should briefly turn to the question of what a gene actually is. Although definitions vary from source to source, a *gene* is often defined to be a functional unit of the DNA that encodes a product, usually a protein, and includes exons, introns, and small regions of DNA immediately preceding or following the transcribed region. The overall genetic makeup of one individual is termed the *genotype*. In contrast, the observable characteristics of the individual are called the *phenotype*. With *coding* DNA the section that will finally be translated into proteins is described, consisting mostly of exons, whereas *non-coding* DNA includes both the introns and the sequences between genes.

It is important to realize that only a small proportion of the DNA is ever transcribed, and only a proportion of the transcribed mRNA is finally translated into protein. With the length of a gene going up to 10,000 bp and the human genome containing about 20,000 to 25,000 genes according to the National Human Genome Research Institute, genes make up only about 1% of the total DNA. The non-coding DNA contains repetitive DNA, pseudo-genes, or regulatory sequences. These provide critical functions such as indicating the beginning and end of a coding sequence or providing binding sites for proteins that turn the transcription on and off. In addition, in different cells, different units of DNA are transcribed.

1.2 HOW IS GENETIC INFORMATION TRANSMITTED FROM GENERATION TO GENERATION?

By now, we know what the genetic information is, where it is located, and what it means. The specific peculiarity, however, is that it is partly transmitted from generation to generation and that information from two individuals, the mother and the father, is combined in the offspring. This sequence of distribution and combination forms the basis for a number of test statistics that will be introduced later.

Two steps can be distinguished in the process of transmission. First, in *meiosis*, the genetic information of each parent is transferred into germ cells, the *gametes*. Second, gametes of the father and the mother fuse to form the *zygote* that carries information from both parents.

Meiosis is a special form of cell division in which a single cell is divided into four daughter cells. Because the resulting gametes each have only half the number of chromosomes of the progenitor cell, these are haploid. To be specific, meiosis includes one round of chromosome duplication and two rounds of cell division (Figure 1.5):

1. The process begins with a regular diploid cell ($2n$), which means that there is one paternal and one maternal double strand or chromosome. Matching maternal and paternal chromosomes are termed to be *homologous*.
2. DNA replication: The DNA is duplicated, and as a result, each chromosome now contains two identical DNA double strands. These are termed *sister chromatids*. Hence, there is a total of four double strands ($4n$) in the cell.
3. Forming of bivalents: Homologous chromosomes are connected to form *bivalents*.
4. Possible crossing over: In this stage, exchange of genetic material between maternal and paternal strands is possible.
5. Meiotic division I: Non-sister chromatids are separated, while sister chromatids remain paired. This results in two diploid cells containing the sister chromatids.
6. Meiotic division II: The sister chromatids are separated, resulting in four haploid cells (gametes).
7. During fertilization, the fusion of the genetic material from two gametes leads to a restitution of the diploid status.

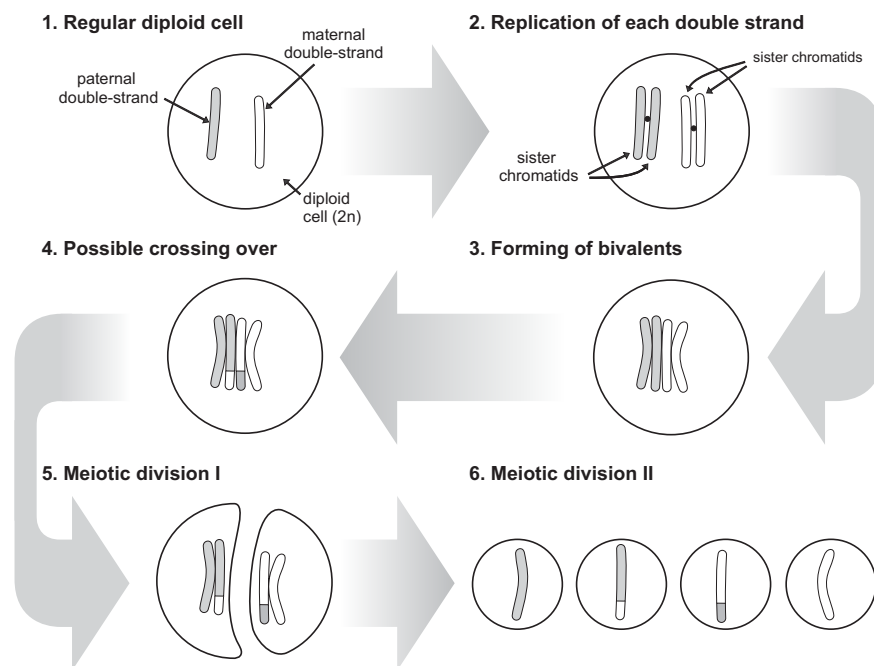


Fig. 1.5 Six stages of the meiotic process. The DNA of a regular diploid cell (1) is replicated (2), and the homologous chromosomes anneal to form bivalents (3). After a possible crossing over (4), the non-sister chromatids are separated into distinct cells (5). Finally, the sister chromatids are distributed to separate haploid gametes (6).

An important feature of meiotic division is that homologous chromosomes are distributed randomly and independent of each other to the gametes. Thus, the resulting gamete most likely contains some chromosomes that the individual has inherited from the mother and others that have come from the father, but the specific combination of chromosomes is random. Hence, there is a large number of possible combinations of the chromosomes in a single cell. With a total of 23 chromosome pairs, the number of possible combinations in one gamete from one parent is 2^{23} . When the fusion of one gamete with another is considered, the number of possible chromosomal combinations is doubled.

In addition, further shuffling of genetic material is possible. In step 4 of the meiotic process, an exchange of genetic material between the maternal and paternal DNA strand, might occur. During meiotic division I, non-sister chromatids are crossed over and form visible chiasmata. At their points of crossing over, the chromatids may break at homologous points and reunite with the non-sister chromatid. This can happen not only once, at one point on the chromosome, but several times, and the results are quite different. If we consider the two chromosomal segments *A* and *B* on the two chromatids 1 and 2, one possibility would be that no crossing over occurs between these segments (see Figure 1.6a). Hence, *A*1 will remain on the same chromatid with *B*1, and so will *A*2 with *B*2. In contrast, one crossing over might have taken place (see Figure 1.6b). The result will be a shuffling of the chromosomal segments with *A*1 and *B*2 on one chromatid and *A*2 and *B*1 on the other. Consequently, a *recombination* between segments *A* and *B* has happened. A third possibility is that two crossing overs occur (see Figure 1.6c). Although the intermediate segment is exchanged, this leads to the same result regarding the upper and lower segments as with no crossing over, namely, *A*1 and *B*1 being on the first chromatid and *A*2 and *B*2 on the second. Hence, no recombination has taken place between the segments. To generalize from this, we can deduce that a recombination between two chromosomal segments can be observed whenever there is an odd number of crossing overs between them. Even numbers of crossing overs, on the other hand, will even out any interchanging.

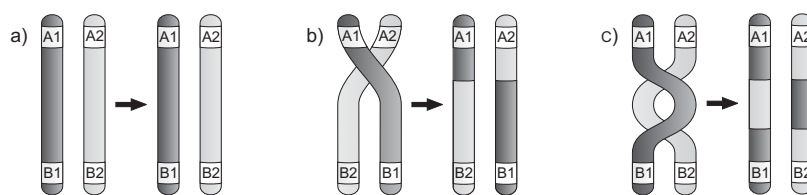


Fig. 1.6 Recombination between chromosomal segments *A* and *B* on chromatids 1 and 2. a) No crossing over might occur. b) One crossing over might occur, leading to a recombination between *A* and *B*. c) Two crossing overs might occur, leading again to no visible recombination between *A* and *B*.

Crossing overs are not rare events. Specifically, the mean number per cell is about 55 in males and is approximately 50% higher in females. The further apart two segments are from each other on the chromosome, the greater the probability is that

a crossing over will occur between them. Hence, the greater the distance, the higher the probability for a recombination to be observed between the segments. Turned around, the probability for a recombination, termed the *recombination fraction* θ , can be used as a measure of the distance of two chromosomal segments. If the segments were located very close to each other, they would almost never be separated by a crossing over, hence θ would approximate 0. If, at the other extreme, they were situated on different chromosomes, the first meiotic division would distribute them to different cells in about half the cases, thus leading to a θ of 0.5. Similarly, two segments on the same chromosome but very far apart will almost certainly be subject to a recombination because of the high number of mean crossing overs. Hence, they will again be separated by the first division with equal probability, and θ again approximates 0.5. This forms the basis for different measures of distance between segments and will be discussed in detail in Chapter 5.

1.3 WHAT IS INDIVIDUAL VARIATION IN GENETIC INFORMATION?

1.3.1 How do individuals differ with regard to their genetic information?

In the previous sections, we have assumed that the DNA strands inherited from father and mother are not identical. Generally, we can state that different individuals usually have the same genes, but not exactly the same DNA sequence. An exception is identical twins, who share the same DNA sequence. To use a more general term, we focus on the variability of a DNA *locus*, which is a specific DNA segment defined by its sequence of bases. Variants at one DNA locus are termed *alleles*. Concerning diploid cells, the two alleles at one locus define the genotype at this locus. If an individual carries the same alleles on both DNA strands, he is said to be *homozygous* at the locus. If the alleles are different, the individual is *heterozygous* for the respective DNA locus. But how do differences at a DNA locus come about in the first place?

There are two main sources for any variation in the DNA sequence:

1. During cell division, an abnormal pairing of the chromatids leads to a rearrangement of chromosomal segments. This in turn can result in insertions, deletions, inversions, translocations, or duplications of whole chromosomal segments.
2. During DNA replication, point mutations with substitutions of one nucleotide for another occur. This may result in transitions (exchanging one purine for another or exchanging one pyrimidine for another) or transversions (exchanging a purine for a pyrimidine or vice versa), leading to a synonymous or a non-synonymous base sequence. If instead one or more (except for multiples of three) nucleotides are inserted or deleted, a frameshift mutation emerges.

Mutation, the second source of variation, is defined as a state that permanently changes the function of the gene. Mutations happen because the DNA replication is not absolutely flawless. Most of the errors are corrected immediately, but if an error

is not detected or repaired, a mutation has occurred. In fact, the mutation frequency for a single gene ranges from around 10^{-4} to 10^{-6} . As a result, the base sequences of any two human individuals differ with a frequency of about one base out of 1000.

The consequences of a mutation depend on the kind and location of the variation. If one nucleotide is substituted for another, the redundancy of the genetic code may yield the same protein being synthesized, which is called a synonymous mutation. If the resulting codon codes for a different amino acid, the synthesized protein will be different and hence the mutation is called non-synonymous. If nucleotides are inserted or deleted from the DNA, we have to distinguish between two cases. First, the insertion or deletion of a multiple of three nucleotides leads to additional or fewer amino acids being coded. Second, any other number of insertions or deletions will lead to a shift of the reading frame; hence, a different protein will be synthesized.

A term that is sometimes used interchangeably with mutation is *polymorphism*. Although there are definitions that include the consequences of the variation, we will use the term polymorphism to signify a variation in DNA sequence with a frequency of at least 1% in at least one human population.

With novel genetic variants being detected and described, it is important to have a standardized nomenclature to denote them [29]. As a general recommendation, variations should be described on the level of DNA. In this case, the naming itself contains the following elements:

- g. for genomic DNA,
- Position of the variation. This is given by the number of the nucleotide in which the variation begins. For this, the base adenine from the start codon is denoted nucleotide +1 and the nucleotide in 5' to +1 is denoted -1. If you want to specify a range, the symbol “-” should be used.
- Specific change. The most common changes are symbolized by > for a substitution, *del* for a deletion, *ins* for an insertion, *dup* for a duplication, and *inv* for an inversion.

Let us consider the following examples: g.23G>T is a substitution of guanine by thymine at position 23, g.3-5delTCC means that the bases thymine, cytosine, and cytosine are deleted from nucleotides 3 to 5, and g.3-4insT symbolizes the inversion of thymine between nucleotides 3 and 4.

If the change has taken place in an intron, the nomenclature suggests to denote the change by *IVS* followed by the number of the intron and the position within the intron. For instance, IVS4 + 2A>C means that adenine was substituted by cytosine in intron 4 at position 2.

It is out of scope of this book to describe the nomenclature of variations on different levels such as genes, RNA, or proteins. These and more complex cases as well as current updates can be found on the web site of the Human Genome Variation Society and/or in Refs. [159, 309, 682].

1.3.2 How can individual differences be detected?

We now know how variations in the DNA sequence may occur and which kinds of mutations are possible. A series of molecular biological laboratory techniques have been developed to detect a novel mutation in a given DNA sequence. In this section, we will describe the basic procedure for sequencing a DNA segment. Because larger amounts of DNA are required for this and other techniques, the amplification of DNA segments plays an important role. Hence, we will explain the basics of the polymerase chain reaction as an amplification technique. Common technologies for single nucleotide polymorphism (SNP) genotyping are described in Section 3.3.

1.3.2.1 Sequencing of DNA segments Different sequencing techniques have been developed over the past few decades. Because our aim is to explain the principles underlying these laboratory techniques, we will focus on one specific technique: the *dideoxy sequencing* method developed in the late 1970s, also termed the *chain-termination method* or, after its developer, the *Sanger method*.

The requirements for the original sequencing reaction are four test tubes with the following reagents:

- A template DNA segment that is to be sequenced. This is prepared as a single-stranded DNA (denatured).
- A set of DNA deoxynucleotides (dATP, dTTP, dCTP, dGTP) to build new segments of DNA.
- Single-stranded labeled DNA segments complementary to a short region on either side of the template sequence. These are called *primers*.
- A specific polymerase enzyme to catalyze the synthesis of the new DNA.
- In each tube, a small amount of one dideoxynucleotide triphosphate that terminates the chain growth wherever it is incorporated (a specific base stop).

The actual sequencing reaction works as follows (see Figure 1.7):

1. Annealing of the primer to the template.
2. Elongation of the primer to complementary DNA using the available deoxynucleotides.
3. Termination of the elongation as soon as a base-specific stop nucleotide is incorporated that cannot form a phosphodiester bond with the next incoming deoxynucleotide.

In each test tube, the result will be a set of new DNA strands of different lengths. To be specific, each length signals the incorporation of a stop nucleotide, hence the occurrence of the respective base in the template DNA. The fragments of different sizes from the four test tubes are then separated by a procedure called gel electrophoresis.

The purpose of gel electrophoresis is to sort a number of DNA segments according to their length. Basically, an agarose gel is poured with small molds and the fragments are poured into the molds, meaning that the gel is loaded with DNA fragments. Then, an electrical field is charged. Because of the current, the fragments start to migrate through the gel. The important point is that their speed of migration depends on their length: the longer they are, the more weight they have and the slower they travel.

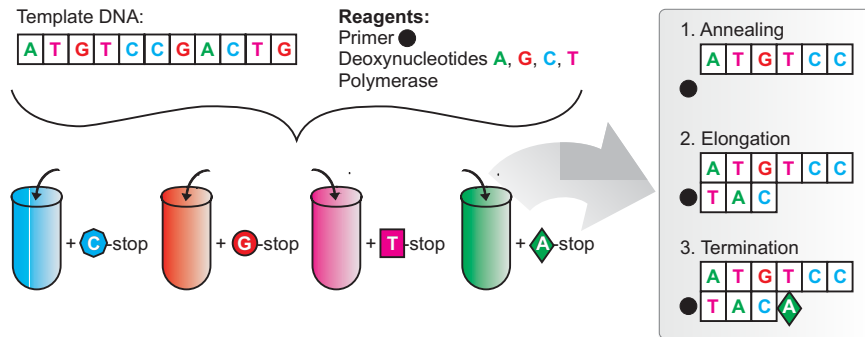


Fig. 1.7 Sequencing of DNA segments. The template DNA and reagents (top left) are mixed in four test tubes that differ only with regard to the added dideoxynucleotide triphosphate serving as a base-specific stop sign (bottom left). In each test tube, the primer will anneal to the template (top right) and be elongated to form a complementary strand (middle right). If, for example, dideoxyadenosine triphosphate (di-dATP) has been added to the mixture, as in the right test tube, instead of the usual deoxyadenosine triphosphate (dATP), di-dATP will be eventually incorporated (bottom right). This will terminate the elongation.

Nowadays, seminal variations of the original procedure have been developed for automated sequencing machines (for reviews, see Refs. [458, 459, 555, 724]). For example, in cycle sequencing the dideoxynucleotides are labeled with fluorescent dyes instead of the primers. Hence, all four reactions can be performed in the same tube, and the products can be loaded into the same lane of the gel. To differentiate the reaction products during electrophoresis, a laser is directed at a fixed position of the gel. As the products migrate past the laser, the laser causes the dyes to fluoresce, and the fluorescence is electronically recorded and interpreted with regard to the wavelength. The result is an electropherogram showing the fluorescence units over time and fragment position (see Figure 1.8).

An important prerequisite for sequencing a DNA segment is that the flanking sequence must be known for primers to be constructed. The two utilized primers have to be specific sequences that are unique in the genome. From a number of approximately 11 bases upward, a locus usually is sufficiently unambiguous in the human genome. Most error prone in this sequencing technique is the reading of the results of the electrophoresis. Hence, it is normally recommended that the output is read twice and additionally once from both directions. The technique is rather expensive and can be performed more efficiently if several loci are simultaneously sequenced. Here, in order to correctly read the result, it is important that they are unambiguously distinguishable with regard to their length. Several factors influence the success of the sequencing. For example, DNA with high *guanine-cytosine content* (*GC-content*) which are termed *GC-rich* regions can be difficult to sequence. GC-content is the percentage of guanine or cytosine bases on a DNA molecule. GC pairs are bound by three hydrogen bonds, while AT pairs are bound by two hydrogen bonds. DNA with high GC-content is more stable than DNA with low GC-content.

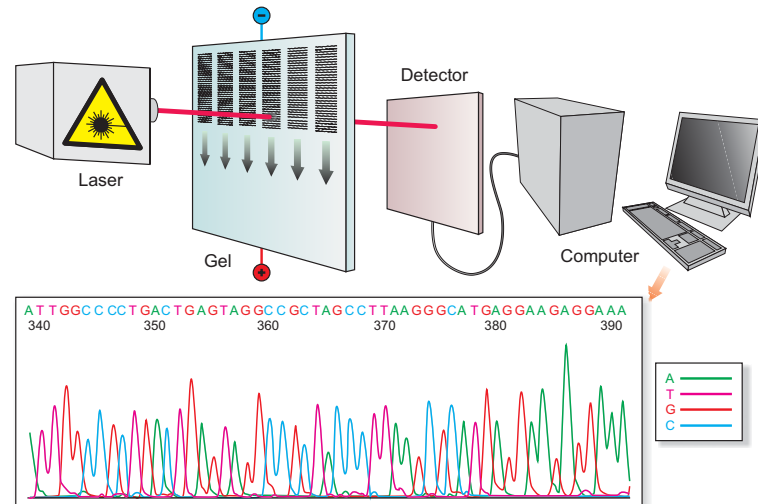


Fig. 1.8 An electropherogram depicting the results of an automated DNA sequencing using fluorescent primers. During electrophoresis, DNA fragments migrate in an electrically charged field, with their speed depending on their length: longer fragments are slower. A laser beam directed at a fixed position of the gel causes the dyes to fluoresce. The recorded fluorescence units over time, thus the position on the fragment, are given in the electropherogram. The example represents sequencing of clone *RP11-386/14* from chromosome 1. This figure was kindly provided by Jeanette Erdmann.

Therefore, the sequence will start out strong but the signal strength will rapidly decrease until there is no sequence data. As a consequence, read lengths are typically shorter for these templates. Problems can also arise for homopolymeric regions and in the presence of high repetitions.

1.3.2.2 Amplification of DNA segments Laboratory techniques such as sequencing require a certain amount of DNA that may not always be available. One possibility for amplifying DNA segments *in vitro* is the *polymerase chain reaction (PCR)*.

For the PCR, the following reagents are required:

- A template DNA segment containing the target sequence that is to be amplified.
- A set of DNA nucleotides to build the new DNA.
- Primers complementary to a short region on either side of the target sequence.
- A specific polymerase to catalyze the synthesis of the new DNA.

The PCR procedure is run in a number of cycles (see Figure 1.9, upper half). In each cycle, the following steps are performed:

1. Denaturation: At a high temperature, the template DNA is melted and thus denatured, i.e., the two strands are unzipped.

2. Hybridization: At a low temperature, the primers anneal to the complementary sequences at the sides of the target sequence.
3. Elongation: At a medium temperature, the new DNA is synthesized by the polymerase enzyme. The elongation of the two primers is performed in the direction of the amplified sequence.

As a result, in the first cycle both primers are elongated along the original template DNA until the end of the template. In the second cycle (see Figure 1.9, lower half), the primers additionally initiate synthesis on the products of the first cycle. Hence, half the strands can be elongated as before without a specified stop. The others terminate where the original primers end. In further replications, amplification is more and more restricted to the sequence between the sites defined by the primers, which is the target sequence.

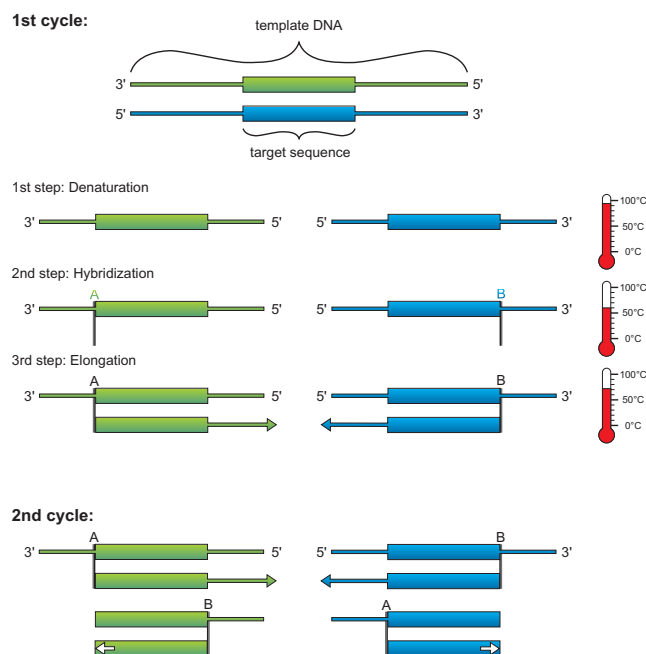


Fig. 1.9 Polymerase chain reaction (PCR) for DNA amplification. A DNA segment containing the target sequence (first from top) is denatured at high temperature so that the two strands become separated (second). Then, the primers hybridize to the ends of the target sequence at low temperature (third). At medium temperature, the primers are elongated unidirectionally (fourth). This cycle of high, low, and medium temperature is repeated (bottom). Importantly, the DNA segment of the previous cycle now serves as an additional template, the only difference being that it is restricted on one side by the primer. Hence, the following synthesized DNA segments will increasingly be restricted to the region between the two primers.

The number of DNA fragments that are thus created increases exponentially with the number of cycles. After 30 to 50 cycles, about 2^{30} to 2^{40} copies are produced.

As in the sequencing reaction, the two primers have to be unique sequences. However, their unambiguity being granted, the length is no longer important for the success of the PCR. What is more important is that their chemical features are similar; specifically, they should have comparable melting temperatures.

As polymerase, different enzymes are used. The most common is the Taq polymerase from the bacterium *Thermus aquaticus*. Because the bacterium lives in hot springs, its polymerase is adapted to high temperatures and is not destroyed during the melting phase of the PCR. Hence, it is not necessary to add new polymerase in each cycle, allowing the entire procedure to be automatized. However, because the Taq polymerase is often connected with a higher error rate, polymerase from different bacteria, for example, *Pyrococcus furiosus*, which has a better correction of errors, is used.

The advantages of the PCR are that the reaction is fast, highly specific, easily automated, and capable of amplifying small amounts of the DNA segment. As before, the sequence to be amplified has to be known in advance in order to construct primers.

1.3.3 How likely is it that individual differences are detected?

Although we now know how to amplify a variation, the last question still has to be answered: How probable is it to detect a variation that has a specific frequency in a given sample? Alternatively, one might be interested in calculating the sample size required to detect a variation with a pre-specified confidence. For simplicity, we first assume that all subjects are homozygous with one of the variants. Furthermore, we assume that the variation is the so-called single nucleotide polymorphism (SNP), i.e., that the variation has only two states (for more details on this, see Section 3.2). If the frequency of the less frequent variant is p and $1 - p$ for the more frequent variant, then the probability of observing only the rare variant is p^n and the probability of observing only the frequent variant is $1 - p^n$. Subsequently, the probability of observing only one variant ($1v$) in n subjects is

$$P(1v) = p^n + (1 - p)^n. \quad (1.1)$$

So far, we have assumed that all subjects are homozygous. If subjects can be heterozygous and if the frequency for homozygous and heterozygous subjects is in Hardy–Weinberg equilibrium (HWE; see Chapter 2), we do not need to consider subjects but can base our calculations on alleles. In this case, the probability of observing only one variant ($1v_{\text{HWE}}$) in n subjects with $2n$ alleles is

$$P(1v_{\text{HWE}}) = p^{2n} + (1 - p)^{2n}. \quad (1.2)$$

Equations (1.1) and (1.2) can be used to calculate the number of samples that need to be investigated until both variants are detected. To this end, $1 - P(1v)$ needs to be considered and solved for n given p . This is best achieved by an algorithm, and

several solutions including MINSAGE and the package GENETICS in R are available. These packages are able to do the necessary calculations in case of multiple variants as discussed in detail by Gregorius [263].

Table 1.1 Probability for observing both variants of a single nucleotide polymorphism (SNP) as a function of the frequency of the rarer variant (p) and the sample size (n). The third column gives the probability $P(1v)$ when only homozygous subjects are available, and the probabilities $P(1v_{\text{HWE}})$ reported in the fourth column are based on alleles, i.e., under the assumption of Hardy–Weinberg equilibrium.

p	n	$P(1v)$	$P(1v_{\text{HWE}})$	p	n	$P(1v)$	$P(1v_{\text{HWE}})$
0.5	10	0.9980	1.0000	0.01	100	0.6340	0.8660
0.4	10	0.9938	1.0000	0.01	200	0.8660	0.9820
0.3	10	0.9717	0.9992	0.01	500	0.9934	1.0000
0.3	20	0.9992	1.0000	0.005	10	0.0489	0.0954
0.2	10	0.8926	0.9885	0.005	20	0.0954	0.1817
0.2	20	0.9885	0.9999	0.005	30	0.1396	0.2597
0.2	30	0.9988	1.0000	0.005	50	0.2217	0.3942
0.1	10	0.6513	0.8784	0.005	100	0.3942	0.6330
0.1	20	0.8784	0.9852	0.005	200	0.6330	0.8653
0.1	30	0.9576	0.9982	0.005	500	0.9184	0.9933
0.1	50	0.9948	1.0000	0.005	1000	0.9933	1.0000
0.05	10	0.4013	0.6415	0.001	10	0.0100	0.0198
0.05	20	0.6415	0.8715	0.001	20	0.0198	0.0392
0.05	30	0.7854	0.9539	0.001	30	0.0296	0.0583
0.05	50	0.9231	0.9941	0.001	50	0.0488	0.0952
0.05	100	0.9941	1.0000	0.001	100	0.0952	0.1814
0.01	10	0.0956	0.1821	0.001	200	0.1814	0.3298
0.01	20	0.1821	0.3310	0.001	500	0.3936	0.6323
0.01	30	0.2603	0.4528	0.001	1000	0.6323	0.8648
0.01	50	0.3950	0.6340	0.001	2000	0.8648	0.9817

Table 1.1 displays the number of subjects, the frequency of the rarer variant p , the probability $P(1v)$ of not observing both variants when only homozygous subjects are available, and the probability $P(1v_{\text{HWE}})$ of not observing both variants under the assumption of HWE. If the frequency of both variants is high, only few subjects like 10 or 20 need to be screened. However, if the frequency of the rarer variant is as low as 1%, several hundreds need to be screened for observing both variants with a probability of at least 99%. If the frequency of the rare variant is as low as 0.001,

the probability for observing both variants is 0.8648 for a sample size as large as 1000 if HWE holds. This would represent the best case because it allows to consider the single alleles. When only homozygous subjects can be observed, the probability of not observing both variants is 0.6323, thus substantially higher. Although this situation is unrealistic in practice, the two scenarios considered here are the two extremes. While the assumption of HWE yields the lowest number of subjects that need to be screened for identifying both variants, the assumption of homozygosity of all subjects represents the worst case scenario and gives the highest numbers.

1.4 PROBLEMS

Problem 1.1

Problem 1.1.1. Answer the following questions:

1. What are the main constituents of a DNA molecule?
2. How are the two strands of one DNA molecule connected?
3. What are the main differences between DNA and RNA?

Problem 1.1.2. Define the following terms:

1. Nucleotide and nucleoside
2. Diploid and haploid
3. Gene, genotype, and phenotype
4. Introns, exons, and promoters

Problem 1.1.3. Consider the following base sequence of a given DNA segment:

T A C A A T G A T C T G A C G A T T

1. What is the sequence of the complementary DNA strand?
2. Given the original template strand, what would be the sequence of an RNA strand after transcription?
3. What would be the result of translation?

Problem 1.2

Problem 1.2.1. Describe the process of meiosis. Specifically, recall how many chromosomes each cell contains before meiosis, after each stage, and after meiosis.

Problem 1.2.2. Consider the following five chromosomal segments *A*, *B*, *C*, *D*, and *E* on two chromatids (Figure 1.10).

1. Between which segments would you observe recombinations if there is one crossing over between segments *B* and *C*?
2. Between which segments would you observe recombinations if there is one crossing over between segments *A* and *B* and a second between segments *C* and *D*?

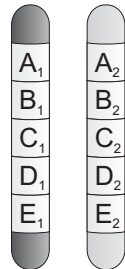


Fig. 1.10 Illustration of five chromosomal segments on two chromatids for Problem 1.2.2.

Problem 1.3

Problem 1.3.1. Define the following terms:

1. Locus and allele
2. Homozygous and heterozygous
3. Mutation and polymorphism

Problem 1.3.2. What would the resulting peptide look like if the following modifications of the original template strand from Problem 1.1.3 had occurred?

1. Insertion of the single base adenine at position 8, between adenine and thymine
2. Substitution of the single base thymine at position 9 with the base adenine
3. Deletion of bases adenine, thymine, and cytosine at positions 8, 9, and 10

Problem 1.3.3. Describe the following techniques:

1. Dideoxy sequencing
2. Polymerase chain reaction

Problem 1.4 What is the probability of not observing both alleles when a variant is expected to exist in 1 out of 10,000 subjects? Assume that 1000, 10,000, 20,000, and 500,000 subjects are available for screening.

URLs

DECIPHERING THE GENETIC CODE

<http://history.nih.gov/exhibits/nirenberg/>

DNA FROM THE BEGINNING

<http://www.dnafb.org/dnafb/>

GENETICS

<http://cran.r-project.org/web/packages/genetics/index.html>

MINSAGE

<http://www.imbs-luebeck.de/software/minsage.html>

HUMAN GENOME VARIATION SOCIETY

<http://www.dmd.nl/mutnomen.html>

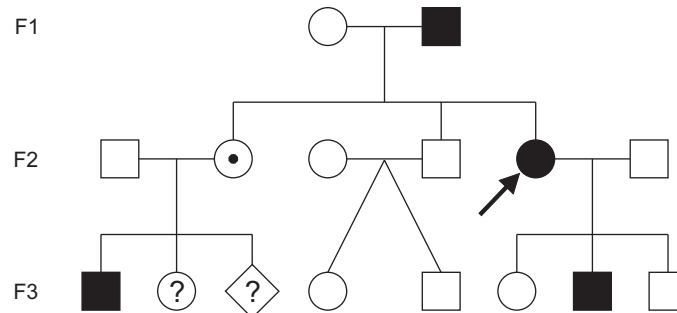
2

Formal Genetics

In the previous chapter, we described DNA and RNA as carriers of genetic information. However, before these structures were known, and even before chromosomes had been detected, some regularities had been discovered that still form the basis for formal genetics. Hence, in this chapter we will introduce some of these regularities and discuss their consequences.

In the subsequent sections, we will focus on Gregor Mendel and look at the laws he derived from the observations in his cloister garden (Section 2.1). On the basis of that, we will address the question of how, in general, phenotypes are inherited in families (Section 2.2) and which complications to these general patterns exist (Section 2.3). We will conclude with a description of another regularity that was independently detected by Godfrey Harold Hardy and Wilhelm Weinberg (Section 2.4).

To facilitate further description of pedigrees and the transmission of genetic factors, we introduce the conventional symbols for displaying pedigrees (Figure 2.1). More complex issues are shown in the literature [55]. In this exemplary pedigree, we depict how the affection with a certain disease is transmitted from generation to generation. The *index proband* is the individual via whom the pedigree has been ascertained. In Figure 2.1, the index proband is an affected female in generation F_2 . She has three offspring. The brother of the index proband has dizygotic twins. Her sister is a carrier of the mutation and has three offspring, but the information on two of these is fairly incomplete with the affection status of both and the gender of one being unknown.



F = generation; alternatively numbering of generations with Roman numerals.

○/□/◇ = unaffected female / male / unknown gender

●/■/◆ = affected female / male / unknown gender

⊙/⊠/◇ = female / male / unknown gender with carrier status

⊙/⊠/◇ = female / male / unknown gender with unknown affection status

○—□ = coupling, ○—○ = twins, ○—○ = monozygotic twins

∅/∅/∅ = deceased, ↗ = index proband

Fig. 2.1 Conventional symbols for displaying pedigrees.

2.1 WHAT ARE MENDEL'S LAWS?

Gregor Mendel was an Augustinian monk living in Brunn, Czech Republic, from 1822 to 1884. After studying natural sciences at the University of Vienna for two years, he performed large-scale cross-breeding experiments in the monastery gardens. Using pea pod plants or seeds, he observed the transmission of easy-to-distinguish traits such as blossom color or seed color. Through this, he developed some basic concepts on genetic information units he called “factors” that still remain valid today. Unfortunately for Mendel, his observations published in 1866 did not receive due attention until they were rediscovered in 1900 by Robert deVries. For detailed information on the history, the interested reader is referred to Ref. [647].

Based on his cross-breeding results and without knowing the physical basics of the genetic factors, Mendel derived regularities that were later restated as the following Mendelian laws:

1. Law of uniformity
2. Law of segregation
3. Law of independent assortment

The first and second laws describe the distribution of traits controlled by a single gene. Under ordinary circumstances, the first generation after crossing two different

homozygotes contains heterozygotes that look the same. For example, when investigating pea color, Mendel crossed homozygous yellow peas with homozygous green peas, and all of the offspring ended up being heterozygous yellow. However, crossing individuals from this generation led to different traits in the next generation. In this special case, both yellow and green peas emerged. To be specific, with a ratio of 1:2:1, the offspring was homozygous yellow, heterozygous yellow, and homozygous green.

The *law of independence* states that two genetic factors are transmitted independent of each other. Consider as an example the pedigree in Figure 2.2. Here, we have two genetic loci, the first showing the alleles 1, 2, 3, and 4, the second with *A*, *B*, *C*, and *D*. Both parents are heterozygous at each locus. Mendel's law now states that if the mother has alleles 1 and 2 at the first and *A* and *B* at the second locus, her children will inherit 1 with *A*, 1 with *B*, 2 with *A*, and 2 with *B* at about equal frequencies, as is shown in the second generation of the pedigree. The same holds for the alleles of the father.

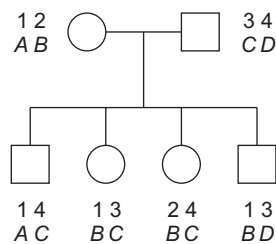


Fig. 2.2 Mendel's third law, the law of independent assortment. Two genetic loci are shown, the first having alleles 1 to 4, the second *A* to *D*. Both parents are heterozygous at each locus. Mendel's law states that the children will inherit parental alleles from both loci independently.

Now, is this a realistic assumption for human genetic loci? After describing the organization of genes in chromosomes in Chapter 1, we know that this is true only for genetic loci that are located on different chromosomes. In this situation, it is a matter of pure chance whether after meiosis they might end up in the same gamete or not. The situation is different for genes on the same chromosome. Here, we would expect them to be transmitted together, as they are physically on the same chromatid. However, because of the recombinations in crossing over events (see Section 1.2), genetic loci on the same chromatid might be separated and not inherited together. These deviations from Mendel's law of independent assortment will lay the foundation for genetic linkage analysis (see Chapter 7).

2.2 HOW ARE PHENOTYPES TRANSMITTED IN FAMILIES?

In the example in Section 2.1 on the law of uniformity, it was observed that all heterozygous individuals were of the same phenotype, being yellow, as one of the homozygous parents. This means that if in one individual both genetic factors for the yellow and the green color are present, the factor for the color yellow *dominates* the factor for the color green. Conversely, the factor determining the color green *recesses* compared to the one for the color yellow. It should be noted that dominant and recessive are two sides of the same coin. Fascinatingly, these observations made

in pea pod plants can be translated to inheritance patterns in humans. In the following section, we will describe the difference between dominant and recessive inheritance patterns. Furthermore, it will be distinguished whether the phenotype of interest is transmitted with an autosome or with a sex chromosome. Because most of the examples stem from clinical genetics, rather than focusing on a phenotype in general, we will look at a disease or an *affection status*.

For simplification, we will in this section describe only diseases that are determined by the genotype at a single locus. At this specific locus, we assume that there are two alleles: one causing the disease, later denoted by D , the other not, later termed the *normal* allele N . The locus is hence termed to be *diallelic*. For the time being, we will neglect all other possible influences on the etiology of the disease such as environmental or behavioral factors.

In the following sections, we will give examples of diseases following a specific mode of inheritance. More detailed information on these diseases and the disease-causing genes can be found in *OMIM*, the *Online Mendelian Inheritance in Man*. This is a catalog of human genes and genetic disorders, with links to literature references, sequence records, maps, and related databases. Originally, it was an online continuation of the book *Mendelian Inheritance in Man* [454] by Victor McKusick, and this online version is updated daily. More information on the online database OMIM and the book can be found in the frequently asked questions section of the web site.

2.2.1 What is autosomal dominant inheritance?

If a disease is transmitted in an autosomal dominant fashion (see Figure 2.3), this means that it is determined by a gene on one of the 22 autosomes. In addition, being dominant, one allele is sufficient to cause the affection. Hence, both individuals who are homozygous for the phenotype allele and r who are heterozygous will be affected. Examples of diseases that follow this pattern are Huntington's disease (OMIM +143100), different forms of non-syndromic deafness (OMIM #600965, #601544), familial hypercholesterolemia (OMIM #143890), neurofibromatosis (OMIM +162200), and myotonic dystrophy (OMIM #160900).

Several characteristics make it easy to recognize autosomal dominant inheritance in a given family:

- Both genders are affected at similar frequencies.
- The disease is visible in every generation of the family.
- If a person is affected, at least one of the parents is also affected, and this can be either the father or the mother.
- If an affected person mates with an unaffected person, about half the children will be affected.

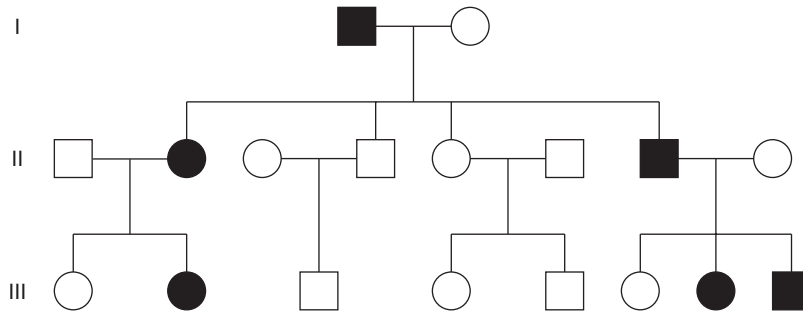


Fig. 2.3 Pedigree with autosomal dominant inheritance. The characteristics are that both genders are affected at similar frequencies; disease is visible in every generation; at least one parent of an affected individual is affected; about 50% of the offspring of an affected–unaffected couple are affected.

2.2.2 What is autosomal recessive inheritance?

Figure 2.4 shows an illustrative example of a disease transmitted in an autosomal recessive way. In contrast to a dominant phenotype, two disease-causing alleles are required for affection. Hence, although the proband in generation III has inherited one causative allele from her father, she does not show the disease. Still, she transmits this allele to both of her sons and one of her daughters. Because the first son also inherits this allele from his father, he is homozygous and hence affected. Important clinical examples of diseases with an autosomal recessive inheritance are spinal muscular atrophy (OMIM #253300), Friedreich's ataxia (OMIM #229300), cystic fibrosis (OMIM #219700), and phenylketonuria (OMIM +261600).

To distinguish an autosomal recessive phenotype, the following features might be present:

- Both genders are affected at similar frequencies.

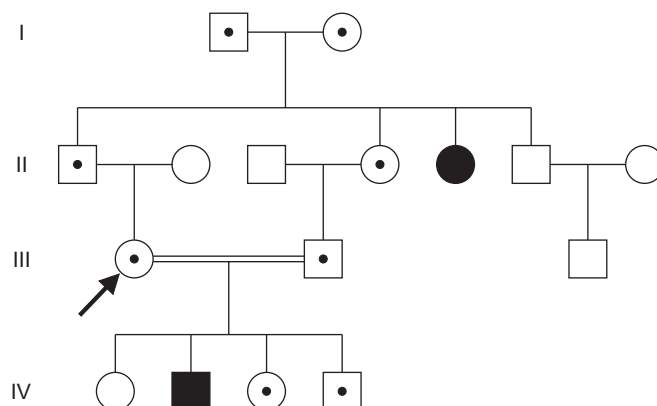


Fig. 2.4 Pedigree with autosomal recessive inheritance. The characteristics are that both genders are affected at similar frequencies; if an individual is affected, about $\frac{1}{4}$ of the siblings are affected; disease is not visible in every generation; parents of the affected are often interrelated.

- The disease is not visible in every generation of the family. Instead, the parents of a person with the disease usually do not carry it.
- If one person is affected, both parents are at least heterozygous for the disease-causing gene. They are hence termed to be *carriers* of the causative allele. As a consequence, about one-quarter of their children will be homozygous for this allele and hence affected.
- Often, the parents of the affected are interrelated. For example, the parents might be cousins.

2.2.3 What is X-chromosomal dominant inheritance?

For diseases inherited on one of the sex chromosomes, the special inheritance pattern of the sex chromosomes has to be kept in mind: that is, mothers always transmit one X chromosome to both their sons and their daughters. Fathers, in contrast, transmit their X chromosome to their daughters and their Y chromosome to their sons. Let us consider the case of a dominant disease caused by a gene on the X chromosome first (see Figure 2.5). Here, if a father carries the disease-causing allele, he transmits this to all of his daughters; hence, all of them will be affected. In contrast, none of his sons will be affected because they never inherit the paternal X chromosome. If, on the other hand, the mother carries one disease-causing allele and therefore is affected, she transmits this to half her children, regardless of their gender. Diseases transmitted in an X-chromosomal dominant pattern are rather rare; examples include the Coffin–Lowry syndrome (OMIM #303600), incontinentia pigmenti (OMIM #308300), and Charcot–Marie–Tooth neuropathy (OMIM #302800).

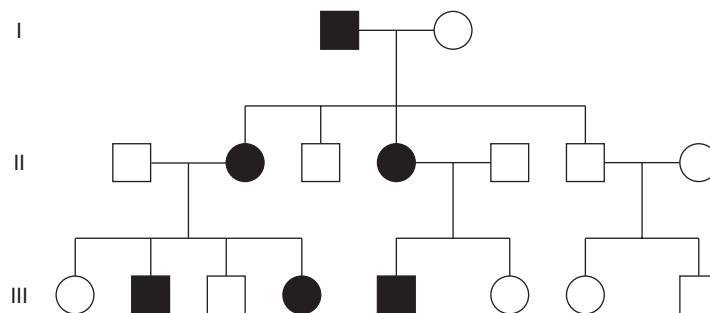


Fig. 2.5 Pedigree with X-chromosomal dominant inheritance. Main characteristics are that the disease is prevalent in every generation; if a father is affected, all of his daughters but none of his sons are affected; if a mother is affected, about half of their children are affected, regardless of their gender.

The following characteristics help us to recognize X-chromosomal dominant inheritance of a disease in a given family:

- Although both sexes are affected, it is more frequent in females.
- As with autosomal dominant transmission, the disease is prevalent in every generation of the family.
- If a father is affected, all of his daughters, but none of his sons, will be affected. In contrast, if a mother is affected, about half her children will be affected, regardless of their gender.

2.2.4 What is X-chromosomal recessive inheritance?

Much more frequent than X-linked dominant diseases are X-linked recessive diseases. The most obvious feature of these diseases is that they are almost exclusively shown by males. This is due to the fact that for loci on the X chromosome, males are *hemizygous*, meaning they have only one allele. Hence, there is no difference between dominant or recessive inheritance, as the single allele determines the expression of the disease. As a consequence, a male can be affected if his mother is a carrier or homozygous. In contrast, a female needs to be homozygous and inherit two alleles to be affected by the disease. This means that her father must be affected and her mother must at least be a carrier (see Figure 2.6). Clinical examples of this inheritance pattern include Duchenne muscular dystrophy (OMIM #310200), hemophilia A (OMIM +306700), red-green color blindness (OMIM +303900, +303800), and Hunter disease (OMIM +309900).

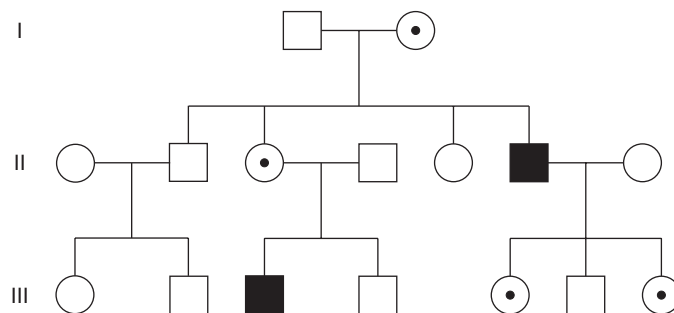


Fig. 2.6 Pedigree with X-chromosomal recessive inheritance. Main characteristics are that almost exclusively males are affected; parents of an affected male are usually unaffected; if a male is affected, about half his brothers are affected.

X-chromosomal recessive inheritance of a disease in a given family can be recognized by the following:

- Almost exclusively males are affected.
- The disease is not prevalent in every generation of the family. Instead, parents of an affected male usually are unaffected.
- If a male is affected, his mother is a carrier. Conversely, about half her sons will be affected and about half her daughters will be carriers.

- If a female is affected, this means that in addition to her mother being a carrier, her father is affected as well. He, in turn, transmits the allele to all of his daughters but none of his sons.

2.2.5 What is Y-chromosomal inheritance?

For human Y-chromosomal inheritance, no distinction between dominant and recessive can be made, as only males inherit a Y chromosome and are hemizygous. The Y chromosome carries only very few genes, and only a few phenotypes are described that follow a Y-chromosomal pattern. The distinctive features are (see Figure 2.7) as follows:

- Only males are affected.
- The disease is visible in every generation of the family.
- If a person is affected, so is the father.
- If a person is affected, all of his sons are also affected.

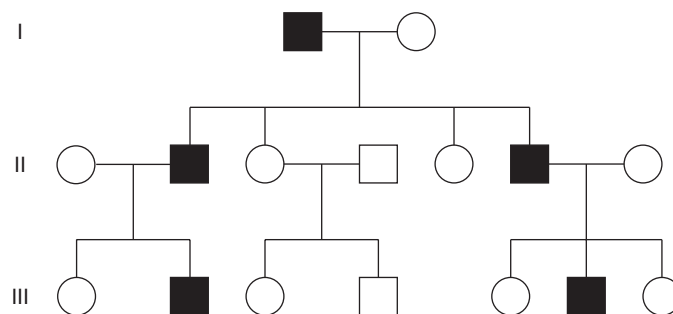


Fig. 2.7 Pedigree with Y-chromosomal inheritance. Main characteristics are that only males are affected; disease is visible in every generation; the father of an affected is affected; all sons of an affected individual are affected.

2.3 WHICH COMPLICATIONS TO THE GENERAL INHERITANCE PATTERNS EXIST?

In the previous section, we focused on diseases with quite simple inheritance patterns that can be easily recognized if the pedigree is large enough. But do all phenotypes that are determined genetically follow these straightforward patterns? The answer, unfortunately, is no. Although for these the genetic contribution is most obvious, deviations are the rule that concern the following issues:

1. Even though a person carries the disease-causing genotype, he might not become affected. Conversely, even though a person does not carry the disease-causing genotype, she might become affected. Less extremely, even though

different members of the family might be affected, the appearance of the disease might differ. This will be described in detail in Sections 2.3.1 and 2.3.2.

2. The affection with or expression of the disease might depend on whether the responsible allele was inherited from the mother or father (Section 2.3.3).
3. The same disease might be caused by different genes or different alleles in different families. In addition, the disease might not be due to only one genetic locus, which would be termed *monogenic* or *unilocal*. Instead, it might be determined by several, being called *oligogenic* or *paucilocal*, or even many loci, termed *polygenic* or *multilocal*. Furthermore, it may also be caused by environmental influences. These issues will be discussed in Section 2.3.4.

2.3.1 What is the penetrance function?

Let us return to Figure 2.5. In this family with a disease that follows an X-chromosomal dominant inheritance pattern, it was observed that the grandfather from generation I and all of his daughters are affected. If, however, one of his daughters in generation II were unaffected, this might be due to reduced *penetrance*. Formally, penetrance is the conditional probability $P(x|g)$ of being affected with disease x given a specific genotype g . If we assume a disease-causing genotype with one disease-causing allele D and one allele not causing the disease N , the probabilities of being affected depending on the genotypes can be expressed as follows:

$$f_0 = P(\text{affected}|NN) \quad f_1 = P(\text{affected}|DN) \quad f_2 = P(\text{affected}|DD) \quad (2.1)$$

Specifically, f_0 is termed the frequency of *phenocopies*, indicating individuals who are affected without carrying a disease-causing allele.

How can we fit the previously described inheritance patterns into these equations? The answer is shown in Table 2.1. Here, the phenotypes show full penetrance and no phenocopies. This means that every individual with and no individual without the disease-causing genotype will become affected.

Table 2.1 Penetrances for simple Mendelian inheritance patterns.

Genotype	Genetic model		
	General	Recessive	Dominant
NN	f_0	0	0
ND	f_1	0	1
DD	f_2	1	1

In addition to the recessive and dominant model, *co-dominant* inheritance is described for some phenotypes. The specific feature of this is that the phenotype of a heterozygote is clearly distinguishable from either homozygote. As an example, consider the ABO blood group system presented in Table 2.2.

In this system, three different alleles determine four different phenotypes. Although alleles A and B clearly dominate allele 0 , individuals being heterozygous AB display a phenotype that is different from all others; hence, alleles A and B are co-dominant.

Let us now part from the assumptions of full penetrance and no phenocopies. If phenocopies occur and $f_0 > 0$, the above functions f_0 , f_1 , and f_2 can be used to formulate *genotypic relative risks* (GRR):

$$\begin{aligned}\gamma_1 = GRR_1 &= \frac{f_1}{f_0} = \frac{P(x|DN)}{P(x|NN)} \\ \gamma_2 = GRR_2 &= \frac{f_2}{f_0} = \frac{P(x|DD)}{P(x|NN)}\end{aligned}\quad (2.2)$$

Hence, the GRRs denote the increased risk that a person with the specific genotype is affected over a person with no disease-causing alleles. Through this, the three parameters f_0 , f_1 , and f_2 are now reduced to the two parameters γ_1 and γ_2 . By taking into account the relationships for specific modes of inheritance, another parameter is rendered unnecessary, as Table 2.3 shows.

In addition to the simple models of recessive and dominant inheritance, other models are conceivable. In the right-hand columns of Table 2.3, additive and multiplicative genetic models are included. If a disease is transmitted in an additive fashion, this means that the risk for a heterozygous person to be affected is half that of a person who is homozygous D compared to an individual who is homozygous N , and thus $f_1 = (f_0 + f_2)/2$. This implies $\gamma_2 = 2\gamma_1 - 1$. In an analogous way, a multiplicative mode of inheritance can be defined [565].

The specification of GRRs in these ways has different uses in genetic epidemiological projects. For one, the values can be used to characterize the familial transmission of a given disease. The resulting values will be important, for instance, in the genetic counseling situation. Second, GRRs present an estimation of the effect size in genetic epidemiological studies and are therefore used for sample size calculations (see, e.g., Section 11.3).

Table 2.2 Penetrances in the ABO blood group system.

Genotype	Phenotype			
	A	B	AB	0
AA	1	0	0	0
A0	1	0	0	0
AB	0	0	1	0
BB	0	1	0	0
B0	0	1	0	0
00	0	0	0	1

Table 2.3 Genotypic relative risks under the assumption of phenocopies.

Genotype	GRR	Genetic model			
		Recessive	Dominant	Additive	Multiplicative
DD	γ_2	γ	γ	$2\gamma - 1$	γ^2
DN	γ_1	1	γ	γ	γ
Restriction		$\gamma_1 = 1$	$\gamma_1 = \gamma_2$	$\gamma_2 = 2\gamma_1 - 1$	$\gamma_2 = \gamma_1^2$

Example 2.1. *Retinoblastoma is a rare form of cancer, with a prevalence of roughly 1 in 20,000 persons. As the proportion of sporadic cases is very high, the frequency of phenocopies f_0 can be assumed to equal about $\frac{1}{20000} = 0.00005$. It has been shown that in familial cases, the disease follows an autosomal dominant inheritance*

pattern, hence $f_1 = f_2$ and $\gamma_1 = \gamma_2$. However, penetrance is reduced, with about 85% of individuals carrying the disease-causing allele developing retinoblastoma. Therefore, $f_1 = f_2 = 0.85$. From this, we can calculate the GRRs with $\gamma_1 = \gamma_2 = \gamma = 0.85/0.00005 = 17,000$.

A term that should not be confused with variable penetrance is the possible *variable expressivity* of a disease. The former refers to the probability that the disease will be manifested at all, whereas the latter indicates variations in the degree of manifestation. A clinical example of this is neurofibromatosis (OMIM +162200), where the expression in affected patients might vary from merely showing spots on the skin, the so-called café-au-lait spots, to a systemic involvement and tumors of the skin.

A special case of variable expressivity is *anticipation*. This term describes the phenomenon that the severity of the expression becomes greater and manifests earlier in successive generations. However, it has long been debated whether anticipation is an observational artifact. This was fueled by several test statistics that had been proposed to test for age at onset anticipation [319, 671, 677]. (Age of onset will be discussed in the next section in more detail.) Most of these utilize *affected parent–affected child* (APAC) pairs, which are oversampled in most applications because pedigrees are preferentially selected for the presence of multiple affected individuals. This sampling procedure can lead to a substantial increase in type I error fractions, causing interpretation of test results to become problematic; therefore, the results of age-at-onset testing is not meaningful in most APAC studies [677].

Nevertheless, the occurrence of anticipation, especially with regard to disease severity, is not doubted anymore. For example, in myotonic dystrophy (OMIM #160900) it has been discovered that the genetic defect in one form of the disorder is an amplified trinucleotide repeat in the 3' untranslated region of a protein kinase gene on chromosome 19. This repeat explains many of the unusual features of the disorder. For example, severity varies with the number of repeats: normal individuals have from five to 30 repeat copies; mildly affected persons, from 50 to 80; and severely affected individuals, 2,000 or more copies. A similar explanation has been found for myotonic dystrophy 2, also known as proximal myotonic myopathy (PROMM), which is caused by expansion of a CCTG repeat in intron 1 of the zinc finger protein-9 gene on chromosome 3q13.3-q24 (OMIM #602668).

2.3.2 What is the role of age-dependent penetrance?

In the previous section we described the concept of variable penetrance rather generally. In that context, the probability of affection was assumed to be similar for each individual carrying a specific genotype. Let us now consider diseases that are not manifested at birth but develop only later in life. A prominent example of this is Huntington's disease. Although its inheritance follows an autosomal dominant pattern, some patients become affected in the first years of life, whereas others not until after they have reached the age of 60 years or more. Therefore, the penetrance depends on the age of the individual: not all individuals carrying the disease-causing

genotype are affected, but the more older they are, the higher the probability for affection. Hence, the disease shows a variable age of onset, and there are many reasons to have a closer look at this phenomenon (for a review, see Ref. [117]).

The first quite obvious area where age of disease onset plays a role is genetic counseling. Here, risks need to be estimated in different individuals. These risks can be lifetime manifestation risks, which may depend on specific genotypes. For example, the risk of lifetime manifestation of breast cancer is 40–80% if the individual has a mutation in the *BRCA1/BRCA2* genes, and it is 100% for Huntington's disease if an individual has a mutation in the *huntingtin gene (HD)*. Apart from directly or indirectly estimating lifetime manifestation risks, it can also be of interest to estimate the age-of-onset distribution among the individuals inheriting a specific genotype. Put differently, it is of interest not only whether someone develops the disease at all but also when. For different genotypes, this corresponds to estimating genotype-specific hazard functions in survival analysis.

The second area of application is the differentiation of etiologies of the same phenotype. Here, the specific disease-causing variant is generally unknown, and other covariates such as positive family history and age may contribute to the individual genetic risk. In this case, age of onset may also be used to distinguish different etiologies of the same phenotype. For example, it has been shown for breast cancer that in the general population there are a small number of genetic cases carrying a susceptibility genotype that is very rare. The vast majority, however, are probably non-genetic cases [125]. The genetic risk here is a clear function of age, with the age-specific relative risks being about 100 for ages younger than 30 years and dropping down to two at ages 80 years and older. Similarly, in families with a proband suffering from prostate cancer, an early age of disease onset and having multiple affected family members predict a higher risk of prostate cancer in the family [107]. Again, this obviously genetic type accounts for only a small proportion of prostate cancer cases in the general population. Hence, knowledge of age of onset in these examples helps us to distinguish between two subforms of breast and prostate cancer as well as diseases such as diabetes mellitus [384], Alzheimer's disease [464], and alcoholism [134]. Interestingly, studies on Alzheimer's disease and Parkinson's disease revealed that although both are probably caused by unique major genes, another common genetic variant might be responsible for the age of onset in both the diseases [412].

Because age-of-onset information can be used to distinguish different etiologies, incorporation of this information in linkage and/or association analysis may be helpful in identifying disease genes. For example, in a family study on breast cancer, a significant linkage was found on chromosome 17 in early-onset families only. In families with late-onset disease, this locus was excluded [276]. Subsequent analyses have led to the identification of the *BRCA1* gene.

By including information on age of disease onset in further studies, two applications can be distinguished. Classically, the aim is to identify a genetic variant that determines affection with the disease. If variable age of onset is not considered, the resulting test statistics can be biased in both directions. Alternatively, instead of thus looking for liability genes, interest has recently shifted to directly detecting genetic

variants responsible for age of onset. In the latter case, the onset age is the dependent variable. It is important to note that genes controlling the age of onset may or may not be the same as those controlling an individual's liability.

Although the focus of this last area is not the age distribution itself, some methods still rely on the precise estimation of it in order to be properly accounted for. Therefore, we need to concern ourselves with the assessment of age-of-onset distributions, regardless of for which of the above reasons we are interested in it.

The first group of methods estimates age-of-onset distributions in a sample of independently recruited probands. An advantage of this is that, in contrast to other approaches described below, no genetic mode of inheritance has to be assumed. Still, appropriating onset data from independently recruited probands has to deal with several problems that are well known from observational studies. A typical problem is caused by lead-bias, meaning that the observed ages of onset are biased toward younger ages, as cases are obviously not recruited before their onset. It does not suffice to increase the proportion of older individuals in the sample; instead, explicit assumptions about the age distribution in the susceptible population have to be incorporated; these are usually based on the general population or on relatives of the probands. The first is appropriate if the respective disease has no effect on either mortality or fertility, while the latter assumes that there is no correlation between the ages at onset among siblings and that the mortality pattern in susceptible and non-susceptible siblings is the same.

The second group of methods to assess the age-of-onset distribution is based on pedigree data and requires data from the relatives of independently ascertained probands. Here, unaffected individuals are "censored," the estimation has to adjust for censoring among relatives, and ascertainment bias in the data must be taken into account. As an example, Cupples [149] adapted Kaplan–Meier methods to simultaneously estimate the recurrence risk and age-of-onset distribution in samples consisting of children and/or siblings of affected individuals. An overview on estimation methods has been given by Siegmund [615].

To summarize, we have seen that the penetrance of a disease-causing genotype might depend on a variety of personal factors, the most important being age. Knowing the age-of-onset distribution is crucial in genetic counseling, but this knowledge must be incorporated in further analysis in genetic epidemiological studies to reduce genetic heterogeneity. This also has important implications for the recruitment scheme of studies aiming at detecting disease-causing genetic variants. For instance, after obtaining age-specific genetic relative risks for lung cancer of more than 100 for individuals younger than 60 years as opposed to a risk of 1.6 by age 80 years or older, Gauderman [237] recommended that preferentially younger cases should be recruited because their disease is more likely to be due to genetic factors. More recently, age of onset has emerged to be an interesting target variable itself.

2.3.3 What are parental effects?

A phenomenon that is still poorly understood is the effect of *imprinting* or the *parent of origin effect*. This refers to the phenomenon that for some genetic loci,

the expression depends on who inherited the allele, the father or the mother. As an important example, both Prader–Willi and Angelman syndromes are caused by mutations in a gene on chromosome 15q11–q13. However, if the mutations are inherited from the father, the child will be affected with Prader–Willi syndrome (OMIM #176270); if they come from the mother, it will be Angelman syndrome (OMIM #105830).

Formally, maternal imprinting means that heterozygous individuals will be affected with the disease only if they inherited the causal mutation from father. Hence, maternal imprinting is synonymous to paternal expression, and this is the case for Prader–Willi syndrome. Analogously, paternal imprinting or maternal expression describes the situation in which a heterozygous child will become affected only if the allele is inherited from the mother for which the Angelman syndrome is an example. Our experience shows that it is difficult to remember the meaning of imprinting, as paternal imprinting means that the maternal allele is expressed. Our mnemonic device here is that imprinting means “suppression.” Thus, paternal imprinting means that the paternal allele is suppressed and, subsequently, the maternal allele is expressed.

Other diseases besides the classical Prader–Willi and Angelman syndromes for which imprinting has been shown to play a role are the rare Beckwith–Wiedemann syndrome (OMIM #130650) and common diseases such as atopy. As described above, an imprinting effect has repeatedly been described for Huntington’s disease in which late-onset cases have inherited the mutation preferentially from their mother and early-onset cases from their father. Although probably more than 1% of all mammalian genetic loci are imprinted, for most diseases the effect of imprinting has never been thoroughly investigated. A catalog of imprinted genes and parent-of-origin effects in humans and animals is given in the online *Imprinted Gene Catalogue*.

The underlying cause for this effect is that the maternal and paternal genomes influence the development of the embryo in different ways. To be specific, the imprinting is established during the development of sperms or eggs by methylation (see also Section 1.1.3). This process is then maintained during cell division and segregation of chromosomes. Although in somatic cells the imprinting will be maintained throughout the life cycle, in the germ cells the imprints will be erased by a demethylation process (for a detailed recent review, see Ref. [558]).

Imprinting should be distinguished from other *parental effects*, and the latter might be non-genetic. For example, homozygosity of the T allele of the C677T polymorphism of the gene encoding the folate-dependent enzyme 5,10-Methylenetetrahydrofolate reductase (*MTHFR*) is a risk factor for neural tube defects [78]. Maternal C677T homozygosity also appears to be a moderate risk factor. Furthermore, maternal periconceptional use of vitamin supplements containing folic acid substantially reduces the risk of neural tube defects (NTDs) in the offspring. This formed the basis for a discussion on the possible interaction between the C677T genotype and maternal folic acid intake in the occurrence of spina bifida, and the role of paternal imprinting [609].

If it is assumed that in a given situation imprinting plays a role, how can this be dealt with? For the definition of penetrances and genotypic relative risks, it is clear that in diseases with imprinting effects the penetrance in heterozygous individuals

depends on whom the allele was inherited from. Therefore, it does not suffice to define three different penetrances f_0 , f_1 , and f_2 as in Section 2.3.1. Instead, we now need two penetrances for heterozygous individuals where the parent is distinguished (see Ref. [643]), resulting in the following four penetrances:

$$\begin{aligned} f_0 &= f_{NN} = P(\text{affected}|NN) \\ f_{1P} &= f_{DN} = P(\text{affected}|DN \text{ with } D \text{ transmitted paternally}) \\ f_{1M} &= f_{ND} = P(\text{affected}|DN \text{ with } D \text{ transmitted maternally}) \\ f_2 &= f_{DD} = P(\text{affected}|DD) \end{aligned}$$

From these, GRRs can be derived as above (Section 2.3.1), and these functions can subsequently be used for further calculations and analyses.

2.3.4 What are other sources of heterogeneity?

2.3.4.1 Heterogeneity The previous sections gave a flavor of the heterogeneity involved in the study of genetic diseases where complications are the rule rather than the exception. For example, we have seen that the same disease might onset at different ages. In addition, we have to acknowledge that the same disease might show diverse characteristics. Even when focusing on familial Alzheimer's disease, for instance, families can be distinguished with regard to not only age of onset but also other clinical features such as occurrence of neurofibrillary tangles or amyloid plaques, or simultaneous evidence for anterior horn cell disease [60]. This is an illustration of *phenotypic heterogeneity*, where the same disease shows different features in different families or subgroups of patients. As with Alzheimer's disease, it is often assumed that the cause of this heterogeneity is that different genetic factors are involved. A related but different effect is *pleiotropy*. Here, the same known mutation causes symptoms or even multiple unrelated effects.

In contrast, *genetic heterogeneity* refers to the phenomenon of different mutations causing the same disease in different subgroups or families. Hence, the characteristics of the phenotype may be identical, but the cause is not. For the selection of an appropriate study design, it will be important to further differentiate between *allelic* and *locus heterogeneity*. The first means that there may be different disease-causing alleles at the same genetic locus. For example, different alleles have been detected at the major histocompatibility complex (MHC) locus on chromosome 6p21 that all play important roles in susceptibility to multiple sclerosis. The latter, in contrast, describes the situation when mutations at different loci in either the same or in different genes cause the same phenotype. For this, the different breast cancer genes *BRCA1* and *BRCA2* are classic examples.

2.3.4.2 Polygenic influences Although allowing a wide variety of heterogeneity, we have so far assumed that for a single individual, only one genetic variant is the cause of the disease. We have thus limited our description to *monogenic* diseases. However, this is often not the case. In contrast, it is more probable that the complex phenotype requires the simultaneous presence of functional variants at

a large number of different genetic loci, a condition termed *polygenic* inheritance. From a more theoretical point of view, polygenic inheritance is related to an infinite number of loci, thus representing the limit of a mixture of discrete distributions. This may be differentiated from *oligogenic* inheritance, where more than one but a few loci are involved simultaneously. In the easiest of these intricate constellations, two loci are involved that can act interactively in a number of different ways [561]. In the model of genetic heterogeneity described in the previous section, different individuals have different mutations, but one of them suffices for affection. In a multiplicative model, the presence of mutations at both loci is needed for the disease, and an additive model describes the situation where the effect at both loci is the sum of both single effects. More situations are conceivable, and they all lead to the definition of more complex penetrance functions.

Although sometimes being defined very differently, the effect of *epistasis* can be described within a multiplicative two-locus model. Mathematically, it refers to an interaction between different loci, and an example taken from Risch [561] is presented in Table 2.4. Here, two diallelic genetic loci *A* and *B*, each with the alleles *D* and *N*, are considered. *D* is used to denote the dominant disease-causing allele, and *N* is the allele that is not causative.

The penetrances shown exemplify that the effect of locus *B* can be observed only when the allele D_A at locus *A* is present, but not otherwise. The depicted situation is symmetric; hence, the effect of *A* is visible only in individuals carrying the allele D_B at locus *B*. As before, these more complicated penetrance functions can be utilized for further analyses and estimations of GRRs. In principle, similar relationships could be defined for more than two genetic loci. In practice, however, in phenotypes assumed to be determined by many loci, these are analyzed separately.

Table 2.4 Example for penetrances at two loci acting epistatically.

Locus <i>B</i>	Locus <i>A</i>		
	$D_A D_A$	$D_A N_A$	$N_A N_A$
$D_B D_B$	1	1	0
$D_B N_B$	1	1	0
$N_B N_B$	0	0	0

2.3.5 What is the difference between simple and complex diseases?

In the previous sections, we introduced a variety of possible complications to a simple inheritance pattern. All of those included merely genetic effects, although, in practice, environmental influences often play a major role as well. Based on all those complications, a distinction is often made between *simple* and *complex* diseases, with simple diseases following a straightforward Mendelian inheritance pattern, as described in Section 2.2. In contrast, complex diseases are caused by multiple genetic and environmental factors, as delineated in previous sections. However, this differentiation is somewhat arbitrary because there is no clear transition from one to the other. Rather, each case presents more or fewer complications to the simple inheritance scheme. Nevertheless, it is important to reflect upon the nature of the disease one is interested in. If there is only a single factor that causes the disease, it should have a rather large effect and be relatively easy to unveil. If, on the other hand,

there are many factors contributing to the affection, the effect of each single one might be only very small. Hence, large sample sizes would be required to reliably detect these effects (see, e.g., Section 8.5). Consequently, after first successes in discovering genes underlying simple genetic diseases, the search for factors underlying complex diseases has been of less avail.

Because most common diseases are complex, we need to pose the following questions: Can their genetic cause be detected with a feasible effort? And how can the chances for success be increased? Whether a genetic effect is detected clearly hinges upon the power of the study. This, in turn, is closely connected to the frequency of the disease-causing variant in the population. In this context, the *common disease—common variant (CDCV)* hypothesis has been formulated. It states that even though a common disease is usually determined by many loci, most of the genetic risk is due to loci with merely one common variant. For example, it was found that a common polymorphism in the peroxisome proliferator-activated receptor- γ gene was associated with a decreased risk of type 2 diabetes [22]. Furthermore, it is assumed that although there are more rare than common polymorphisms in the human genome, the majority of the variant alleles stem from common variants. If this hypothesis were true, it would imply that genetic effects might be detected with a reasonable effort. On the other hand, it has been claimed that common variants have been more frequently detected simply because it is easier to find them. However, the body of available empirical evidence is still small, and simulation results have been controversial (a more thorough discussion on this topic can be found in Ref. [543]).

To improve the prospect of success, a reasonable strategy for studies is to try to simplify the disease under study. Specifically, the complexity of a disease might be reduced by selecting more homogeneous subgroups of patients. An obvious selection, as discussed in Section 2.3.2, is based on age of disease onset. An impressive example of this is early-onset familial breast cancer, where the genetic effect of a locus on chromosome 17q21 could be observed only when families with a young mean age of onset were considered [276]. Other selections of homogenous subsets might be founded on the specific characteristics of the phenotype, such as disease severity. As an example, when children with more extreme scores for reading disability were selected, genetic effects for dyslexia were verified for loci on chromosome 6p21 [158].

Similarly, it might be more promising not to search for causes of the final disease status itself but rather to analyze intermediate phenotypes that may be less subject to genetic or environmental impact. Put simply, the central dogma of molecular biology (see Section 1.1.3) can be modified to

DNA makes RNA, RNA makes proteins, proteins make intermediate phenotypes, intermediate phenotypes make diseases.

For instance, the investigation of the intermediate phenotype of the angiotensin I-converting enzyme level has led to the identification of a polymorphism that is a causative factor in myocardial infarction [622]. Similar strategies have been suggested for other complex diseases, such as eating disorders, but not proven successful [13].

To summarize, in the early years of genetic epidemiology, successes in the identification of genes underlying simple Mendelian characters and complex diseases, such as *BRCA1/BRCA2* for breast cancer, have created an overall optimistic atmosphere. The following years saw a focus on studies of complex diseases with a comparative lack of positive results in the 1990s. However, strategies as delineated above in combination with further technological developments in the area of genotyping throughput and novel genetic markers (see Chapter 3) have again turned the tide to successful identification of genes underlying complex diseases.

2.4 WHAT IS THE LAW DETECTED BY HARDY AND WEINBERG?

As a conclusion to this chapter on formal genetics, we will part from the segregation of phenotypes in families and turn to a fundamental principle from the realm of population genetics. This principle describes the relationship between allele and genotype frequencies in a population and was detected independently by Godfrey Harold Hardy and Wilhelm Weinberg in 1908; hence, it is usually known as the *Hardy–Weinberg law (HWL)*. It has proven useful in a number of ways, and we will show its possible applications in genetic counseling (see Example 2.2) and will discuss it for checking data quality in genetic epidemiological studies (see Section 4.3) as well as for testing genetic association (see Chapter 10).

The general statement of the *Hardy–Weinberg equilibrium (HWE)* is that under certain assumptions, the genotype and allele frequencies in a large, randomly mating population remain stable over generations and that there is a fixed relationship between allele and genotype frequencies. To understand this, consider a diallelic autosomal locus with alleles A_1 and A_2 . The possible genotypes at this locus then are A_1A_1 , A_1A_2 , and A_2A_2 , and we define the genotype frequencies in the population

$$P(A_1A_1) = p_{11}, \quad P(A_1A_2) = p_{12}, \quad P(A_2A_2) = p_{22}$$

with $p_{11} + p_{12} + p_{22} = 1$. From these genotype frequencies, we can calculate the allele frequencies $P(A_1) = p$ and $P(A_2) = q$ by

$$P(A_1) = p = p_{11} + \frac{1}{2}p_{12}, \quad P(A_2) = q = p_{22} + \frac{1}{2}p_{12}$$

with $p + q = 1$.

Under very general assumptions that we will elucidate below, we now consider matings from within this population. Nine mating types, as shown in Table 2.5, are possible. Assuming that the parental genotypes are independent, the frequencies of these matings are shown together with the probabilities for all possible genotypes in the offspring.

Combining the frequencies of the parental mating types with the probabilities for the offspring's genotypes, the frequencies for the different genotypes occurring in the offspring can be calculated by

$$\begin{aligned}
 P(A_1A_1) &= 1 \cdot p_{11}^2 + \frac{1}{2} \cdot p_{11}p_{12} + \frac{1}{2} \cdot p_{11}p_{12} + \frac{1}{4} \cdot p_{12}^2 = (p_{11} + \frac{1}{2}p_{12})^2 \\
 &= p^2, \\
 P(A_1A_2) &= \frac{1}{2} \cdot p_{11}p_{12} + 1 \cdot p_{11}p_{22} + \frac{1}{2} \cdot p_{11}p_{12} + \frac{1}{2} \cdot p_{12}^2 + \frac{1}{2} \cdot p_{12}p_{22} \\
 &\quad + 1 \cdot p_{11}p_{22} + \frac{1}{2} \cdot p_{12}p_{22} \\
 &= p_{11}p_{12} + 2p_{11}p_{22} + \frac{1}{2}p_{12}^2 + p_{12}p_{22} \\
 &= p_{11}(p_{12} + 2p_{22}) + p_{12}(\frac{1}{2}p_{12} + p_{22}) \\
 &= 2pq, \\
 P(A_2A_2) &= \frac{1}{4} \cdot p_{12}^2 + \frac{1}{2} \cdot p_{12}p_{22} + \frac{1}{2} \cdot p_{12}p_{22} + 1 \cdot p_{22}^2 = (\frac{1}{2}p_{12} + p_{22})^2 \\
 &= q^2.
 \end{aligned}$$

Table 2.5 Mating frequencies and offspring genotype probabilities given parental genotypes.

Mating		Frequency	Offspring genotype		
Father	Mother		A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	p_{11}^2	1	0	0
	A_1A_2	$p_{11}p_{12}$	$\frac{1}{2}$	$\frac{1}{2}$	0
	A_2A_2	$p_{11}p_{22}$	0	1	0
A_1A_2	A_1A_1	$p_{11}p_{12}$	$\frac{1}{2}$	$\frac{1}{2}$	0
	A_1A_2	p_{12}^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	A_2A_2	$p_{12}p_{22}$	0	$\frac{1}{2}$	$\frac{1}{2}$
A_2A_2	A_1A_1	$p_{11}p_{22}$	0	1	0
	A_1A_2	$p_{12}p_{22}$	0	$\frac{1}{2}$	$\frac{1}{2}$
	A_2A_2	p_{22}^2	0	0	1

Therefore, the genotype frequencies in the offspring can be calculated directly from the allele frequencies in the general population from which the parents came. This simple relationship between genotype and allele frequencies is the core statement of the HWL.

In addition, the allele frequencies in the offspring can be deduced as before and will lead to the same values as in the parental generation:

$$\begin{aligned}
 P(A_1) &= P(A_1A_1) + \frac{1}{2}P(A_1A_2) = p^2 + \frac{1}{2} \cdot 2pq = p(p+q) = p \\
 P(A_2) &= P(A_2A_2) + \frac{1}{2}P(A_1A_2) = q^2 + \frac{1}{2} \cdot 2pq = q(p+q) = q
 \end{aligned}$$

If we extend the model to a locus with m alleles, we use $p_1, p_2, \dots, p_i, \dots, p_m$ to denote the frequencies of the alleles $A_1, A_2, \dots, A_i, \dots, A_m$. The $\frac{m(m+1)}{2}$ possible genotypes are then given by $A_1A_1, A_1A_2, \dots, A_iA_j, \dots, A_mA_m$. The HWL then states the following relationship between the genotypic and allelic frequencies:

$$P(A_iA_i) = p_i^2 \quad i = 1, \dots, m \quad P(A_iA_j) = 2p_i p_j \quad i \neq j, i, j = 1, \dots, m$$

This definition can be easily extended to a finite number of loci [347]. The multilocus HWL is based on haplotype frequencies instead of allele frequencies and considers combinations of haplotypes instead of one-locus genotypes.

The resulting relationship between allele and genotype frequencies is shown in Figure 2.8, which displays a de Finetti triangle.

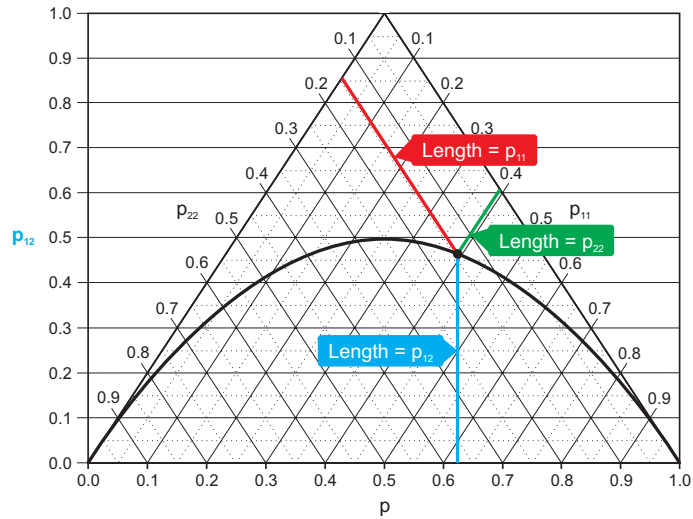


Fig. 2.8 De Finetti triangle. A point within the triangle represents a specific combination of genotype frequencies: p_{11} and p_{22} equal the projection along the side of the triangle from the point to the left and the right side of the triangle; p_{12} is the projection to the bottom side. The approximate value of p_{11} and p_{22} is obtained by following the grid line to the right and the left sides of the triangle to read the indicated frequencies. For the exemplary point, $p_{11} \approx 0.4$, $p_{22} \approx 0.15$, and $p_{12} \approx 0.45$ given by the height. For each point within this triangle, the sum of the genotype frequencies will be equal to 1. The horizontal axis shows the allele frequency for the given constellation of genotype frequencies. For the example, $p = p_{11} + \frac{1}{2}p_{12} = 0.4 + 0.225 = 0.625$, which is shown by the point where the perpendicular to the horizontal axis intersects the axis. Finally, the bold curved line depicts, for each allele frequency, the genotype frequencies that are in Hardy–Weinberg equilibrium (HWE).

Here, any point within the triangle represents a specific combination of genotype frequencies. Specifically, p_{11} is given by the projection along the side of the triangle from the point to the left side, p_{22} by the projection to the right side, and p_{12} by the projection to the bottom side of the triangle. For example, for the point marked in Figure 2.8, p_{11} is given by the length of the projection to the left along the side of

the triangle as indicated in the figure. The approximate value of this can be obtained by following the diagonal grid to the right, ending at the value 0.4. Hence, for the point indicated, $p_{11} \approx 0.4$. Similarly, p_{22} is given by the length of the projection to the right and can be read by following the diagonal grid to the left, resulting in $p_{22} \approx 0.15$. Finally, p_{12} equals the height of the point that is indicated on the vertical axis. In this case, $p_{12} \approx 0.45$. It should be noted that for each point within this triangle, the sum of the genotype frequencies will be equal to 1.

So far, we have seen what the values both on the left and right sides of the triangle and on the vertical axis stand for. The values on the horizontal axis, the bottom side of the triangle, use the specific relationship between genotype and allele frequencies. For each given constellation of genotype frequencies, i.e., for each point within the triangle, the horizontal axis depicts the resulting allele frequency p . For the exemplary point, $p = p_{11} + \frac{1}{2}p_{12} = 0.4 + 0.225 = 0.625$, which is shown by the point where the perpendicular to the horizontal axis intersects the axis.

Beginning with an arbitrary allele frequency p , exactly one constellation of genotype frequencies will be in HWE. Hence, for each point on the horizontal axis of Figure 2.8, precisely one point within the triangle represents the genotype frequencies that are in HWE. These genotype frequencies are shown by the bold curved line. Hence, even though each point within the triangle gives a possible combination of genotype frequencies, only those on the curved line are in HWE.

As pointed out before, the HWL strictly holds only under the following assumptions, which will be explained below:

- Random mating
- No selection or migration
- No mutation
- No population stratification
- Infinite population size

Random mating with respect to the genetic locus of interest means that the mating does not depend on the genotype. Hence, every individual has the same chance of mating with any individual of the opposite gender, a state that is termed *panmixia*. Non-random mating, sometimes also referred to as *assortative mating*, leads to a distortion of the equilibrium because the mating frequencies are no longer products of the genotype frequencies in the parents. This might be the case in matings within families. Also, if a genetic locus is responsible for phenotypes such as intelligence or obesity, assortative mating is common [16, 494].

In a similar way, if the probability to reproduce in general depends on the genotype at the relevant locus, a *selection* takes place. Again, this leads to mating frequencies that differ from those given in Table 2.5. For example, in many autosomal recessive diseases there is a selection against homozygotes, meaning that individuals carrying both disease-causing alleles have a reduced chance to reproduce compared to the other genotype carriers. If there is *migration* into the population, with individuals coming from populations with different genotype frequencies or individuals with certain genotypes preferentially leaving the population, this might also distort the frequency equilibrium.

Mutations (see Section 1.3) also result in a shift in the equilibrium. Here, the frequencies of the offspring's genotypes differ from those expected from the parental generation. However, mutation frequencies usually are too low to play a relevant role.

A different mechanism might result in deviations from the HWL if the investigated population is not homogeneous with respect to the analyzed genetic locus. Consider, for example, the genotype frequencies shown in the de Finetti triangle of Figure 2.9. Here, population A has the genotype frequencies $p_{11} = 0.05$, $p_{12} = 0.35$, and $p_{22} = 0.6$. As denoted by the point lying on the curved line, this is consistent with HWL. Assuming a different population B with the frequencies of p_{11} and p_{22} being interchanged, we also obtain a frequency distribution conforming to HWL. However, if we analyze a sample T that comprises individuals from both population A and population B, and if both populations contribute equally to the sample under investigation, the resulting genotype frequencies of this total population will be $p_{11} = 0.325$, $p_{12} = 0.35$, and $p_{22} = 0.325$. As shown in Figure 2.9, this mixture clearly deviates from HWL, and this phenomenon is caused by *population stratification* (a detailed explanation is given in Section 11.4).

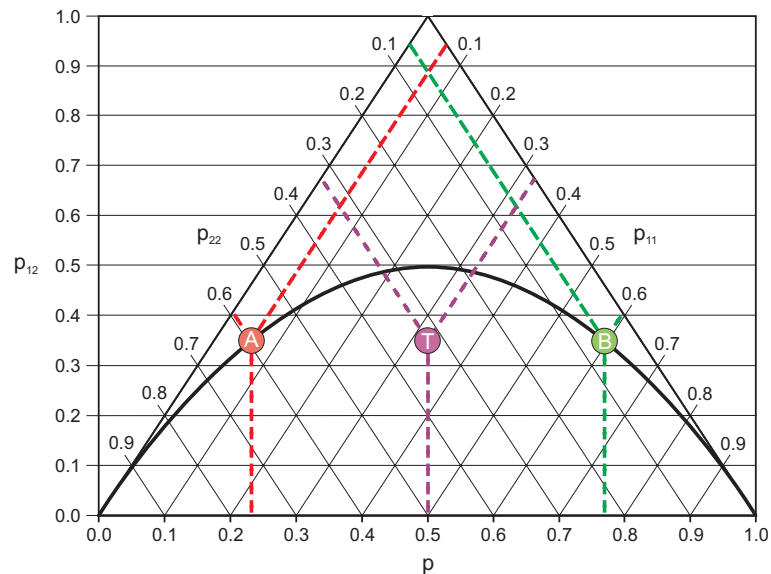


Fig. 2.9 De Finetti triangle with an example of population stratification. Both populations A and B lie on the curve indicating Hardy–Weinberg equilibrium (HWE). However, genotype frequencies from the total population T being composed of A and B in equal parts deviate markedly from HWE.

Finally, to guard against sampling errors, the population must be large enough. Theoretically, an *infinite population size* is usually assumed.

It is important to note that even if deviations from these assumptions occur, only one generation of random mating is required to restore the HWE at the single-locus

level. However, Hardy–Weinberg proportions between two or more loci cannot be restored after one generation [114].

Example 2.2. Cystic fibrosis (CF), also termed mucoviscidosis (OMIM #219700), is a recessively inherited disease with a prevalence of about 1 in 2500 live births in Caucasians. If we assume that the underlying disease locus has two alleles with D causing the disease and N not causing the disease, this means that the prevalence equals the frequency of individuals homozygous for D . Hence, $P(DD) = \frac{1}{2500} = 0.0004$. If, in a situation of genetic counseling, we are now interested in the frequency of individuals carrying the disease allele, we can use the results from above in the following way.

We will denote the frequency of the alleles D and N by p and $q = 1 - p$. Given HWE, $P(DD) = p^2$. Hence, $p^2 = 0.0004$, $p = 0.02$, and $q = 0.98$. Thus, we can calculate the frequency of carriers with $P(DN) = 2pq = 2 \cdot 0.02 \cdot 0.98 = 0.0392$. As a result, we have to assume that approximately 4% of the population carries the disease-causing allele for cystic fibrosis.

Another important application of the HWE is in the context of data-quality checking in genetic epidemiological studies. This will be explored in Section 4.3.

2.5 PROBLEMS

Problem 2.1 Consider the following pedigrees and decide according to which segregation pattern the depicted phenotype is probably inherited.

Problem 2.1.1. The shown phenotype in Figure 2.10 has a prevalence of about 1 in 10,000.

Which of the following segregation patterns is the most probable?

1. Autosomal recessive
2. Autosomal dominant
3. X-chromosomal recessive
4. X-chromosomal dominant
5. Y-chromosomal

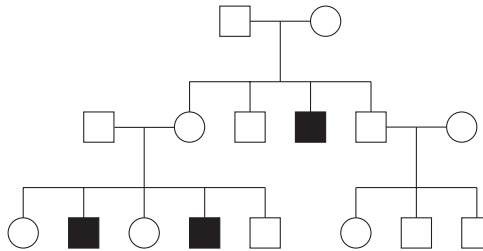


Fig. 2.10 Pedigree for Problem 2.1.1.

Problem 2.1.2. The shown phenotype in Figure 2.11 has a prevalence of about 1 in 16,000. Which of the following segregation patterns is the most probable?

1. Autosomal recessive
2. Autosomal dominant
3. X-chromosomal recessive

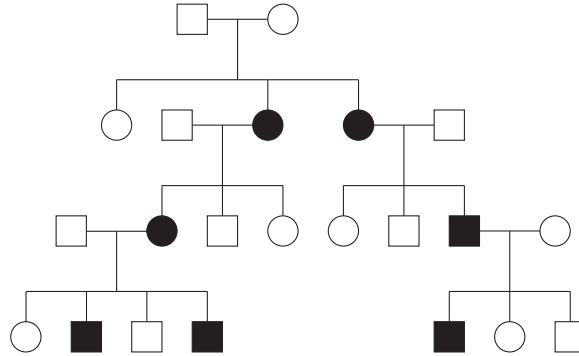


Fig. 2.11 Pedigree for Problem 2.1.2.

4. X-chromosomal dominant with reduced penetrance

Problem 2.1.3. Which of the following segregation patterns is the most probable for the depicted phenotype in Figure 2.12?

1. Autosomal recessive
2. Autosomal dominant
3. X-chromosomal recessive
4. X-chromosomal dominant
5. Y-chromosomal

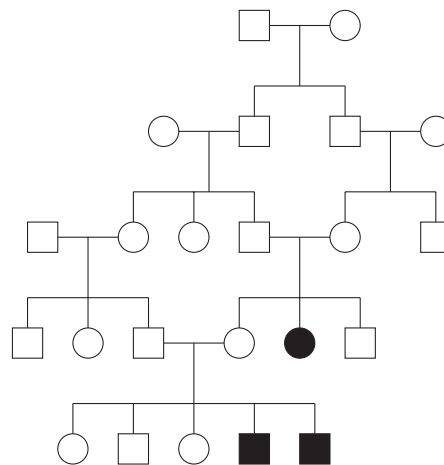


Fig. 2.12 Pedigree for Problem 2.1.3.

Problem 2.1.4. Which of the following segregation patterns is the most probable for the depicted phenotype in Figure 2.13?

1. Autosomal recessive
2. Autosomal dominant
3. X-chromosomal recessive
4. X-chromosomal dominant
5. Y-chromosomal

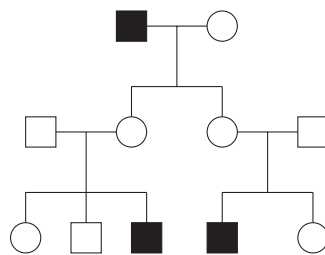


Fig. 2.13 Pedigree for Problem 2.1.4.

Problem 2.1.5. In the displayed pedigree in Figure 2.14, the index patient is interested in which kind of segregation pattern the disease might be transmitted in her family. What do you tell her?

1. The disease is transmitted with the X chromosome because only females are affected.
2. The disease follows an autosomal dominant pattern.
3. The disease follows an autosomal dominant pattern with reduced penetrance or gender-specific expression.
4. The disease follows an autosomal recessive pattern.

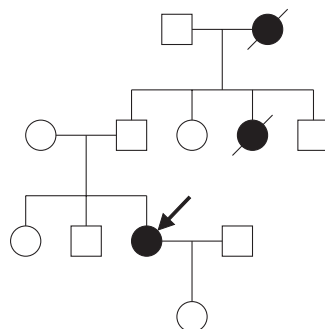


Fig. 2.14 Pedigree for Problem 2.1.5.

Problem 2.2**Problem 2.2.1.** Define the following terms:

1. Penetrance and phenocopy
2. Variable expressivity, anticipation, and imprinting
3. Genetic heterogeneity, pleiotropy, and epistasis

Problem 2.2.2. Consider the following descriptions of diseases or general phenotypes. Which phenomena might be the cause?

1. In polydactyly, some affected individuals show additional fingers, some additional toes, and some both.
2. In familial hypercholesterolemia, the number of a certain kind of receptors depends on a specific genotype. Individuals who are heterozygous typically have about half the number that individuals carrying two normal alleles have. Individuals homozygous for the disease-causing allele usually have no receptors at all.
3. Albinism follows an autosomal recessive inheritance pattern. However, if two albino individuals have a child, this child may not be affected.

Problem 2.2.3. Consider a disease with a frequency of sporadic cases in the general population of about 1 in 10,000. Concerning a known genetic diallelic locus L with the alleles A and a , assume that the probability to become affected is about 98% in individuals with the genotype aa . For heterozygous individuals, the probability depends on whether the a allele was transmitted from the father or from the mother. If the father transmitted a and the mother transmitted A , the probability for the disease is 65%. However, if the mother transmitted a and the father transmitted A , the probability reduces to 35%. Calculate the genotypic relative risks.**Problem 2.3****Problem 2.3.1.** State the main causes for deviations from HWE.**Problem 2.3.2.** Assume observance of a disease with an autosomal dominant inheritance pattern with complete penetrance and no phenocopies. The disease has a prevalence of about 1% in the general population, and the disease-causing genetic locus is in HWE. What is the frequency of the disease-causing allele? What is the frequency of carriers?**URLs**

OMIM – ONLINE MENDELIAN INHERITANCE IN MAN

<http://www.ncbi.nlm.nih.gov/omim/>

OMIM FAQ – FREQUENTLY ASKED QUESTIONS

<http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html>

IMPRINTED GENE CATALOGUE

<http://www.otago.ac.nz/igc>

3

Genetic Markers

The common starting point in genetic epidemiology is that we are faced with a disease known to have a genetic background. For example, we might have observed a pedigree as in Figure 2.3 and can clearly see a transmission pattern. Having no further information, the crucial question will be: Which gene or combination of genes is responsible for this? Or, because we want to avoid scrutinizing every single basepair of the DNA: Where in the genome should we start looking?

In order to guide our search of the genome or even parts of it, we need landmarks. To be precise, we need genetic loci with a specified and known location in the genome and that show some variability across individuals. This chapter is devoted to the description of these genetic markers and measures of their utility or informativity (Section 3.1). For simplicity, we consider informativity measures for only autosomal loci. The reader may refer to Ref. [113] for a definition of informativity measures on X-chromosomal loci. Furthermore, the different types of genetic markers that are utilized will be described in Section 3.2, together with their advantages and disadvantages. In the URLs at the end of the chapter, databases as resources of genetic markers will be listed. Detailed problems and errors that subsequently occur, as well as error checking, will be the topic of Chapter 4.

3.1 WHAT IS A GENETIC MARKER?

A *genetic marker* is a special DNA locus with at least one base being different between at least two individuals. In principle, only the features of being detectable and having a known location in the human genome are required for a locus to serve

as a marker. However, there is a list of qualities that are especially desirable in a genetic marker:

- It needs to be heritable in a simple Mendelian fashion.
- It should have only a low mutation frequency.
- It should follow a co-dominant mode of inheritance. Hence, all genotypes can be distinguished.
- It should be in Hardy–Weinberg equilibrium (HWE).
- It should be easy to ascertain reliably.
- It must be polymorphic, i.e., several distinguishable variants must exist in the population. Usually, the more the variants and the higher the proportion of heterozygotes there are, the better.

The last quality points at which markers should be preferred over others. But how can we quantify this? The usefulness of a marker hinges upon its *informativity*, which in turn is determined by the number of heterozygous individuals in a given population. If a parent is homozygous, it cannot be determined which allele is transmitted to the child, and the meiosis is not informative. Therefore, the heterozygosity of a marker, defined as the probability that a random person will be heterozygous, is a basic measure of its informativity. For an m -allelic marker with alleles A_i and allele frequencies p_i , $i = 1, \dots, m$, the *mean heterozygosity* (HET) is given by

$$HET = 1 - \sum_{i=1}^m p_i^2,$$

reducing to $HET = 1 - \frac{1}{m}$ if the marker consists in m equally frequent alleles. This definition is based on the assumption of HWE (see Section 2.4) at the marker locus because genotype frequencies can be predicted from the allele frequencies in this case.

For estimating HET , we first need estimators of the allele frequencies p_i . Consider a sample of n independent individuals so that the number of alleles is $2n$. A natural and in fact the maximum likelihood (ML) estimator of the allele frequency p_i is given by $\hat{p}_i = (2 \cdot \#(A_i A_i) + 1 \cdot \#(A_i A_j)) / (2n)$, where $\#(A_i A_j)$ denotes the number of genotypes $A_i A_j$ in the sample. The uniformly minimum variance unbiased (UMVU) estimator of HET is [614]

$$\widehat{HET} = \frac{2n}{2n-1} \left(1 - \sum_{i=1}^m \hat{p}_i^2 \right).$$

An unbiased estimator of the variance of $\widehat{HET}$ is [614]

$$\mathbb{V}\text{ar}(\widehat{HET}) = \frac{2}{2n(2n-1)} \left((3-4n) \left(\sum_{i=1}^m \hat{p}_i^2 \right)^2 + 2(2n-2) \sum_{i=1}^m \hat{p}_i^3 + \sum_{i=1}^m \hat{p}_i^2 \right).$$

These estimators can be used for estimating HET and an approximate 95% confidence interval for HET by $\widehat{HET} \pm 1.96 \sqrt{\widehat{\mathbb{V}\text{ar}}(\widehat{HET})}$, where p_i is replaced by $\hat{p}_i$ for

estimating $\text{Var}(\widehat{HET})$. These calculations can be carried out with the HET program available from the Web.

However, consider the case in which both parents have the same heterozygous genotype: in about half of these situations, the offspring will be heterozygous as well, and it cannot be ascertained which allele is of maternal and which allele is of paternal origin. Hence, a more sophisticated measure of informativity is the *polymorphism information content (PIC)* [77, 269], which also considers couples who are both heterozygous and takes into account that half of their children will be heterozygous, thus uninformative. As before, we consider a genetic marker with m alleles A_i , $i = 1, \dots, m$. Here, the mating types as shown in the left column of Table 3.1 are possible. With the respective allele frequencies denoted as above, both

Table 3.1 Mating types, mating type frequencies, and possible offspring genotypes given the mating type; conditional probability I_{1t} that one can deduce which allele was transmitted by the first parent given the mating type; and informativity of the parental source I_{PS} .

Mating type	Frequency	Offspring	I_{1t}	I_{PS}
1 $A_i A_i \times A_i A_i$	$\sum_{i=1}^m p_i^4$	$A_i A_i$	0	0
2 $A_i A_i \times A_j A_j$	$\sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2$	$A_i A_j$	0	1
3 $A_i A_i \times A_i A_j$	$2 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^3 p_j$	$A_i A_i,$ $A_i A_j$	0	$\frac{1}{2}$
4 $A_i A_j \times A_i A_i$	$2 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^3 p_j$	$A_i A_i,$ $A_i A_j$	1	$\frac{1}{2}$
5 $A_i A_i \times A_j A_k$	$\sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m \sum_{\substack{k=1 \\ i \neq k \\ j \neq k}}^m p_i^2 p_j p_k$	$A_i A_j,$ $A_i A_k$	0	1
6 $A_j A_k \times A_i A_i$	$\sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m \sum_{\substack{k=1 \\ i \neq k \\ j \neq k}}^m p_i^2 p_j p_k$	$A_j A_i,$ $A_k A_i$	1	1
7 $A_i A_j \times A_i A_j$	$2 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2$	$A_i A_i,$ $A_i A_j,$ $A_j A_j$	$\frac{1}{2}$	0
8 $A_i A_j \times A_i A_k$	$4 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m \sum_{\substack{k=1 \\ i \neq k \\ j \neq k}}^m p_i^2 p_j p_k$	$A_i A_i,$ $A_i A_k,$ $A_j A_k,$ $A_i A_j$	1	$\frac{3}{4}$
9 $A_i A_j \times A_k A_l$	$\sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m \sum_{\substack{k=1 \\ i \neq k \\ j \neq k}}^m \sum_{\substack{l=1 \\ i \neq l \\ j \neq l \\ k \neq l}}^m p_i p_j p_k p_l$	$A_i A_k,$ $A_i A_l,$ $A_j A_k,$ $A_j A_l$	1	1

the frequencies expected under the assumption of random mating and the possible genotypes for the offspring are shown. For informativity, we are now interested in the probability I_t that we are able to deduce which allele was transmitted by an arbitrarily chosen parent.

These calculations can be simplified by looking at the transmission informativeness of the first parent I_{1t} in Table 3.1. Specifically, I_{1t} denotes the probability that we are able to deduce which of the two alleles of the first parent has been transmitted to his/her offspring. This transmission cannot be determined if the first parent is homozygous. This occurs for mating types 1–3 and 5 with a total frequency of

$$\sum_{i=1}^m p_i^4 + \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2 + 2 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^3 p_j + \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m \sum_{\substack{k=1 \\ i \neq k \\ j \neq k}}^m p_i^2 p_j p_k = \sum_{i=1}^m p_i^2.$$

In addition, half of the matings where both parents have the same heterozygous genotype, i.e., mating type 7, are not informative. Putting these together, the frequency of matings in which it is known which allele was transmitted from the first parent is calculated to

$$1 - \sum_{i=1}^m p_i^2 - \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2.$$

This probability will be the same if the second parent is considered instead of the first parent. Hence, the average informativeness is given by

$$PIC = 1 - \sum_{i=1}^m p_i^2 - \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2.$$

Interestingly, the PIC value can also be interpreted as the probability of knowing the parental source of each of a child's alleles [269]. In the first step, this probability is determined for each mating type. For example, four offspring genotypes are possible for mating type 8. We can determine unambiguously which allele has been inherited from which parent for all genotypes except the $A_i A_i$. This is impossible for the $A_i A_i$ genotype because we cannot determine which of the two A_i alleles in the offspring has been inherited from the first parent. The informativity of the parental source I_{PS} therefore equals $\frac{3}{4}$ for mating type 8 (right column of Table 3.1). Following this, the PIC can also be obtained by

$$\begin{aligned} PIC &= 1 - \sum_{i=1}^m p_i^4 - \frac{1}{2} \cdot 2 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^3 p_j - \frac{1}{2} \cdot 2 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2 \\ &\quad - 2 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2 - \frac{1}{4} \cdot 4 \cdot \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m \sum_{\substack{k=1 \\ i \neq k \\ j \neq k}}^m p_i^2 p_j p_k \\ &= 1 - \sum_{i=1}^m p_i^2 - \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m p_i^2 p_j^2. \end{aligned}$$

The UMVU estimator and the ML estimator of the PIC value and their exact variances can be found in Ref. [614].

Many analyses that will be introduced in Chapters 8 and 9 focus on the genetic similarity of sib-pairs within a family. Specifically, the core of those analyses will be the number of marker alleles the sibs share that are identically transmitted from father or mother. Conforming to this, the *linkage information content* (LIC) has been defined as the probability of knowing how many marker alleles are identical in the sibs. To avoid reproducing the extensive tables, the reader is referred to Ref. [269], Table 2, which was corrected in Ref. [271], Table 1. As a result, the LIC for sib-pairs is given by

$$LIC = 1 - \sum_i^m p_i^2 + \frac{1}{2} \sum_i^m p_i^4 - \frac{1}{2} \left(\sum_i^m p_i^2 \right)^2.$$

For a m allelic marker with identical allele frequencies, the LIC reduces to

$$LIC = \frac{(m-1)(2m^2-1)}{2m^3}. \quad (3.1)$$

This definition can be extended to other pairs of unilineal relatives [271]. However, the UMVU estimator and its variance for the LIC have not been provided in the literature so far.

Because the LIC is still an uncommon informativity measure, it is important to see that it can be easily expressed in terms of the two more common measures of marker informativity, HET and PIC , if the genetic marker has equally frequent alleles. For example, $LIC = \frac{HET}{2}((\sqrt{2}-1+HET)(\sqrt{2}+1-HET))$ can be easily shown. Furthermore, the relationship between LIC and PIC can be established using the following detour: in the first step, the PIC value is utilized to calculate the corresponding number of equipotent alleles m by

$$m = \frac{1}{3(1-PIC)} + 2\sqrt{\frac{4-3PIC}{9(1-PIC)^2}} \cos\left(\frac{1}{3} \arcsin \sqrt{1 + \frac{(-2-9(1-PIC)+27(1-PIC)^2)^2}{4(-4+3PIC)^3}}\right). \quad (3.2)$$

In the second step, this can then be entered into Eq. (3.1) to compute the LIC from the number of m equipotent alleles. Equation (3.2) has been derived and kindly provided by Xiuqing Guo in order to avoid computations that involve complex numbers using the theory described in Section 5.3 of Ref. [672]. These more complicated calculations are required when using the formula from p. 103 in Ref. [270].

In most applications, not only single genetic markers but also a number of markers that are close to each other will be analyzed at the same time. Hence, it might be of interest to calculate the informativity of clusters of markers, and a measure for this is the *multilocus polymorphic information content* ($MPIC$) suggested in Ref. [249].

Example 3.1. To compare the informativity of a marker with different numbers of alleles and different allele frequencies, we consider two diallelic markers A and B and two markers with five alleles, C and D. Let us assume that the respective alleles have frequencies as shown in Table 3.2.

Table 3.2 Example of informativeness of different markers.

Marker	Allele frequencies	HET^a	PIC	LIC
A	0.5, 0.5	0.50	0.3750	0.4375
B	0.9, 0.1	0.18	0.1638	0.1719
C	0.2, 0.2, 0.2, 0.2, 0.2	0.80	0.7680	0.7840
D	0.5, 0.2, 0.1, 0.1, 0.1	0.68	0.6420	0.6610

^a HET is the mean heterozygosity, PIC the polymorphism information content, and LIC the linkage information content.

When the informativeness of the different imaginary markers is compared, two interesting issues can be learned. First, with a greater number of alleles, the informativeness is higher. Second, the highest informativity can be obtained by using markers that have alleles of about equal frequencies. Calling into mind the relationship between allele and genotype frequencies that was depicted in the de Finetti triangle in Chapter 2, Figure 2.8, this does not come as a surprise. There, we saw that the number of heterozygous genotypes is greatest when the alleles are of equal frequency.

3.2 WHAT TYPES OF GENETIC MARKERS ARE THERE?

The types of genetic markers utilized is a rapidly changing topic, and the history of the markers that are used today is quite young [593]. However, we refrain from giving a historical overview on the development of markers but instead focus our description on the two types that are most often seen today: microsatellites and single nucleotide polymorphisms. Excellent and thorough information on other types of markers can be found in textbooks on human genetics, e.g., by Strachan and Read [636].

3.2.1 What are short tandem repeats?

In the non-coding regions of the DNA, large areas can be found where parts of the sequence are repeated in tandem. Because the number of repeated units varies between individuals, these can be utilized as genetic markers. Depending on the size of the single repeated unit, these are called *variable number of tandem repeats* (VNTRs) or *microsatellites*, which is used synonymously for *short tandem repeats* (STRs). In contrast to other repeat polymorphisms, STRs consist of shorter sequences of typically two to four nucleotides. Based on the number of repeated nucleotides, they are termed *mononucleotides*, *dinucleotides*, *trinucleotides*, or *tetranucleotides*.

Some STRs are neither true mononucleotides nor dinucleotides and are therefore termed *non-integer STRs*. The total expansion size of STRs ranges from about 10–100 kilobases (Kb). Across all genomic regions, mononucleotide repeats are most frequent, followed by di- and tetranucleotide repeats. In exons, by contrast, trinucleotide repeats are most frequent.

Most STRs are short enough to be amplifiable via standard polymerase chain reaction (PCR) methods and separable in high-resolution media (see Section 1.3.2). Through this, alleles that differ in size by only a single repeat unit can be resolved unambiguously, e.g., by using gel electrophoresis. To enhance efficiency, compatible STRs can be produced that can be amplified together with non-overlapping allele sizes and different coloring. For example, if the reading frame of a microsatellite is 130–160 basepairs (bp) and 170–190 bp for a second microsatellite marker, both can be separated on the same gel. It is, however, important to know that PCR reactions should be performed separately for the two markers. In practice, laboratories are able to separate alleles from 5–10 STR markers in the same step.

The naming of STRs has been standardized to include some conventional information. Hence, a typical STR might be called *D15S650*, where D stands for DNA, 15 for the chromosome 15, S for single copy, and 650 for the 650th system described on chromosome 15. The number of repeat units that determine the length of an allele is utilized to name the specific allele. For instance, three alleles of some trinucleotide STR might be 136, 139, and 145. Note that the alleles already indicate that the marker is a trinucleotide marker, as the differences between the allele lengths $139 - 136 = 3$ and $145 - 139 = 6$ are divisible by 3.

STRs have the distinct advantage of being highly polymorphic, although there are usually only one or two highly frequent alleles. Concerning the resulting informativeness, this means that the large number of alleles leads to high informativity. Hence, the detection and subsequent use of STRs enabled the construction of genetic maps, and for a long period of time, they have been viewed as ideal genetic markers.

On the other hand, there are some disadvantages with the use of microsatellites:

- Their high polymorphism is mainly due to their rather high mutation frequency of 10^{-2} – 10^{-6} events per locus per generation [411].
- Because of technical problems, they do not lend themselves easily to automatic scoring.
- They are frequent but possibly still not dense enough for some applications.

The problems and frequent errors associated with genotyping of STRs will be discussed in detail in Chapter 4. However, we want to stress at this point that reading of STR genotypes is generally time-consuming, is error prone, and requires technical experience (compare Section 4.2.1).

It has been debated whether STRs have biological functions. On the one hand, it has been assumed that they have no biological use but are useful only in the areas of forensic DNA profiling and genetic linkage analysis [54]. However, biological consequences have been assigned as trinucleotides have been found to be associated with diseases such as Huntington's disease, myotonic dystrophy, and certain types of spinocerebellar ataxia. This group of diseases is characterized by the primary

cause, the expansion of a trinucleotide repeat far beyond the normal polymorphic range. Interestingly, the disease severity often correlates with the extent of abnormal expansion, and the longer these repeats are, the more prone they are to still further expansion. This phenomenon might explain the anticipation that has been observed in these diseases (see Section 2.3.1).

More generally, STRs are often found at frequencies much higher than would be predicted purely on the grounds of base composition, and functional significance of a substantial number of STRs has been proven by critical tests in various biological phenomena [411]. Hence, the question of neutrality versus functionality might not be a contradiction but merely a question of quantity.

To select markers for a certain chromosomal region or to obtain further information on a specific STR, several databases are available on the World Wide Web. For example, details on markers can be found in the nucleotide database, headed by the National Center for Biotechnology Information (NCBI). Similarly, markers and locations can be browsed in the Human Genome Database (GDB). Other databases include UniSTS, available via NCBI, the location database (LDB), which gives locations for expressed sequences and polymorphic markers; and the Marshfield Web page, maintained by the Center for Medical Genetics of the Marshfield Clinic Research Foundation, containing much information on human DNA polymorphisms.

3.2.2 What are single nucleotide polymorphisms?

The most abundant form of variation in the human genome, about 90%, are *single nucleotide polymorphisms* (SNPs). Basically, these are variations at only a single base, meaning that one base is substituted with another. Loose and strict definitions for SNPs have to be distinguished in practice. While the looser definition makes no assumption about the *minor allele frequency* (MAF), i.e., the frequency of the rarer allele, the strict definition requires that the MAF be at least 1% in the population in order for a polymorphism to be called a SNP. Because for a SNP one base is exchanged for another, in theory four alleles are possible; however, in humans almost only diallelic SNPs occur, with about two-thirds having cytosine and thymine (C and T) or guanine and adenine (G and A) alleles.

There are several nomenclatures for SNPs, and they can be quite confusing for statisticians. The most commonly used nomenclature is based on reference SNP ID numbers, or *rs* IDs. The *rs*-number is an identification tag assigned by the NCBI in the United States of America to a group or cluster of SNPs, each described by a *submitted SNP number* (*ss-number*) that map to an identical location. *rs*-numbers can be searched in dbSNP and are mapped to external resources or databases, e.g., Ensembl. Like UniSTS, dbSNP is a public domain archive for a broad collection of nucleotide variations [534, 612]. The *rs*-number is noted on the records of these external resources and NCBI databases in order to direct users back to the original dbSNP records. Although *rs*-numbers are unique and therefore ideally suited for search in databases, they do not provide information about the possible function of a SNP. This is overcome by the nomenclature of the Human Genome Variation Society [159, 309]. For example, IVS4-2A>C describes an intron variation with alleles A

and C (“IVS”) in intron 4 at position -2 , i.e., position 2 starting from the G of the AG splice acceptor site. Similarly, IVS4+1G>T denotes a G to T substitution at position +1 of intron 4.

SNPS are very frequent, and the steep rise in their investigation is shown in the following figures. In 1998, 5000 SNPs had been known. This rose to 2.2 million and 8.8 million SNPs in 2002 and 2004. The last number was obtained from the latest build 130 of dbSNP and includes 17.8 million SNPs with rs-numbers, of which 9.5 million were validated.

Although very frequent, SNPs are not uniformly distributed across the genome. In regions such as the non-coding human leukocyte antigens (HLAs), the frequency of these polymorphisms can go up to 10%. Nonetheless, SNPs are found in various parts of the DNA. Accordingly, coding SNPs (cSNPs) that occur in coding regions of genes are distinguished from non-coding SNPs in the non-coding regions around the immediate neighborhood of genes and from random non-coding SNPs in the intergenic area. Even more detailed, cSNPs are differentiated with regard to their consequences: synonymous changes that do not alter the coded amino acid are less consequential than non-synonymous changes that lead to a different amino acid coding. Interestingly, it has been shown for a number of SNPs within genes that fewer synonymous SNPs occur than would be expected by chance, which hints at a selection against these major changes [105]. The different types of SNPs, together with their estimated frequencies, are shown in Table 3.3.

Table 3.3 Typology of single nucleotide polymorphisms (SNPs) and their frequency.

Description	Frequency (in thousands)
Coding, non-synonymous, non-conservative	60–100
Coding, non-synonymous, conservative	100–180
Coding, synonymous	200–240
Non-coding, 5' UTR	140
Non-coding, 3' UTR	300
Other non-coding	> 1000

Adapted from Ref. [567], Table 2.

As with STRs, novel sequence variations may emerge. However, because of their low mutation frequency, most SNPs originated some time before the emergence of our different human populations. Hence, we usually do not share our SNPs with other primates, but about 85% are common to all human populations, with differing allele frequencies [89].

Because of their short length, SNPs can be easily typed by PCR methods. Specifically, a variant termed *PCR amplification of specific alleles (PASA)*, also called *amplification refractory mutation system (ARMS)*, can be utilized. Here, two PCRs are carried out. One primer is the same in both PCRs; the other exists in two different versions. The first version binds only to the “normal” DNA sequence, while the other only binds to the “variant” sequence. Additional control primers are usually included

for amplifying some unrelated sequence from every sample to check the PCR reaction has worked. The result of the genotyping can then be read via electrophoresis. This gives two bands, one for the control sequence and one for the sequence of interest. Therefore, automated single reading is generally sufficient for SNPs typed by ARMS in contrast to standard typing, where greater effort is required.

The major advantage of SNPs is that the concept underlying ARMS can be extended so that SNPs can be determined with techniques other than gel electrophoresis. This allows a large degree of automatization if primers are designed carefully. For example, primers can be designed in a way that the PCR products have slightly different molecular weights. If these ions are then accelerated in a vacuum toward a target, one can determine their time of flight (TOF). Molecules with different weights generally require different TOFs, thus allowing separation of alleles by measuring the TOF. This idea is utilized in the so-called *matrix assisted laser desorption/ionization time of flight mass spectrometry* (MALDI-TOF MS) [636]. Here, single SNPs of many individuals are rapidly determined.

An alternative approach is to color-label the two variant specific primers. The genotype is then called by the color so that homozygotes are, e.g., either green or red, while heterozygotes are yellow [336]. With this level of automatization, high-throughput typing using DNA microarrays or DNA microtiters with 10, 000 SNPs or more at a time becomes possible.

Thus, although the informativeness of a single SNP is low, the decisive strength lies in the automated high-throughput typing of many SNPs. In summary, the advantages of SNPs are as follows:

- SNPs are very frequent and hence provide a dense marker spacing.
- Although most SNPs are found in non-coding regions, some of them are located in genes or in the promoter of genes and can therefore be directly viewed as candidate variations for disease.
- Whereas STRs show a high mutation frequency, SNPs are much more stable.
- SNP typing can be highly automated in a high-throughput genotyping environment.

These advantages have led to SNPs being heralded as the catalyst to unravel the genetic basis of complex diseases and interindividual variation in drug response. A number of resources as listed in the URLs section at the end of this chapter are available on the World Wide Web where SNPs can be selected for specific purposes.

As SNPs are always diallelic, they are inherently less informative than STRs. As seen in Example 3.1, the maximal informativity of a marker is given when alleles are equally distributed. Hence, the maximal informativity a SNP can reach is a *PIC* of 0.3750 or a *HET* of 0.5. In contrast, a STR with five alleles has a maximal informativity of 0.7680 or 0.8, respectively. Consequently, more SNPs need to be genotyped for the same study in order to achieve a comparable level of informativity. The question of how many more SNPs need to be genotyped to match the information content for linkage analysis of an average multiallelic STR has been investigated by different groups, and estimates range from two to over five [696]. In other words, a genome-wide linkage study that uses 300 STRs with a genetic distance of about 10

centiMorgan (cM) from each other (see Chapter 5) might be replaced with a study design utilizing 750 to 1500 SNPs with an average distance of 5 cM–2 cM from each other.

3.3 WHAT ARE GENOTYPING METHODS FOR SINGLE NUCLEOTIDE POLYMORPHISMS?

Jeanette Erdmann

In this section, we give a short overview about commonly used methods for genotyping large collections of DNA samples because several technologies for rapid genotyping have emerged in the last few years (Figure 3.1). These are reviewed in detail, e.g., in Refs. [97, 443, 553, 554, 610, 652] and the various chapters in Ref. [383]. Depending on the specific experimental setup, the principal investigators need to carefully choose the most appropriate method for the project. The final decision will be based on

- the total number of samples to be analyzed,
- the total number of SNPs to be analyzed,
- the amount of DNA available for the project,
- the quality of DNA,
- the time for processing the DNA samples, and, finally,
- the total cost.

Because these aspects are crucial for the success of any genetic epidemiological study that involves genotyping, we specifically focus on the pros and cons of the different technologies. At the end of this section, we illustrate the decision-making process in such genotyping projects using three real experiments.



Fig. 3.1 Evolution of genotyping methods. Starting in the 1980s with restriction fragment length polymorphism (RFLP) analysis, a method that allowed genotyping of just one single nucleotide polymorphism (SNP) in a small number of samples. In the early 2000s, the invention of the chip-technologies allowed the simultaneous genotyping of 10,000 SNPs in hundreds of patients. This technique has been further developed and currently allows to genotype 1,000,000 SNPs in thousands of patients almost simultaneously. In the near future, techniques such as next generation sequencing will allow to determine ones complete genetic variability in a short time at reasonable cost.

Genotyping refers to the process of determining the genotype of an individual by the use of any biological assay. In the context of genetic epidemiological studies,

genotyping is important for the identification of genes that are associated with diseases or quantitative traits (see Chapter 11). Present technologies only allow the genotyping at specific pre-specified chromosomal positions. However, in the near future, it will be possible to genotype the entire genome of a subject at once through whole-genome genotyping. Here, we restrict our focus on current technologies, i.e., SNP genotyping at a specific chromosomal position. Table 3.4 gives an overview on the most commonly applied genotyping methods.

3.3.1 What is restriction fragment length polymorphism analysis?

The first method to genotype a proband's DNA was developed in the early 1980s. This method was called restriction fragment length polymorphism (RFLP) analysis. This method relies on the fact that specific restriction enzymes are able to cut (digest) DNA depending on a specific DNA sequence. This means that by using RFLP analysis the DNA sample is cut in different fragments (digested) by a specific restriction enzyme and the resulting restriction fragments are separated according to their lengths by gel electrophoresis (Figure 3.2). The method is almost obsolete nowadays, and, for details, the interested reader is referred to Ref. [574].

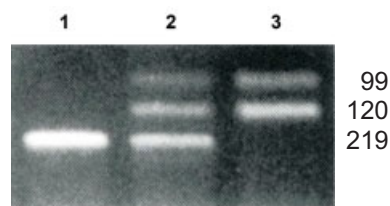


Fig. 3.2 Example of the result from restriction fraction length polymorphism analysis using the restriction enzyme *MspI*. Subject 1 is homozygous *A/A* as can be seen from the single bright band at 219 bp. Subject 3 is homozygous *B/B* and has two bright bands at 99 bp and 120 bp. Note that the sum is 219 bp. Subject 2 in the middle is heterozygous. This person has three bands, the one at 219 bp for the *A* allele as well as the two bands at 99 bp and 120 bp for the *B* allele. The figure has been kindly provided by Jeanette Erdmann.

3.3.2 What is the principle of genotyping methods relying on quantitative, real-time polymerase chain reaction?

The *real-time polymerase chain reaction (RT-PCR)*, also called *quantitative real-time PCR (Q-PCR/qPCR)* or *kinetic PCR*, is a laboratory technique based on PCR, and it is used to amplify and simultaneously quantify a target DNA molecule. It enables both detection and quantification of a specific sequence in a DNA sample, as absolute number of copies or relative amount when normalized to DNA input or additional normalizing genes.

The method has been further developed by Applied Biosystems[™] and is nowadays referred to as *TaqMan*[®] technology. Details are described in Figure 3.3. The

Table 3.4 Genotyping methods for low to ultra-high-throughput genotyping.

Method	Principle	Throughput	Reference
Allele-specific PCR	A common reverse primer and two forward allele-specific primers with different tails amplify two allele-specific PCR products of different lengths, which are further separated by agarose gel electrophoresis	Low	[238]
RFLP analysis	The DNA sample is digested by restriction enzymes and the resulting restriction fragments are separated according to their lengths by gel electrophoresis	Low	[574]
Pyrosequencing	Sequencing by synthesis, involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically	Middle	[509]
SNPlex	Oligonucleotide ligation/polymerase chain reaction and capillary electrophoresis	Middle	[663]
TaqMan	Quantitative real-time PCR, allele-specific TaqMan probes	Middle	[427]
SNPstream	Single-base primer extension technology	Middle to high	[51]
Affymetrix	Microarray based, fluorescence-labeled DNA	Ultra-high	[451]
Illumina	Microarray based, fluorescence-labeled extension products	Ultra-high	[629]

Low: single SNP in up to 500 of samples per day.
Middle: 3–10 SNPs in up to 1000 samples per week.
High: up to 50 SNPs in thousands of samples per week.
Ultra-high: thousands of SNPs in thousands of samples per week.

fundamental idea of the TaqMan is to use two different fluorescence labeled probes. Each probe matches one specific allele of a SNP. The fluorescence signal of each probe can be automatically detected, and the TaqMan method is therefore suitable for automated reading of fluorescence intensities, which, in turn, can be translated into genotypes (see Chapter 4).

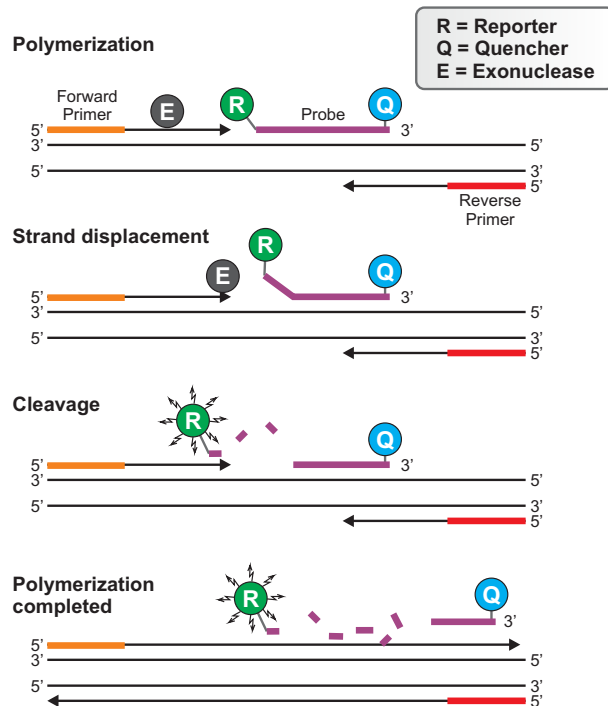


Fig. 3.3 The TaqMan probes are so designed that they anneal within a DNA region amplified by a specific set of primers. These probes typically span 25–30 bp at a SNP position. a) The Taq polymerase extends the primer and synthesizes the nascent strand. b) The 5' to 3' exonuclease activity of the polymerase degrades the probe that has annealed to the DNA template. c) Degradation of the probe releases the fluorophore from it and breaks the proximity to the quencher, thus relieving the quenching effect and allowing fluorescence of the fluorophore. d) Hence, fluorescence detected in the real-time polymerase chain reaction thermal cyclers is directly proportional to the fluorophore released and the amount of DNA template present in the PCR. For TaqMan genotyping two probes with different fluorophores are used, each matching one of the two alleles of the respective SNP.

3.3.3 What is the matrix assisted laser desorption/ionization time of flight technology?

While the TaqMan method uses fluorescence signals to detect the presence of a specific allele, the *matrix assisted laser desorption/ionization time of flight (MALDI-TOF)* genotyping method aims at detecting different masses of a molecule by mass spectrometry. Specifically, the heavier the molecule, the higher its *time of flight (TOF)*. One specific platform utilizing the MALDI-TOF principle is the MassARRAY® iPLEX® of Sequenom®.

The entire MassARRAY iPLEX process can be completely automated including assay development, PCR setup, post-PCR treatment, transfer of extension products to chips, serial reading of chip positions in the mass spectrometer, and, finally, analytical interpretation. The method allows multiplexing with up to 25–30 different SNPs to be genotyped in a small portion of DNA. The multiplexing allows high-throughput genotyping of thousands of genotypes per day at relatively low cost. The principle of the MassARRAY iPLEX method is described in Figure 3.4 and nicely reviewed in Ref. [554].

3.3.4 What are chip-based genotyping methods?

In the last few years, the potential of genotyping methods has been further revolutionized by the development of microarray-based platforms that facilitate very high-throughput and cost-efficient genotyping of hundreds of thousands of SNPs on a single microarray. There are two platforms commonly used for these genome-wide association (GWA) studies based on this microarray technique, Illumina and Affymetrix, which fundamentally differ in their SNP detection system.

The SNP genotyping on an Affymetrix microarray platform is based on the detection of both alleles of a SNP by a hybridization reaction. Each allele of a SNP is represented by a short DNA probe spotted on a microarray. To determine the genotypes of a DNA sample, the DNA is first amplified, fragmented and labeled with one single fluorescence dye. The second step in this genotyping process is hybridization in which the DNA binds the complementary DNA probes spotted on the microarray. The genotyping relies on the fact that a binding between the DNA to be genotyped and the probes is possible only in case of a perfect match. In this case, we will get a fluorescence signal that can be detected by a reader.

The Illumina microarray platform relies upon enzymatic extension of genomic DNA captured at specific sequences spotted on the arrays. The fragmented and amplified DNA anneals to these locus-specific fragments. Thereby, each SNP is represented by two allele-specific fragments, and after the annealing step the allelic specificity is conferred by enzymatic extension and subsequent fluorescent labeling with two allele-specific dyes. The intensities of the allele-specific fluorescence can be automatically detected by a reader. In contrast to the Affymetrix platform in which the genotyping relies on the intensity of only one fluorescence dye, the Illumina platform determines the genotype of each SNP based on the relative signal observed for each allele.

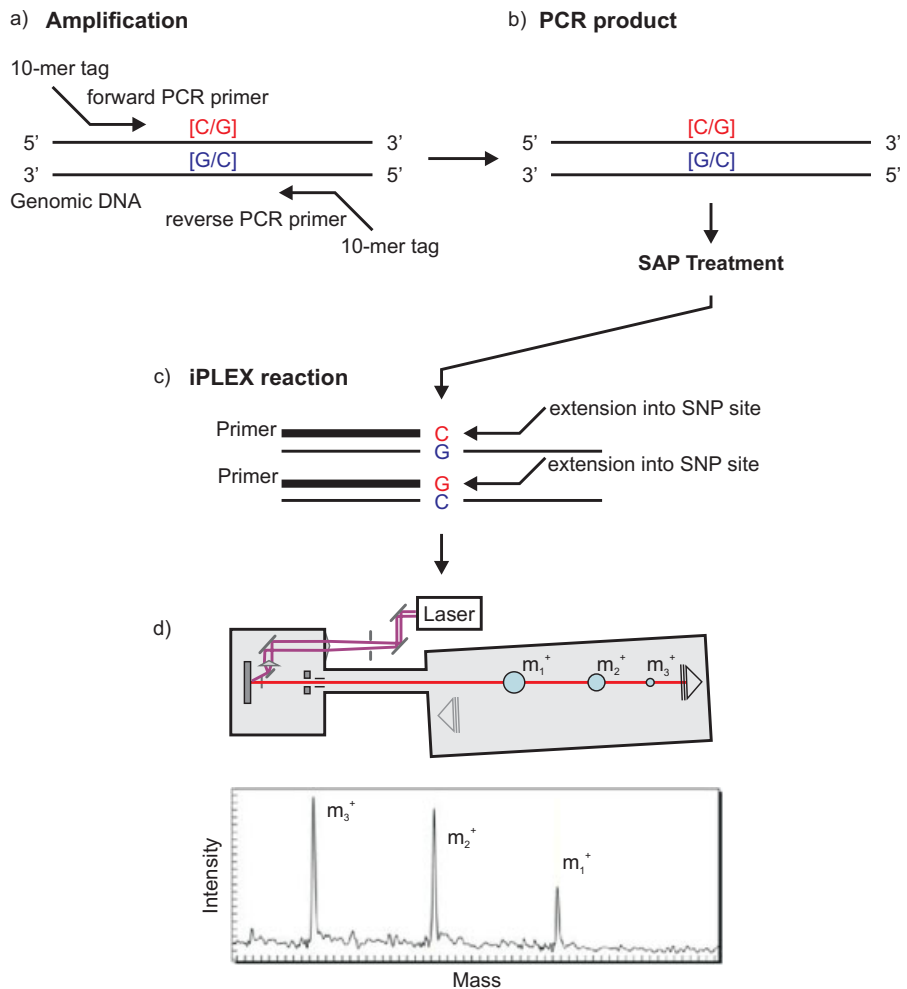


Fig. 3.4 MassARRAY iPLEX technology. The genotyping process consists of four steps. a) A polymerase chain reaction (PCR) is performed to amplify the region surrounding the SNP. b) Treatment of PCR reaction with shrimp alkaline phosphatase to dephosphorylate unincorporated nucleotides. c) A single base extension reaction is performed using thermo sequenase to incorporate mass modified nucleotides. This post-PCR primer extension reaction generates small DNA products that have unique mass values that are different for each allele. d) Mass extension products are spotted on a chip consisting of matrix that helps in the ionization of the DNA products when exposed to high intensity laser. The time of flight (TOF) of these ionized products will depend on the mass of each allele that is measured by the mass spectrometer.

Both platforms offer advantages and disadvantages with regard to genotyping calling and SNP content. However, both techniques are ideally suited for conducting GWA studies. For further reading, the reader may refer, e.g., to Ref. [443].

3.3.5 When to use which genotyping methodology?

The choice of a specific genotyping method relies on the purpose of the experiment. Here, one must keep in mind that the different methods differ in costs and, in most instances, in the technologies available in the laboratory. We therefore describe three real experiments in this section for illustrating the decision-making process, and we explain the pros and cons.

Example 3.2 (Experimental setup 1).

Background *In a small set of patients, a potentially functional mutation was identified by sequencing. This potentially functional mutation was an amino acid substitution in a highly conserved functionally relevant position. The next steps would be to further explore if this genetic variant is indeed a causal variant, capable of causing the disease, or if the variant is only rare without having an effect on the disease. In the latter case, this variant will also be found in healthy controls; for details of this experiment, see Ref. [468].*

Aim *The aim of this study therefore was to screen a large sample of healthy controls and to determine the allele frequency of this variant. If this variant is also found in controls, one has to assume that the variant has no disease-causing function. If, instead, the variant cannot be observed in more than 500 chromosomes of healthy controls, this variant is likely disease-relevant. Of course, further functional studies are required to prove the causal relationship.*

Task *To study a sample of 250 DNAs for the presence of a single genetic variant with highest possible accuracy. In Section 1.3.3, the general approach is described how the number of subjects required for screening can be determined.*

Decision *The gold standard for this task would be the sequencing of the relevant DNA sequence following PCR amplification. However, sequencing is a relatively expensive method, and a DNA sequencer is not available in every laboratory. It is time-consuming to amplify, purify and sequence the samples. Although sequencing is widely used in this setting, an alternative method would be to design an RFLP assay. In most instances, a base-pair substitution creates or destroys a restriction site for a restriction enzyme.*

Pros and Cons *Establishing an RFLP assay depends on the specific DNA context and, therefore, in rare cases it might not be possible to establish such an assay. In addition, establishing the optimal digest conditions is time-consuming in some instances. The second best method is the direct sequencing of PCR fragments. Sequencing quality, however, depends on sequence context. For example, when there is a high GC content (see Chapter 1), it is sometimes almost impossible to get an accurate sequence. The costs between establishing an RFLP assay and direct sequencing vary between 3 € (for a RFLP assay)*

and 16 € for sequencing as per sample costs. If money is no constraint, the decision will go for sequencing. In most applications, the principal investigator will decide for an RFLP assay.

Example 3.3 (Experimental setup 2).

Background A GWA study (see Chapter 14) was undertaken with subsequent in-silico replication steps [191]. To rule out that the association identified in the GWA study and in the subsequent in-silico replication step is not a false positive, a wet-lab replication phase was planned. The reader should note that the initial in-silico replication step substantially reduced the number of SNPs that was needed to be genotyped in the wet-lab replication phase.

Aim Wet-lab replication of three SNPs, identified in a GWA study; genotyping was to be performed in more than 20,000 DNA samples.

Task Rapid and reliable genotyping of three to ≤ 10 SNPs in $> 20,000$ DNA samples.

Decision In this study, the most commonly chosen genotyping methodology is the TaqMan. It allows rapid genotyping of several hundreds to thousands of DNA samples per day at intermediate costs. The accuracy is very high; the requirements regarding DNA quality and amount do not need to be excellent. Specifically, this method does not need a specific preparation of the DNA samples.

Pros and Cons The TaqMan technology is the method of choice if the principal investigator needs to genotype a small number of SNPs in large samples. Using this method on 384 well-plates allows to genotype up to 1300 DNA samples per day, yielding 6500 genotypes per week. The method thus is not a high-throughput method. Nevertheless, taking costs and expenditure of work into account, this method is the best choice for such a middle-scale project.

Example 3.4 (Experimental setup 3).

Background A GWA study was performed in five steps in total [348]. In stage 3 of this specific study, a total of 33 SNPs were chosen for further genotyping in 8000 samples.

Aim Replication of approximately 30 SNPs in 8000 DNA samples.

Task Rapid and reliable genotyping at low costs of 33 SNPs in 8000 DNA samples.

Decision In this setting, the principal investigator decided to use the MALDI-TOF allowing rapid genotyping of multiple of 30 SNPs in multiplex genotyping reactions. The possibility to multiplex the genotyping reactions leads to high throughput with thousands of genotypes per day.

Pros and Cons The MALDI-TOF technology is a relatively cheap genotyping method with high throughput. This method requires, however, an expensive investment in the technology platform per se. From a technical point of view, this method generates high-quality genotypes. Unfortunately, some SNPs cannot

be processed on this platform because of cross-reactions of primers during the PCR. In this case, the principal investigator is forced to find alternative SNPs, so called proxies. This can be easily achieved in most cases.

3.4 PROBLEMS

Problem 3.1 Answer the following questions regarding genetic markers:

1. Which genetic markers are most commonly used nowadays?
2. How do these differ with regard to frequency in the human genome, informativity, mutation frequency, and ease of ascertainment?
3. Which functionality can be assigned to different genetic markers?

Problem 3.2 Consider two STRs A and B that have four alleles each. While the alleles of A are equifrequent, the alleles of B have the frequencies 0.8, 0.07, 0.07, and 0.06. For both markers, calculate HET , PIC , and LIC values.

Problem 3.3 Assume that you have, for a given STR, calculated a PIC of 0.6. What would be the corresponding LIC value?

URLs

DATABASE FOR SNPs (dbSNP)

This database serves as a central repository for single base nucleotide substitutions, deletion, and insertion polymorphisms. dbSNP takes the looser “variation” definition for SNPs, so there is no requirement or assumption about minimum allele frequency.

<http://www.ncbi.nlm.nih.gov/projects/SNP>

ENSEMBL

Ensembl is a comprehensive database of the genomes of various species ranging from primitive chordates to mammals, whose genomes have been sequenced to completion.

<http://www.ensembl.org>

HET CALCULATIONS

<http://linkage.rockefeller.edu/ott/linkutil.htm>

HUMAN GENOME DATABASE (GDB)

The GDB is the official central repository for genomic mapping data resulting from the Human Genome Initiative.

<http://www.gdb.org>

HUMAN GENOME VARIATION SOCIETY

The Human Genome Variation Society, which was formerly the Human Genome Organization (HUGO) Mutation Database Initiative, maintains a resource for mutations in human and other genomes containing links to numerous single-locus databases.

<http://www.genomic.unimelb.edu.au/mdi/>

LOCATION DATABASE

The genetic location database (LDB) gives locations of expressed sequences and polymorphic markers. Locations are obtained by integrating data of different types (genetic linkage maps, radiation hybrid maps, physical maps, cytogenetic data, and mouse homology) and constructing a single “summary” map.

http://cedar.genetics.soton.ac.uk/public_html/LDB2000/release1.html

MARSHFIELD WEB PAGE AND DATABASES

Information on human DNA polymorphisms is provided by the Center for Medical Genetics. It contains a list of over 200,000 confirmed and candidate human insertion/deletion polymorphisms. The list includes both microsatellites and SNPs. Information on the Marshfield Genetic Map is also available.

<http://research.marshfieldclinic.org/genetics>

NUCLEOTIDE DATABASE FROM THE NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI)

The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>

SNP CONSORTIUM LTD.

The SNP Consortium Ltd. was founded in 1999 with the aim of detecting a high number of SNPs evenly distributed throughout the genome.

<http://snp.cshl.org>

UNISTS—INTEGRATING MARKER AND MAPS

UniSTS is a comprehensive database of sequence-tagged sites (STSs) derived from STS-based maps and other experiments. STSs are defined by PCR primer pairs and are associated with additional information as genomic position, genes, and sequences.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=UniSTS>

4

Data Quality

Intensive data cleaning and appropriate quality control are fundamental to statistical analyses. Specifically, laboratory conditions can substantially affect the generated genotype data (see, Refs. [131, 228, 539]). For example, if only 1% of the genotypes are erroneous, 53–58% of the linkage information (see Chapters 3 and 7) for a trait locus can be lost [166], and error fractions as low as 3% can have serious effects on linkage disequilibrium analyses [7] and on association analyses (see Chapter 11). In association studies, errors can decrease the power for detecting associations and the probability for a false association (see Chapter 11; for an overview, see Ref. [539], also see [228, 657]). Furthermore, for genotype imputation [524] that often is the basis for meta-analyses of genome-wide association (GWA) studies, only single nucleotide polymorphisms (SNPs) of high quality should be used (see Chapter 14). Finally, when machine learning approaches or genome-wide haplotype analyses are used on GWA data [668, 763], all SNPs should be quality assured (see Chapter 13).

At the beginning of this chapter, we noted that the quality control procedures substantially changed in the last few years. Genetic epidemiologists usually obtained typed genotypes from a collaborating laboratory. Before the new array technologies were introduced, i.e., at the time of genome-wide linkage scans utilizing short tandem repeat markers (STRs) or candidate gene analyses with a handful SNPs, data were quality controlled in the laboratory. Thus, genotype data were transferred to genetic epidemiologists after extensive checks had been performed by people working in the laboratory.

Traditionally, the work flow was as follows. In the first step, people working in the laboratory checked the assays and optimized the reaction conditions. In the second step, when the actual genotyping was performed on all samples, they performed

an intensive quality control. Still, the data required further cleaning because some errors, e.g., can be identified only in the interplay of several genetic markers.

With the advent of the array technologies, the work-flow has substantially changed (for details see Chapter 14). And this also applies to the quality control procedures and the sharing of the workload. Specifically, genotypes, i.e., repeat lengths of STRs and nucleotide information for SNPs were and are transferred for STR-based linkage studies. Nowadays, signal intensities (for a detailed discussion, see Section 4.4) are now commonly distributed in high-throughput genotyping studies, including GWA studies or candidate gene chip studies. This means that the genetic epidemiologist does not only receive more data but also data of a different kind.

According to the differentiation between classical and recent quality control procedures, we discuss both approaches in this chapter. The first part of this chapter deals with family studies and the second with population-based studies. Specific to family studies are *pedigree errors* that are described in Section 4.1. *Genotyping errors* happen in both family and population-based studies, and they may partly be detected and removed in family studies, as will be detailed in Section 4.2. The situation is more difficult for studies, where no familial information is available. The traditional approach for identifying genotyping errors when genotypes from a few or just a single marker are available is investigating Hardy–Weinberg equilibrium (HWE; Section 4.3). Finally, we describe the various issues of standard quality control procedures in high-throughput genotyping studies (Section 4.4) and of investigating signal intensities (Section 4.5).

This chapter therefore deals with data quality issues, methods for detecting errors, and practical approaches for dealing with erroneous or missing genotypes.

4.1 HOW CAN PEDIGREE ERRORS BE DETECTED?

A *pedigree error* is the misclassification of a relationship between individuals and can have serious consequences for linkage studies. It can lead to false-positive evidence for linkage and can also reduce the power to detect linkage. Pedigree errors are frequent in animal breeding or plant genetics. However, they also occur in genetic epidemiology. In human studies, the causes usually are

- non-paternity,
- unknown adoptions, or
- sample swaps on the laboratory level, e.g., by DNA sample mislabeling.

Non-paternity usually results in half-sib pairs instead of full-sib pairs. Its frequency may depend on factors such as cultural background, social class, or the disease under study. Adoptions may be unknown to descendants or more distantly related family members. In the worst case, an unknown adoption may lead to a pedigree that is completely useless for further investigations.

Finally, sample swaps on the laboratory level are usually observed when working with single tubes. Here, sample swaps may occur regardless of the number or whether many or only few genetic markers have been genotyped. As soon as samples have

been transferred onto master plates, sample swaps involve the whole well plate and are thus unlikely. For the following, it is important to note that the likelihood for detection increases with the number of genotyped markers. For example, for a single genetic marker, a pedigree error would most likely be counted as a genotyping error if detected by checking for Mendelian inheritance (see Section 4.2). If, however, several markers show inconsistencies for a single individual, a pedigree error becomes likely.

The basic step for identifying pedigree errors is to examine whether the genotypes are congruent with Mendelian inheritance. However, if some individuals in the pedigree are not genotyped, then Mendelian errors may not result. In this case, statistical methods are required to test for deviation from the reported relationship. Most previous work has been concentrated on methods for detecting pedigree errors in sib-pairs (see, e.g., Ref. [98] for a comparison of algorithms). The reason for this focus is simple: many studies are based on sib pairs only. In addition, if parental data are unavailable in these studies, which is especially true for late-onset diseases, a strong bias can occur. Methods have, however, been extended to detect errors in other relationships [456, 651].

The statistical models required to rigorously address the problem of pedigree error detection are beyond the scope of this textbook. We refer to McPeck and colleagues [456, 651] for details. They provide methods for relationship estimation, hypothesis testing, and confidence interval construction, and they give exact likelihood calculations for all relationships considered. They describe the implementation of four different tests for ten different relationships in the software PREST (*Pedigree Relationship Statistical Test*) and ALTERTEST. The relationships tested in PREST are full-sib, half-sib, parent-offspring, grandparent-grandchild, avuncular, first cousin, half avuncular (Figure 4.1, left), half first cousin (Figure 4.1, middle) and half-sib plus first cousin (Figure 4.1, right), and unrelated. Furthermore, monozygotic (MZ) twins can be checked using ALTERTEST. Because the detection of pedigree errors increases with the number of genetic markers, PREST and ALTERTEST can easily take data from an STR genome-wide scan on autosomal chromosomes.

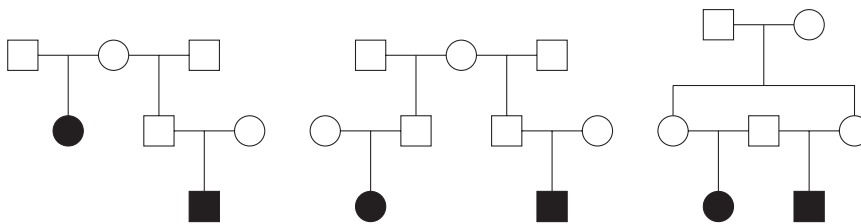


Fig. 4.1 Pedigree for a half-avuncular pair (left, filled in), for a half-first-cousin pair (middle, filled in) and for a half-sib-plus-first-cousin pair (right, filled in)

4.2 HOW CAN GENOTYPING ERRORS BE DETECTED IN FAMILY-BASED STUDIES?

It has been repeatedly reported that genotyping errors can increase the estimated recombination fraction between markers or between marker and disease locus (more generally, an inflation of the map distance for multiple markers, see Section 7.2.2.2). Also, they can enhance the type I error frequency for linkage and decrease the power to detect linkage. It is therefore important to know the frequency of genotyping errors, their reasons, and automatic approaches for their detection.

4.2.1 What is the frequency of genotyping errors?

Genotyping errors depend on many factors, including

- personnel,
- DNA degradation,
- marker type,
- specific marker conditions, e.g., buffer solution,
- genotyping hardware,
- the quality of the well plate,
- the specific chip,
- the use-by-date of the chip,
- the freezing of the chip after hybridization but prior to image reading,
- the genotype calling algorithm, and
- the quality control process.

When reporting genotyping errors, it is important to distinguish between *genotyping errors*, *repeatability errors*, and *reproducibility errors*. A genotyping error occurs if the observed marker genotype deviates from the true marker genotype. According to the International Organization for Standardization (ISO), *reproducibility* refers to genotyping with the same method on identical subjects in different laboratories with different personnel using different equipment [333, 332, also see ISO web site]. In contrast, *repeatability* refers to genotyping with the same method on identical subjects in the same laboratory by the same personnel using the same equipment within short intervals of time. In genetic epidemiology, this distinction is commonly not made. Here, errors are called reproducibility errors if two genotyping techniques are applied or if the same technology is used twice, irrespective of the time frame. However, the reader should be aware that the distinction between reproducibility and repeatability is important for diagnostic tests.

Cross-platform comparisons are of interest in some applications. If genotypes agree, the term “concordance rate” is often used in the literature although a *rate* is defined as a number of cases *in a period* divided by the average population size in that period. Therefore, the terms *concordance fraction* or *concordance proportion* would be more appropriate.

Genotyping error frequencies varied from 0.4% to 3% for STR markers [95, 96, 168, 696]. Dinucleotide and non-integer STRs have higher error frequencies than tri- and tetranucleotide STRs. This is the reason for the emphasis on tri- and tetranucleotide markers in several STR marker panels. With the use of improved technology, good genetic markers, and excellent quality control procedures, genotyping accuracy may be far above 99%. A reasonable estimate for accuracy fraction of genotypes in genome-wide scans is at least 99.5%.

With the advent of the microarray technology for SNPs, it is almost impossible to report valid and reliable error frequencies because there are too many factors affecting them. Frequencies have been reported in a few publications [286, 489, 501], and we summarize the current knowledge as follows. For the Affymetrix GeneChip Mapping 10K (SNP) array, genotype error frequencies have been estimated at about 0.1% [286], and the proportion of discordances was found to be $< 0.4\%$. For the Affymetrix Genome-Wide SNP 6.0 Array and cross-platform comparisons, concordance was $\geq 99.8\%$ [5, 501]. For the BeadArray technology used by Illumina, the repeatability was $\geq 99.99\%$, the genotype error frequency was 0.01%, and the frequency of Mendelian inconsistencies was 0.005% [489, also see Genizon web site].

Not all genotyping errors can be detected by Mendel checks (Section 4.2.4). Depending on the allele frequency, pedigree structure, genotyping system, and error model, detection fractions using Mendelian laws vary between 13% and 75% for SNPs [167, 241]. Furthermore, the fraction of mistypings consistent with Mendelian inheritance will increase with the more widespread use of SNPs and other diallelic markers. This problem is particularly acute in genome scans involving sib-pairs without parents [167]. As pointed out by Oliphant and colleagues [512], in linkage studies a 1% error fraction may double the number of samples required to achieve a given power of analysis. Unfortunately, research studies seldom undertake complete detection of genotyping errors. Furthermore, only a very few linkage and association models and programs explicitly allow inclusion of genotyping errors. Thus, most analysis programs will not run unless the data have no Mendelian errors. They must be removed in advance.

For applications, two situations having an impact on fractions of genotyping errors should be distinguished. In genome-wide scans, error frequencies are generally low as these markers have been chosen carefully because they give reliable results, are easier to call, and have high heterozygosity. If fine mapping is performed, which usually requires the establishment of new genetic markers, higher genotyping errors, up to 15%, have to be expected. If STR markers are to be used in a fine mapping project, tri- or tetranucleotide STR markers should be preferred over dinucleotide or non-integer STRs.

4.2.2 What are the reasons for genotyping errors?

The sources for mistypings are manifold, as is discussed in several papers [211, 539, 675, 696]. *Technological factors* for genotyping errors include the following:

- The occurrence of mutations at primer sites. In this case, certain alleles may not be amplified, termed *null alleles*. If an individual is heterozygous at the

respective marker and one allele is not amplified, the individuals will be falsely counted as homozygous for the other allele; hence, null alleles result in false homozygotes. The mutation rate is lower for SNPs (on the order of 10^{-7} to 10^{-9} per gamete per generation) than for STR markers (on the order of 10^{-3} to 10^{-5}).

- Bands running off the gel. This is specific for STR markers and also leads to null alleles.
- The preferential amplification of small alleles. In this case, termed *large allele dropout* or *short allele dominance*, the larger allele fails to amplify, again resulting in false homozygotes.
- Low template DNA concentrations, which may result in an allele failing to amplify because of stochastic sampling error.
- *Electrophoresis artifacts* (EA): Extra alleles may arise if high concentrations of polymerase chain reaction (PCR) products are capillary electrophoresed. This EA can lead to identification of a false heterozygote or to incorrect allele assignment in a heterozygote.
- *Stutter bands*: This is the most common problem for STR markers. Slippage during PCR amplification can produce additional stutter products that differ from the original template. This PCR artifact is therefore also termed *DNA polymerase slippage product*. Stutter bands make it difficult to discriminate between homozygotes and heterozygotes, as is illustrated in Figure 4.2. They usually appear as minor bands one repeat sequence below the true allele in electrophoresis.

Human causes for genotyping errors include

- unintentional duplication of samples,
- mishandling of samples in the process of DNA extraction or DNA amplification, and
- confusion in the data entry process if genotype data are entered by hand.

Although human causes lead to fewer systematic errors than technological cases, human factors should not be underestimated as reasons for genotyping errors. An example is a recent study in which human causes accounted for many of the typing errors [72]. Sample swaps, pipetting errors, or confusion in the data entry after scoring are common errors and will never be totally avoided when dealing with many individuals.

4.2.3 How can genotyping errors be detected using Mendel checks?

It was pointed out in Section 4.2.1 that Mendelian errors need to be removed prior to linkage and association analyses. Mendelian inheritance checks require genotype data from relatives. Hence, Mendelian error checking is not possible in population-based samples (see Section 4.3). It is also difficult or even impossible in sib-pair studies without parental genotypes.

In principle, Mendelian errors can be detected by visual inspection. However, checking can be done more efficiently with diagnostic software like PEDCHECK [511]. It investigates Mendelian inheritance errors on autosomes and sex chromosomes and gives detailed diagnostic information on the source of error by summarizing the detected genotyping errors by pedigree and by marker at the end of its output. If the number of genotyping errors is particularly large for a pedigree, then it is very likely that a pedigree error has caused most of the Mendelian genotyping errors (see Section 4.1).

PEDCHECK [511] handles one locus at a time so that it is not limited by the number of investigated markers. It works with four different error-checking algorithms to deal with different degrees of difficulty in the identification of an error. For example, to check raw laboratory data that might have numerous genotype-scoring or data entry errors, a likelihood-based check would be inefficient. It may offer no additional information if simple parent–offspring relations can reveal these errors. Likewise, if a large pedigree has a subtle error that simple parent–child relationships cannot identify, then a more powerful check is required. The first algorithm is probably the most important one for routine applications and therefore is described in more detail. It is termed nuclear family algorithm because it checks for inconsistencies between parents and offspring. An error is flagged if one or more of the following conditions is true:

- The alleles of a child and a parent are *incompatible*, i.e., they do not match Mendel’s laws (Chapter 2).
- The child is compatible with each parent separately but not when both parents are simultaneously considered.
- There are more than four alleles in a sibship.

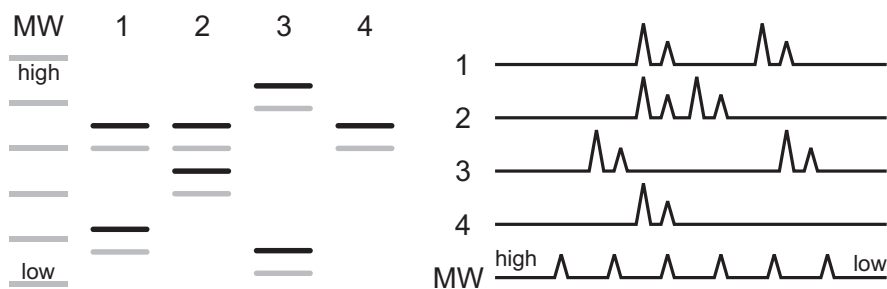


Fig. 4.2 Stylized examples of short tandem repeat (STR) data. Left: four sets of data were produced by gel electrophoresis, and the major (black) and stutter (gray) bands can be seen. MW: molecular weight standards. Right: these data were produced by analysis on an automated capillary electrophoresis-based DNA sequencer. The data are line graphs, with the location of each peak on the X-axis representing a different sized polymerase chain reaction (PCR) product and the height of each peak indicating the amount of PCR product. The major bands produce higher peaks than the stutter peaks. Figures have been adapted from www.bio.davidson.edu/courses/genomics/method/microsatellite.html with kind permission of A. Malcolm Campbell.

- There are more than three alleles in a sibship with a homozygous child.
- There are more than two alleles in a sibship with two different homozygotes among the sibs.
- An allele is out of bounds of any specified range.
- At an X-linked locus, a male is not coded as homozygous.
- An individual has only one allele defined in an autosomal system.

The last item is not an error, but most genetic epidemiological software packages expect both alleles to be specified. The nuclear family algorithm is appropriate as a first check, but it ignores loop information, i.e., interrelatedness of individuals. This information is utilized by the genotype elimination algorithm [398], the second level of error checking in PEDCHECK. The other two error-checking algorithms may be used to identify the source of the genotyping error more easily. A detailed description of all four different error-checking algorithms implemented in PEDCHECK can be found in Ref. [511].

For STRs, the proportion of genotyping errors is greatly reduced by simple Mendel checks. However, with this approach consistency of genotypes is investigated only per family. This is particularly suitable for linkage analyses (Chapter 5), but consistency across families and gels is required for association analyses. The latter requires the use of control DNA with known genotypes.

4.2.4 What are double recombinants, and how can they be checked?

As stated in Section 4.2.3, PEDCHECK performs single-locus checks. It thus cannot detect errors that may become apparent only by jointly investigating closely neighboring markers. Figure 4.3 illustrates this problem using five STR markers on one chromosome. For simplicity, we assume that parental haplotypes have been determined. The mother is heterozygous with alleles 1 and 2 at all markers. All 1 alleles form one haplotype, while all 2 alleles form the other. The situation is similar for the father. He is heterozygous with alleles 2 and 3; his haplotypes consist of the 2 and 3 alleles, respectively. As shown in Figure 4.3, the offspring has received the paternal haplotype consisting of 2 alleles, only. However, two recombinations must have occurred for the maternal haplotype: while the 1 allele is transmitted for the first two and the last two markers, a 2 allele has been genotyped for marker 3. Of course, no Mendelian error can be found. However, two recombinations on a single haplotype, termed *double recombinant*, are unlikely if the markers are closely neighboring.

If, for example, the recombination fraction θ_{23} between markers 2 and 3 is 0.1, and if θ_{34} is equal to 0.05, then the probability for two recombinations between markers 2 and 4 is $0.1 \cdot 0.05 = 0.005$ assuming independence of recombinations. However, recombinations do not occur independently, and this phenomenon is termed *interference*. During meiosis, the presence of one chiasma usually inhibits formation of a second chiasma nearby, resulting in *positive interference*. If interference is taken into account, the estimate for a double crossing over between markers 2 and 4 is approximately 0.0006 (Section 5.2.2), thus very unlikely. Considering also that

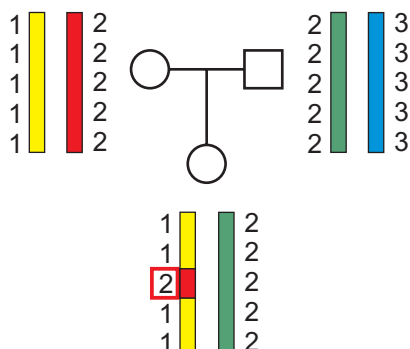


Fig. 4.3 Pedigree with three individuals where the parental haplotype has been determined. The pedigree has a genotyping error in the offspring for marker 3 that is consistent with Mendelian inheritance. The mistyping is detected by checking for double recombinants.

many genotyping errors lead to a false homozygote, a genotyping error at marker 3, for which the child appears homozygous, is more plausible.

Because it cannot be ruled out that a true double recombinant has occurred, even though the probability is slight, checks for double recombinants need to be carried out with caution. However, this search for mistyping errors can be very useful in fine mapping where the marker map is dense. If double recombination events are not detected, the genetic map expands [88]. These undetected double recombinants constitute more than one-quarter of all genotyping errors in fully typed nuclear families consisting of both parents and their offspring [167].

To avoid these excessive recombination events, a different type of genotyping error checking, different from Mendelian error checking, may be carried out. A common approach is to inspect pedigrees manually. This can be done in a customized way by plotting pedigrees with chromosome-wise marker information and the observed recombination events. If, for example, GENEHUNTER is utilized and the HAPLOTYPE option is turned on, the SCAN command reports one of the most likely inferences made regarding the haplotypes of the individuals in each pedigree. In addition, if the POSTSCRIPT OUTPUT option is on, the entire pedigree with haplotypes and recombinations indicated is drawn in a postscript file suitable for printing and displaying.

Instead of visually inspecting double recombinants, they may be assessed by a likelihood ratio approach [2]. Alternative software solutions for automated detection of double recombinants include MENDEL and SIMWALK2. However, these programs use a different approach for the identification of errors. Upon using a Bayes method, they report posterior probabilities of mistypings. These methods do not perform very well for small pedigrees [40] but can be recommended for larger pedigrees.

From a practical point of view, false homozygosity is the most commonly detected error. Though approaches to deal with erroneous genotypes have been proposed (see, e.g., [257]), Mendelian inheritance errors need to be solved in the first step for use in most computer packages. This may be done by carefully rereading genotypes or by re-genotyping. However, in many especially high-throughput applications, it is too expensive to manually reread or even re-genotype single individuals at an already genotyped marker. Nevertheless, genotypes should never be replaced with different

values without laboratory evidence. If an inconsistency cannot be resolved on the laboratory level, both alleles should be set to the missing value code, which is 0 in most genetic software packages.

Two approaches are, in principle, possible if inconsistencies that cannot be resolved are detected in a pedigree. One could eliminate all genotypes of a family at the critical marker. However, it seems to be preferential to remove only unlikely genotypes from individuals in a pedigree, if it is possible to identify the individual with the unlikely genotype, in order to keep as much information as possible for statistical analyses. One may proceed similarly with double recombinants. Though all these familial error-checking approaches are important, one must be aware that not all errors can be detected. Furthermore, the proportion of undetected errors will increase with the more widespread use of SNPs. For some specific genetic markers, mostly SNP markers, the observed fraction of erroneous genotypes is extremely high. One reason can be chemical reactions that do not work properly, e.g., when the ingredients are too old or inadequately stored, or other factors including DNA degradation. A standard approach in these situations is that the whole genetic marker is left out for further analyses.

In summary, algorithms and statistical tools should be used only to highlight pedigrees, individuals, genetic markers, or genotypes that need special consideration. Error-checking software should not resolve erroneous genotypes. Instead, software packages should automatically eliminate critical genotypes if they cannot be resolved in the lab for cost and time reasons. Despite these approaches, the inclusion of genotyping errors in the likelihood procedure is preferable to trying to detect and remove the errors.

4.3 HOW SHOULD GENOTYPING ERRORS BE CHECKED IN POPULATION-BASED STUDIES USING THE HARDY–WEINBERG EQUILIBRIUM?

As described in Section 4.2.2, genotyping errors often cause false homozygotes and hence a heterozygote deficiency. This may lead to deviations from Hardy–Weinberg equilibrium (HWE; see Chapter 2). It is therefore obvious that the traditional routine check for genotyping errors is to test for deviation from Hardy–Weinberg equilibrium. This approach has been and is used when a few or just a single marker are available. However, there are other reasons besides genotyping errors leading to a deviation from HWE (Section 4.3.1). In almost all applications, deviation from HWE is tested for SNPs (Section 4.3.2) and STRs (Section 4.3.3). The standard approaches test the null hypothesis that HWE is fulfilled. With a significant finding, one has proven that the genotype distribution does not conform with HWE. Several statistical approaches require, however, compatibility with HWE. One thus has to prove that HWE holds or deviates from HWE only to a small degree. To this end, we need to quantify this small degree of deviation from HWE by using appropriate measures of deviation

(Section 4.3.4). We finally derive a simple approach for establishing goodness-of-fit, i.e., compatibility with HWE (Section 4.3.5).

4.3.1 What are the causes of deviations from Hardy–Weinberg equilibrium?

Many test statistics have been proposed to test for deviation from HWE. However, great care is needed when using HWE testing as an indicator for mistypings. As explained in Chapter 2, deviations from HWE may arise from factors such as inbreeding caused by consanguinity, assortative, i.e., non-random mating, Wahlund effect caused by population stratification, or selection that may have a dependence of survival on genotypes. A summary of some potential causes of deviations from HWE are provided in Table 4.1.

Table 4.1 Some potential causes for deviations from Hardy–Weinberg equilibrium (HWE). Deviations from HWE lead to an excess of homozygosity, i.e., decrease in heterozygosity (HET), or an excess of heterozygosity.

Decrease in HET caused by	Increase in HET caused by
Selection against heterozygotes	Selection favoring heterozygotes
Inbreeding	Outbreeding
Positive assortative mating	Negative assortative mating
Null allele	Copy number variants
Wahlund effect	Amplification artifact of new alleles
Allele dropout in old samples	Misclassification of alleles at different loci in multigene families

Adapted from Ref. [291], Table 2.15.

Deviations from HWE can also occur if the probability of ascertainment depends on the genotype. Specifically, in case-control studies genotypes of cases may not be in HWE if genotypes are associated with different disease risks (the concept of association and case-control studies is described in detail in Chapter 11). In fact, Lee [403] has proposed scanning the genome for disease susceptibility gene(s) by testing for deviation from HWE (Hardy–Weinberg disequilibrium; HWD) in affected individuals (also see critical comments in Ref. [703]), and several colleagues proposed to incorporate an HWD measure in the association test [301, 621].

Because HWE may be violated in cases without being caused by genotyping errors, a common recommendation is to check only the control group for compatibility with HWE, with the hope that the controls might represent the general population well for a rare disease. However, when selection is likely to play an important role as, e.g., in infectious diseases such as malaria, deviation from HWE with excess heterozygosity in controls is plausible. Furthermore, if the entire population is in HWE, and a subset

of the population (the cases) that is not in HWE is removed, then the remaining subset (the controls) should also depart from HWE [727]. This deviation should, however, be lower in the controls compared to the cases because the prevalence of the disease typically is low. Therefore, it is difficult to judge whether HWE should be investigated in the entire sample or the controls only. Despite these arguments, HWE checking remains a reasonable approach to detect mistypings in population-based samples.

Repeated genotyping of the same probands is preferable over HWE testing. It is, however, costly, and it may reveal only genotyping errors that are caused by technical artifacts. It will not identify all forms of genotyping errors, such as errors caused by duplicates or triplicates of a region in the genome, contamination of DNA, or sample swap. Another possibility is to genotype a population-based sample in which HWE should hold. However, if deviation from HWE is observed in the population-based sample, it will not be possible to distinguish between genotyping error and violation of one of the assumptions necessary for HWE. Clearly, more options are needed to adequately identify genotyping errors in population-based samples using SNPs.

For STR markers, checking for HWD may be more reasonable than for SNP markers. First, most STR markers will be non-functional so that selection may be rare. Short allele dominance, stuttering, and null alleles yield deficiencies and excesses of particular genotypes. This is in contrast to allelic dropout, which is assumed to be largely independent of allele size [465]. Therefore, discrimination between deviations caused by non-panmixia and genotyping errors is possible in STR markers.

4.3.2 How can deviation from Hardy–Weinberg equilibrium be tested for single nucleotide polymorphisms?

The methods proposed for testing deviations from HWE are either large-sample goodness-of-fit tests such as Pearson’s χ^2 or the comparable likelihood ratio statistic. In order to avoid dependency on asymptotic properties, exact tests have been proposed as alternatives.

With the notation of Table 4.2, the standard asymptotic χ^2 test for deviations from HWE is the classical “observed minus expected” statistic [705]:

$$\chi_{HW}^2 = \frac{(n_{11} - n \cdot \hat{p}^2)^2}{n \cdot \hat{p}^2} + \frac{(n_{12} - n \cdot 2\hat{p}\hat{q})^2}{n \cdot 2\hat{p}\hat{q}} + \frac{(n_{22} - n \cdot \hat{q}^2)^2}{n \cdot \hat{q}^2} \quad (4.1)$$

Here, $\hat{p} = \hat{P}(A_1) = (2n_{11} + n_{12})/(2n)$ is the estimated allele frequency of A_1 , and $\hat{q} = 1 - \hat{p}$ (see Sections 2.4 and 3.1). Under the null hypothesis of no deviation from HWE, χ_{HW}^2 is asymptotically distributed as a central χ^2 . Because the contingency table has three cells, χ_{HW}^2 has 2 degrees of freedom (d.f.) if $\hat{p}$ is not estimated from the current data set but is specified in advance. However, it does not seem meaningful to check for deviation from HWE with the allele frequency $P(A_1) = p$ estimated from an external data set or even specified in advance without any knowledge. Instead, p

Table 4.2 Contingency table summarizing the results of genotyping n individuals with respect to a single nucleotide polymorphism with alleles A_1 and A_2 .

	Genotype			Total
	A_1A_1	A_1A_2	A_2A_2	
Observed count	n_{11}	n_{12}	n_{22}	n
Population frequency	p_{11}	p_{12}	p_{22}	1

is usually estimated from the current data set so that the d.f. have to be reduced by 1 because of the estimated parameter.

The asymptotic test statistic is liberal for small sample sizes and/or for a small genotype frequency. Although several corrections have been made, we recommend relying on an exact test [429] or a Monte-Carlo method. The conventional Monte-Carlo approach is as follows [268]. If HWE holds, then a sample of n genotypes can be interpreted as n random union of two alleles. If the alleles are permuted and arranged into n pairs, then the n pairs can be regarded as another random sample from the same population. This leads to the following algorithm to check for deviations from HWE:

Algorithm 4.1.

1. Compute χ_{HW}^2 using the original n data
2. Set counter $C = 0$
3. For a specified number N of permutations,
 - (a) Permute the alleles and arrange them into n new pairs
 - (b) Compute χ_{HW}^2 using the permuted data
 - (c) If the test statistic from the permuted data is at least as large as χ_{HW}^2 using the original data, add 1 to C
4. The Monte-Carlo p -value is C/N

The final question is how many Monte-Carlo permutations have to be carried out to ensure a prespecified precision of the estimated p -value. Using the central limit theorem, we can assume asymptotic normality of the estimated p -value. Standard text books (see, e.g., [386], p. 3893) show that the number of permutations N required to estimate a p -value assumed to be less than π with precision e at a confidence level of $1 - \alpha$ is given by

$$N = \left(\frac{z_{1-\alpha/2} (\pi(1-\pi))^{1/2}}{e} \right)^2. \quad (4.2)$$

$\pi(1-\pi)$ takes its maximum at $\pi = \frac{1}{2}$ so that for practical purposes Eq. (4.2) is reduced to

$$N = \left(\frac{z_{1-\alpha/2}}{2e} \right)^2 \quad (4.3)$$

if no a priori information about the p -value is available.

Example 4.1. We want to estimate a p -value with a precision of $e = 0.01$ with 99% confidence. With the $1 - \alpha/2 = 0.995$ quantile of the standard normal distribution being 2.576, approximately

$$N = \left(\frac{2.576}{2 \cdot 0.01} \right)^2 \approx 16,589.44 \approx 17,000$$

simulations are required.

Although approximately 17,000 permutations are required for a precision of 1% in the general case, only 4000 permutations are required when the p -value is around 0.05. In fact, with 10,000 permutations the precision is around $e = 0.0058$ when the p -value is close to 5%.

Equation (4.3) also shows that 100 times more simulations have to be conducted if the precision is to be increased by one digit. Therefore, to save computing time, one should increase the number of permutations as required by the observed p -value in several steps and should start with about 17,000 simulations according to Example 4.1. For example, if the p -value is 10% and above, no further simulations are required as the result is not significant. If, however, the observed p -value is low, the number of simulations should be increased to the required precision.

Instead of permuting alleles, an alternative Monte–Carlo approach may be taken by generating new genotypes using the estimated allele frequency $\hat{p}$. This results in the following algorithm:

Algorithm 4.2.

1. Estimate the allele frequency $\hat{p}$ using the original n data
2. Compute χ_{HW}^2 using the original n data
3. Set counter $C = 0$
4. For a specified number N of permutations,
 - (a) Generate n new genotypes under HWE using $\hat{p}$
 - (b) Compute χ_{HW}^2 using the generated data
 - (c) If the test statistic from the generated data is at least as large as χ_{HW}^2 using the original data, add 1 to C
5. The Monte–Carlo p -value is C/N

This algorithm is implemented, e.g., in HWG–TEST.

We illustrate the application of testing for deviations from HWE by using two previously published SNP data sets.

Example 4.2. The first data set has been taken from Reich and colleagues [557], who conducted a candidate gene case-control study on psoriasis, which is a complex inflammatory skin disease. Genotype frequencies of controls are 238 for genotype GG, 103 for genotype AG, and 4 for genotype AA for the G-308A promotor polymorphism in the gene encoding for tumor necrosis factor- α . The estimated allele frequency of the G allele is approximately $0.8391 \approx \frac{2 \cdot 238 + 103}{2 \cdot 345}$. The expected frequencies under HWE are $242.93 \approx 0.84^2 \cdot 345$, $93.14 = 2 \cdot 0.84 \cdot 0.16 \cdot 345$, and $8.93 = 0.16^2 \cdot 345$ for the GG, AG, and AA genotypes, respectively. As can be seen

by comparing observed and expected genotype frequencies, most of the discrepancy is contributed by a deficiency of AA homozygotes with a preference for the AG heterozygotes. The contributions of each genotype to the χ^2 goodness-of-fit test of Eq. (4.1) can be calculated to be 0.10, 1.04, and 2.72 for the respective genotypes. The test statistic is 3.86 and has an asymptotic p -value of 0.0494, which is below the 5%-test level. With the Monte–Carlo approach of Algorithm 4.2, the p -value is 0.071 with 100,000 replications.

This result of a weak deviation from HWE could well be an argument for existence of genotyping errors. However, deviation from HWE at the TNFA-308 locus with under-representation of TNFA-308A homozygotes in the control group has been observed previously and has been interpreted as evidence that these controls are not a representative population sample. Instead, it has been demonstrated that carriage of TNFA-308A is associated with fatal outcome of common infectious diseases, including meningococcal disease in children. These results point to a selection favoring heterozygotes as a possible explanation for this phenomenon.

As a second example, we take data from the same study. Results for the IL1B+3953 polymorphism are given by 189 and 24 for the two homozygous TT and CC genotypes and 132 for the heterozygous CT genotype, respectively. The χ^2 test statistic is 0.02 with an asymptotic p -value of 0.8842. The Monte–Carlo p -value utilizing Algorithm 4.2 gives 0.89 with 17,000 replications. Thus, no deviation from HWE is detected.

4.3.3 How can deviation from Hardy–Weinberg equilibrium be investigated for short tandem repeat markers?

The easiest approach to investigate deviations from HWE in STRs is a visual inspection using Vasarely charts [435]. The chart, which can be produced using VASARELY, is a simple grid of square cells, with the first allele defining the rows and the second allele defining the columns. At the intersection between a row and a column, the expected genotype frequency is represented by shading the full cell. The measured frequency is represented by shading a circle, centered on the cell with a diameter that gives it a comparable visual weight to the remaining cell (about 75% of the cell size). In the absence of null alleles or other problems distorting allele frequencies, the measured and expected genotype frequencies are nearly equal, so the circles have the same shade as their surrounding squares and therefore are not seen as circles (see Figure 4.4 for illustration). Increased homozygosity appears as dark circles along the identity diagonal (alleles D and K in Figure 4.5), with light circles elsewhere.

To formally test deviations from HWE, the χ^2 goodness-of-fit tests described in the previous section may be applied to genetic markers with multiple alleles. However, the contingency tables will be very sparse for a large number of different alleles. Therefore, the Monte–Carlo permutation approach as described in Algorithm 4.1 or an exact test should be preferred over tests relying on the asymptotic distribution.

An alternative to check for mistypings in STR markers uses the often observed *deficiency of heterozygosity* [112, 214]. Let p_i be the frequency of allele A_i , $i = 1, \dots, m$



Fig. 4.4 Vasarely chart for marker D3S3726 with 15 alleles showing no deviation from Hardy–Weinberg equilibrium. The data were kindly provided by Jochen Hampe.

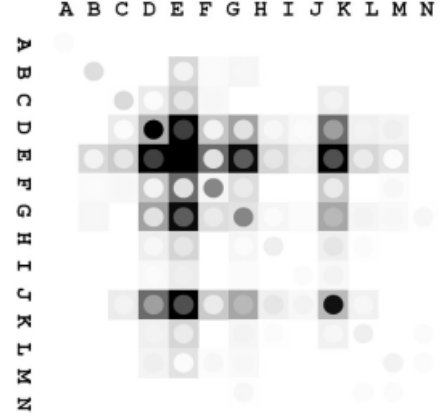


Fig. 4.5 Vasarely chart for marker D3S1293 with 14 alleles showing a deviation from Hardy–Weinberg equilibrium. The data were kindly provided by Jochen Hampe.

for an m -allelic STR. Furthermore, let n_{ij} , $j \geq i$, $i = 1, \dots, m$ be the number of observed genotypes $A_i A_j$. If there is no excess of homozygosity, the probability for an individual to be heterozygous is $\sum_{i \neq j} p_i p_j$, and the number of heterozygous individuals is binomially distributed. One can obtain a practical heterozygote deficiency test, abbreviated as *DH test*, by comparing the expected number $n \sum_{i \neq j} p_i p_j$ and the observed number $\sum_{i \neq j} n_{ij}$ of heterozygous individuals. Instead of relying on the binomial distribution, a Monte–Carlo permutation approach can be carried out using the permutation scheme of Algorithm 4.1, yielding the following algorithm:

Algorithm 4.3.

1. Compute the number of homozygous individuals Hom_O
2. Set counter $C = 0$
3. For a specified number N of permutations,
 - (a) Permute the alleles and arrange them into new pairs
 - (b) Compute the number of homozygous individuals in the permuted data Hom_P
 - (c) If $Hom_O \geq Hom_P$, add 1 to C
4. The Monte–Carlo p -value is C/N

Obviously, this approach is computationally intensive, especially compared to its binomial version. Own simulation studies have shown that the DH test has greater power to detect mistypings in STR markers than does the Monte–Carlo χ^2 test.

If one is willing to assume that the heterozygote deficiency is caused by null alleles, not by other genotyping errors or other factors leading to deviations from HWE, one can estimate the frequency of null alleles from an apparent heterozygote deficiency. Different methods have been proposed for this purpose, depending on whether some samples failed to amplify, and on whether these non-amplified samples

represent null homozygotes or are caused by insufficient DNA or other factors such as DNA degradation (for details, see Refs. [90, 112]).

Finally, it is important to note that the tests to check for deviations from HWE can result in inflated type I error rates when applied to samples with related individuals. Valid test statistics for this case are described in Ref. [81].

4.3.4 How can deviation from Hardy-Weinberg equilibrium be measured for single nucleotide polymorphisms?

In the previous sections, we presented approaches that prove deviations from HWE by testing lack of fit. Thus, a significant deviation from HWE is interpreted as an indication for genotyping errors. However, not detecting a significant deviation does not evince compatibility with HWE. Hence, instead of searching for lack of fit, one might be interested in establishing goodness-of-fit, i.e., compatibility with HWE. Before we can describe (Section 4.3.5) how compatibility with HWE can be investigated, we need to be able to quantify the amount of deviation from HWE.

One approach to measuring deviation from HWE is based on the inbreeding model [291, 434]. As pointed out by Ayres and Balding [37], it may be appropriate if inbreeding is the main cause of deviation from HWE. For this, the *inbreeding coefficient*, f is the probability that both alleles in an individual are identical by descent (IBD). This means that they are derived from one particular allele possessed by a common ancestor. f is sometimes written as F_{IS} according to Wright [732], and this notation refers to the relation of alleles within individuals I relative to a subpopulation S . It is therefore also termed *within-population inbreeding coefficient*. If inbreeding is present, then the probability is F that the second allele drawn from a population is the same as the first allele. If the first allele A_1 has a probability of $P(A_1) = p$ to be drawn, and the probability that the second allele is the same is f , then the probability of getting two identical A_1 alleles is $p \cdot f$.

There further is the chance that a subject has two A_1 alleles that are not IBD but just identical by state (IBS). With p^2 being the probability that two independent consecutive drawings give two A_1 alleles and $1 - f$ being the probability that the second allele is not IBD, we get $p_{11} = P(A_1 A_1) = p^2(1 - f) + pf = p^2 + fpq$ as frequency of genotype $A_1 A_1$. Likewise, with the notation $P(A_2) = q$ we obtain $p_{22} = q^2 + fpq$ as probability for genotype $A_2 A_2$ and $\text{Het} = P(A_1 A_2) = p_{12} = 2pq - 2fpq$ for the heterozygous genotype $A_1 A_2$. Because both homozygous genotype frequencies p_{11} and p_{22} are not negative, one can derive the restrictions $f > -\frac{p}{(1-p)}$ and $f > -\frac{(1-p)}{p}$. In addition, $f \leq 1$ because $p_{12} \geq 0$.

The inbreeding coefficient as defined here is a probability, and $f = 1$ is its maximum. In this case, $p_{11} = p$, $p_{12} = 0$, and $p_{22} = q$. If $f = 0$, only homozygous subjects are available and they occur in the proportions of the allele frequencies. The heterozygote frequency can be written as $p_{12} = 2pq(1 - f)$. If we use f_e for denoting the expected heterozygosity under HWE, and the heterozygosity in subjects with inbreeding f by H_f , then $H_f = H_e(1 - f)$. Subsequently, $f = 1 - H_f/H_e$.

f may thus be used as a measure of deviation from HWE. It can be estimated by maximum likelihood (ML) as ([705], p. 65)

$$\hat{f} = 1 - \frac{2n_{12}n}{(2n_{11} + n_{12})(n_{12} + 2n_{22})} = 1 - \frac{n_{12}}{2n\hat{p}\hat{q}}.$$

$\hat{f}$ is only asymptotically unbiased because its expected value is $-\frac{1}{2n-1}$ if $f = 0$. The approximate variance of $\hat{f}$ is given by

$$\mathbb{V}ar(\hat{f}) = \frac{1}{n}(1-f)^2(1-2f) + \frac{f(1-f)(2-f)}{2np(1-p)},$$

where p is the allele frequency of A_1 . An estimator of $\mathbb{V}ar(\hat{f})$ is readily obtained by replacing f and p by their estimators $\hat{f}$ and $\hat{p} = (2n_{11} + n_{12})/(2n) = \hat{p}_{11} + \hat{p}_{12}/2$.

A second measure of HWD has been proposed by Feder et al. [205]. They consider the *excess homozygosity*

$$F = \frac{H_f - H_e}{1 - H_e}$$

at an SNP. In fact, this is a measure of the inbreeding coefficient that is rescaled and normalized by $1 - H_e$, and it thus lacks adequate statistical properties. In fact, for deriving a test statistic or an asymptotic confidence interval, one would need to transform F back to an unrestricted range. Thus, the seeming advantage of using a rescaled measure over the inbreeding coefficient f is artificial.

A natural measure of HWD is the *disequilibrium coefficient* $\mathcal{D}_{A_1}$ which measures the difference between the observed and the expected genotype frequencies

$$\mathcal{D}_{A_1} = p_{11} - p^2 = p_{22} - q^2 = -\frac{1}{2}p_{12} + pq.$$

The expected value and variance of its estimator $\hat{\mathcal{D}}_{A_1}$ are given by

$$\begin{aligned} \mathbb{E}(\hat{\mathcal{D}}_{A_1}) &= \mathcal{D}_{A_1} - \frac{1}{2n}(pq + \mathcal{D}_{A_1}) \quad \text{and} \\ \mathbb{V}ar(\hat{\mathcal{D}}_{A_1}) &= \frac{1}{n}(p^2(1-p)^2 + (1-2p)^2\mathcal{D}_{A_1} - \mathcal{D}_{A_1}^2). \end{aligned}$$

The ML estimator of $\mathcal{D}_{A_1}$ thus is only asymptotically unbiased.

A HWD measure that is related to $\mathcal{D}_{A_1}$ is Jiang's J

$$J = \frac{p_{11} - p^2 + p_{22} - q^2}{p^2},$$

which has been proposed in the context of fine mapping [342]. It is thus not intended to be used to measure HWD for judging genotyping errors.

After the introduction of $\mathcal{D}_{A_1}$ and subsequent measures, we can note that deviation from HWE does not have an effect on the point estimator $\hat{p} = \hat{P}(A_1) = \hat{p}_{11} + \hat{p}_{12}/2$

for the allele frequency p . However, the variance $\mathbb{V}ar(\hat{p})$ is affected. Specifically, under HWE, $\mathbb{V}ar(\hat{p})$ is given by

$$\mathbb{V}ar(\hat{p}) = \frac{p(1-p)}{2n}, \quad (4.4)$$

while it equals

$$\mathbb{V}ar(\hat{p}) = \frac{p(1-p)}{2n} + \frac{p_{11} - p^2}{2n} = \frac{p(1-p)}{2n} + \frac{\mathcal{D}_{A_1}}{2n} \quad (4.5)$$

when there is deviation from HWE (Problem 4.8; see [460]). Because the variance is important for statistical testing, especially in association testing, when frequencies of cases and controls are compared (see Section 11.2), deviation from HWE may have a strong effect on the validity of allele-based association tests.

Now, we return to measures for HWD. An approach different from the previous ones has been followed by Pereira and Rogatko [530] and Lindley [420]. They considered the ratio

$$\gamma = \frac{4p_{11}p_{22}}{p_{12}^2}$$

as a measure of deviation from HWE. In fact, under HWE, $\gamma = 1$. If there is a heterozygote deficiency, i.e., excess homozygosity, $\gamma > 1$, and if there is an excess heterozygosity, $\gamma < 1$. Lindley has based all subsequent calculations on

$$\alpha = \frac{1}{2} \ln \frac{4p_{11}p_{22}}{p_{12}^2} = \ln \frac{2\sqrt{p_{11}p_{22}}}{p_{12}} = \ln 2 + \frac{1}{2} \ln p_{11} + \frac{1}{2} \ln p_{22} - \ln p_{12}, \quad (4.6)$$

which allows to consider genotyping frequencies additively on the logarithmic basis. Alternatives to γ and α have been proposed from a Bayesian perspective. For example, Verdinelli ([420], p. 320) discussed considering the difference $\text{HWD}_{\text{diff}} = 4p_{11}p_{22} - p_{12}^2$ rather than the ratio $\gamma = 4p_{11}p_{22}/p_{12}^2$.

γ and α are preferable over HWD_{diff} from a statistical perspective because it is a simple monotone transformation of the parameter of the conditional distribution for the number n_{11} of homozygous A_1A_1 subjects given the total number of A_1 alleles. This conditional distribution has been used by several authors, see, e.g., [182]. Because the conditional distribution belongs to the linear exponential family, many standard results can be utilized including those for statistical testing and parameter estimation.

In the next section, we will discuss how compatibility with HWE can be established using a confidence interval-based approach. For this, we will utilize a simple monotone transformation of γ and α . Specifically, we consider

$$\omega = \frac{p_{12}}{2\sqrt{p_{11}p_{22}}} = \gamma^{-1/2} \quad (4.7)$$

as a measure of HWD [711]. If $\omega > 1$, there is a *relative excess heterozygosity (REH)*, and for $\omega < 1$ we see a heterozygote deficiency, i.e., excess homozygosity.

Again, ω , γ , and α have simple interpretations both genetically and statistically and are therefore appealing for use in practice.

4.3.5 How can compatibility with Hardy–Weinberg equilibrium be tested for single nucleotide polymorphisms?

In the previous sections, we presented approaches that prove deviations from HWE by testing lack of fit. Hence, instead of searching for lack of fit, one might be interested in establishing goodness-of-fit, i.e., compatibility with HWE. Wellek [710] proposed two different statistical approaches that establish sufficiently good agreement of the genotype distribution with HWE. His approach follows the basic principles of equivalence testing. Thus, an indifference zone is first constructed, consisting of those genotype distributions that are sufficiently close to a distribution exactly fulfilling HWE. Second, a statistical test is derived meeting the pre-specified criteria.

Equivalence testing has been well developed in classical biostatistics where two therapies are evaluated with respect to their efficacy and/or safety. In clinical trials, two different terms are used, i.e., *non-inferiority* and *true equivalence*. Non-inferiority in the context of clinical trials means that the active treatment to be evaluated is only irrelevantly inferior to the standard treatment. For example, the active treatment should only show an irrelevantly higher number of adverse events compared to the standard treatment. In case of “true” equivalence, the irrelevant inferiority is considered in both directions, e.g., neither a too strong nor a too weak lowering of the blood pressure would be acceptable (two-sided question). “True” equivalence can also be of interest when there is no common standard therapy but two active treatments that have not been compared. Here, it can be of interest to investigate whether both therapies do not differ except for a negligible difference. As in the blood pressure example, a two-sided question is investigated here. All three scenarios have similar implications for the basic statistical test to be performed. And they are therefore typically summarized under the term *equivalence* or *equivalence studies*, except when specific aspects are considered.

To repeat, for equivalence testing it is important to define a region of acceptable indifference. In our case, this means that we need to define a zone around the Hardy–Weinberg curve in the de Finetti triangle that represents an area of acceptable deviation from HWE. This indifference zone is used for formulating the statistical hypothesis. Deviations from the indifference zone are possible in both directions. For use in practice, a series of different statistical approaches have been developed for proving equivalence. The ideas of equivalence testing and the different statistical procedures are described in detail in the excellent textbook by Wellek [709].

In brief, two different concepts are available for establishing equivalence. The first is through a statistical test, and this approach for establishing HWE has been followed by Wellek [710]. More guidelines on the biostatistics area favor the approach of establishing equivalence through the use of confidence intervals [331], and this approach is quite intuitive. We have therefore decided to present this simpler approach that utilizes a confidence interval for the parameter ω here. Because ω is bounded to the left, it is more promising to use $\ln \omega$ for deriving a confidence interval. It can be

shown (Problem 4.9) that $\ln \hat{\omega}$ is asymptotically normally distributed with mean $\ln \omega$ and variance

$$\sigma_{\ln \hat{\omega}}^2 = \frac{1}{n} \left(\frac{1 - p_{12}}{4p_{11}p_{22}} + \frac{1}{p_{12}} \right). \quad (4.8)$$

$\ln \omega$ and its variance can be estimated by plugging in the observed genotype frequencies. An asymptotic confidence interval for $\ln \omega$ is therefore given by

$$\begin{aligned} \underline{C}_{1-\alpha}(\ln \hat{\omega}) &= \ln \hat{\omega} - z \sqrt{\frac{1}{n} \left(\frac{1 - \hat{p}_{12}}{4\hat{p}_{11}\hat{p}_{22}} + \frac{1}{\hat{p}_{12}} \right)}, \\ \overline{C}_{1-\alpha}(\ln \hat{\omega}) &= \ln \hat{\omega} + z \sqrt{\frac{1}{n} \left(\frac{1 - \hat{p}_{12}}{4\hat{p}_{11}\hat{p}_{22}} + \frac{1}{\hat{p}_{12}} \right)}, \end{aligned}$$

where $\underline{C}$ and $\overline{C}$ denote the lower and upper bounds, respectively. Furthermore, z is the appropriate quantile from the standard normal distribution, and its choice will be explained below. The backtransformation to the original scale yields

$$\begin{aligned} \underline{C}_{1-\alpha}(\hat{\omega}) &= \hat{\omega} / \exp \left(z \sqrt{\frac{1}{n} \left(\frac{1 - \hat{p}_{12}}{4\hat{p}_{11}\hat{p}_{22}} + \frac{1}{\hat{p}_{12}} \right)} \right), \\ \overline{C}_{1-\alpha}(\hat{\omega}) &= \hat{\omega} \cdot \exp \left(z \sqrt{\frac{1}{n} \left(\frac{1 - \hat{p}_{12}}{4\hat{p}_{11}\hat{p}_{22}} + \frac{1}{\hat{p}_{12}} \right)} \right). \end{aligned}$$

These confidence intervals can be used for both significance testing and equivalence testing. Specifically, for the standard approach of testing lack of fit, i.e., the approach described in the previous section, the null hypothesis of HWE can be rejected if and only if the confidence interval fails to include unity. Here, the halved quantile $z = z_{1-\alpha/2}$ needs to be used. This procedure is asymptotically valid at an α -test level. If the null hypothesis is rejected, it has been proven that the distribution deviates from HWE.

For establishing goodness-of-fit, i.e., rejecting the null hypothesis that the distribution is not in HWE, the confidence interval can be used as well. This requires pre-specification of a fixed region $(1 - \varepsilon_1, 1 + \varepsilon_2)$ of acceptable deviations of the true value of ω from unity. This interval corresponds to the hypothesis $H: \omega \leq 1 - \varepsilon_1 \vee \omega \geq 1 + \varepsilon_2$. Given the values of the equivalence margins, the principle of confidence interval inclusion allows us to proceed as follows: the null hypothesis of relevant deviations from HWE is rejected in favor of the alternative hypothesis $1 - \varepsilon_1 < \omega < 1 + \varepsilon_2$ of equivalence with a distribution satisfying HWE if and only if the confidence interval is a sub-interval of the interval $[1 - \varepsilon_1, 1 + \varepsilon_2]$. In order to ensure that the test based on this decision rule is likewise asymptotically valid at level α , the confidence interval is used with quantile $z = z_{1-\alpha}$, and not the traditional two-sided $z_{1-\alpha/2}$.

For application, it may be more convenient to display the rejection region of the equivalence test based on ω in a de Finetti triangle instead of the boundary. To this end, given a total sample size n , the allele frequency p may be varied and is displayed

on the x -axis, and the heterozygote frequency p_{12} on the y -axis. This requires the confidence bounds for p_{12} , and these are given by

$$\begin{aligned}\underline{C}_{1-\alpha}(p_{12}) &= \ln \hat{p}_{12} - \frac{1}{2} \left(\ln(\hat{p} - \frac{\hat{p}_{12}}{2}) + \ln(1 - \hat{p} - \frac{\hat{p}_{12}}{2}) \right) \\ &\quad - z \sqrt{\frac{1}{n} \left(\frac{1 - \hat{p}_{12}}{4(\hat{p} - \hat{p}_{12}/2)(1 - \hat{p} - \hat{p}_{12}/2)} + \frac{1}{\hat{p}_{12}} \right)}, \\ \overline{C}_{1-\alpha}(p_{12}) &= \ln \hat{p}_{12} - \frac{1}{2} \left(\ln(\hat{p} - \frac{\hat{p}_{12}}{2}) + \ln(1 - \hat{p} - \frac{\hat{p}_{12}}{2}) \right) \\ &\quad + z \sqrt{\frac{1}{n} \left(\frac{1 - \hat{p}_{12}}{4(\hat{p} - \hat{p}_{12}/2)(1 - \hat{p} - \hat{p}_{12}/2)} + \frac{1}{\hat{p}_{12}} \right)}.\end{aligned}$$

This confidence interval can be used for significance testing using the heterozygote frequency from a sample of size n . Wellek et al. [711] have also described in detail how the confidence interval can be transformed in a rejection region given a pre-specified sample size. The interested reader is deferred to the original publication for this issue.

Although a positive result of any equivalence testing procedure is scientifically more satisfactory when smaller equivalence margins have been chosen, it is not reasonable to narrow the equivalence region to an extent that an unacceptably large proportion of SNPs fail to be compatible with HWE although they perfectly satisfy the model. These general considerations suggest to choose the equivalence margin as small as we can do under the following restrictions:

1. *Power*: Over the whole range of values of the minor allele frequency (MAF) to be taken into account, the probability of rejecting lack-of-fit under a genotype distribution that exactly conforms to the model must not be smaller than 90%.
2. *MAF*: The power to detect an association in a case-control or cohort study is low for a low MAF. In GWA studies, the proportion of SNPs with $\text{MAF} < 0.1$ associated with a disease is extremely low (see Chapter 11, GWA DATABASE of [343]). For SNPs defined eligible for inclusion in association studies, we therefore assume $0.1 \leq \text{MAF} \leq 0.5$, corresponding to $0.01 \leq p_{11} \leq 0.25$ under HWE.
3. *Sample size*: Current GWA studies have up to 3000 control subjects. The only exception are studies from Iceland, where more than 10000 subjects served as controls for a variety of investigated diseases. Therefore, the required sample size should not exceed 3000.

The solution can be easily found using a closed sample size formula. Suppose that the power to detect perfect agreement with HWE in the asymptotic test at nominal level α using the equivalence margins $\pm \varepsilon$ to $\ln \omega$ is to be $1 - \beta$. Then, it follows (Problem 4.10) that the sample size required to satisfy this condition is approximately given by

$$n = \frac{(z_{1-\beta/2} + z_{1-\alpha})^2}{(2\sqrt{p_{11}}(1 - \sqrt{p_{11}}))^2 \varepsilon^2}. \quad (4.9)$$

This equation is decreasing on the interval $]0, 1/2]$. Hence, the smallest equivalence margin leading to a sample size of $n = 3000$ for SNPs satisfying $0.1 \leq \text{MAF} \leq 0.5$, is calculated at $p = 0.1$, and this is

$$\varepsilon = \frac{z_{1-\beta/2} + z_{1-\alpha}}{2 \cdot 0.1 \cdot 0.9 \cdot \sqrt{3000}} = \frac{2 \cdot z_{0.95}}{2 \cdot 0.09 \cdot \sqrt{3000}} = 0.3337.$$

Because ε operates on $\ln \omega$, i.e., $-0.3337 < \ln \omega < 0.3337$, it needs to be transformed back to the original scale and yields an equivalence margin of $e^{0.3337} - 1 = 0.3961$ for ω itself. Therefore, we propose $\varepsilon_0 = 0.4$ as equivalence margin. This margin can also be justified without taking results on power into account using the arguments given by Ref. [709], Section 1.5.

Figure 4.6 displays the de Finetti triangle with the rejection boundaries using $1 - \varepsilon_1 = 1/(1 + \varepsilon_0)$, $1 + \varepsilon_2 = 1 + \varepsilon_0$ with $\varepsilon_0 = 0.4$. Figure 4.7 presents the rejection regions in the de Finetti triangle for the equivalence test using varying sample sizes of 100, 200, 500, 1000, 1500, 2000, and 3000 at the 5% test level and the equivalence margin $\varepsilon_0 = 0.4$. It is important to note that the rejection regions increase with increasing sample size. This is in contrast to the standard lack-of-fit test, where the rejection regions decrease with sample size [250]. Below, we illustrate the use of the equivalence test using the data from Example 4.2.

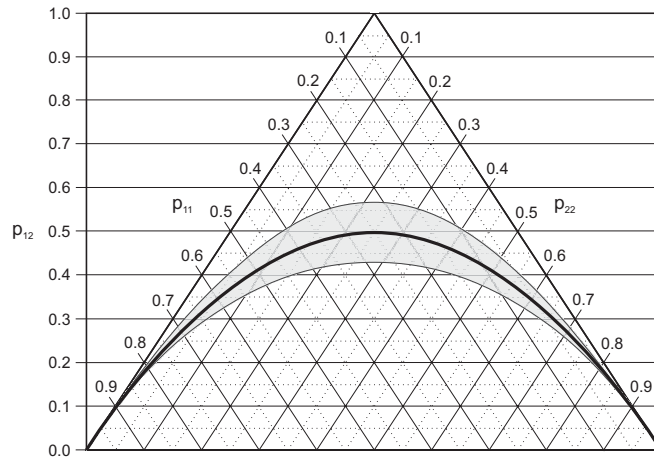


Fig. 4.6 De Finetti triangle representing genotype distributions. Each point in the triangle corresponds to a specific genotype distribution. Genotype distributions strictly fulfilling Hardy–Weinberg equilibrium (HWE) correspond to the black solid line. The gray band represents an indifference zone where deviation from HWE is not too strong. The parameter ε determining the width of the indifference zone is user defined. We argue that $1/1.4 < \omega < 1.4$ is an appropriate choice for this indifference zone.

Example 4.3. The data from Example 4.2 showed deviation from HWE for the TNFA-308A>G. Therefore, equivalence with HWE can be rejected in this case. This can also be seen from the two-sided 95% confidence interval for ω . First, with

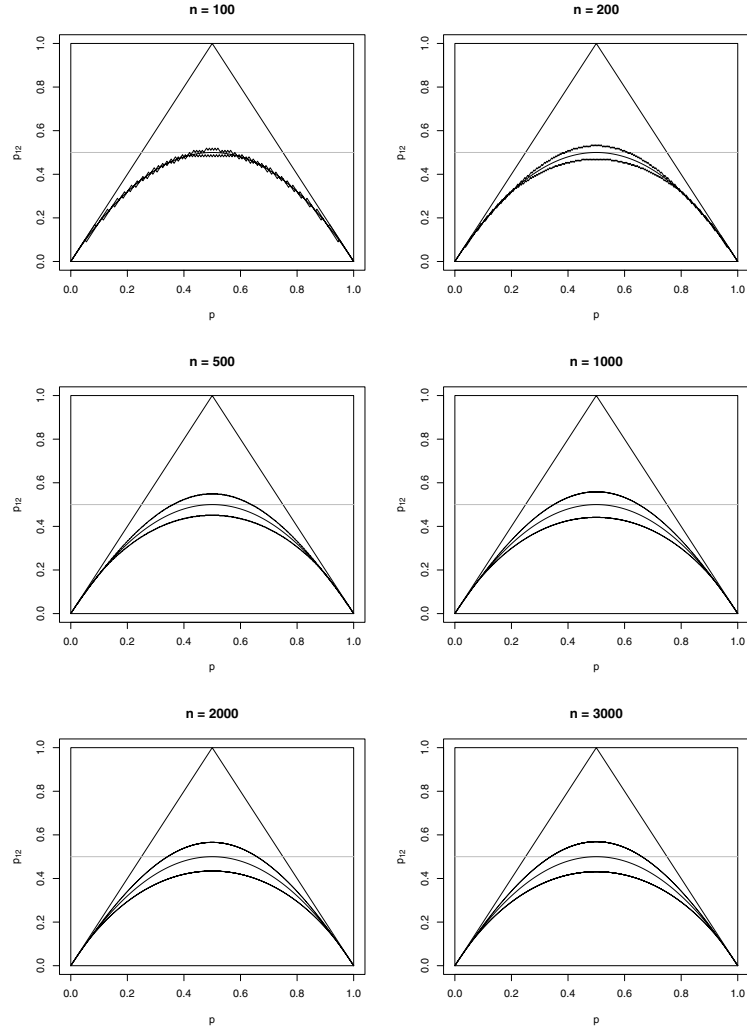


Fig. 4.7 Rejection regions of the Hardy–Weinberg equilibrium equivalence test for varying sample sizes at the 5% test level. The chosen equivalence margin is $\varepsilon_0 = 0.4$, and the sample sizes are 100, 200, 500, 1000, 2000, and 3000. The rejection regions increase with sample size.

$\hat{p}_G = 0.8391$, and estimated genotype frequencies $\hat{p}_{GG} = 0.6899$, $\hat{p}_{AG} = 0.2986$ and $\hat{p}_{AA} = 0.0116$, we obtain $\hat{\omega} = 1.6691$. This point estimate already indicates a fairly high relative excess of heterozygotes. The two-sided 95% confidence interval for ω reveals $[1.1720, 2.3771]$ that does not include the 1. Therefore, the hypothesis of validity of HWE is rejected.

Because the lack-of-fit test already rejects the hypothesis that the data are compatible with HWE, the goodness-of-fit test cannot lead to the decision that the data are in HWE. However, we perform the required calculations for illustrative purposes. To establish compatibility with HWE, we need to consider the one-sided 95% confidence interval that is obtained as $[1.2406, 2.2458]$. This confidence interval is not in the boundary $[\varepsilon_1, \varepsilon_2] = [1/(1 + 0.4), 1 + 0.4] = [1.4000, 1.2358]$. Therefore, compatibility with HWE cannot be established.

Results are different for the second SNP considered by Reich *et al.*: The estimated allele frequency of IL1B+3953T is 0.7391. The estimated genotype frequencies for the T allele at the IL1B+3953 polymorphism are 0.5478, 0.3826, and 0.0696 for the TT, CT, and CC genotype, respectively. Subsequently, $\hat{\omega} = 0.9800$. The standard two-sided 95% confidence interval for ω is obtained as $[0.7433, 1.2920]$. Because 1 is included in this interval, the standard lack-of-fit hypothesis would not be rejected. Furthermore, the one-sided 95% confidence interval reveals $[0.7771, 1.2358]$. This confidence interval is in the boundary $[\varepsilon_1, \varepsilon_2] = [1/(1 + 0.4), 1 + 0.4] = [0.7143; 1.4000]$. Therefore, equivalence with HWE can be established asymptotically for the IL1B+3953 polymorphism in this sample.

Finally, we stress that the applied approach for testing lack of fit and for establishing compatibility with HWE using the confidence interval for ω relies on large sample results. Results from Monte–Carlo simulation studies suggest that the confidence interval has a fairly good coverage for $\text{MAF} \geq 0.1$. However, the convergence of the test statistic to the standard normal distribution function is fairly slow, an exact test might be preferable. The technical level for the derivation of the exact confidence interval exceeds the level of this book. The interested reader may therefore refer to original publications [710, 711]. A set of SAS macros GOFHWEX for equivalence testing and the confidence interval approach are available. For the asymptotic version of the confidence interval approach, a simple function in R or even Excel can be used.

4.4 HOW CAN GENOTYPING ERRORS BE DETECTED IN HIGH-THROUGHPUT GENOTYPING STUDIES?

In the beginning of this chapter, we indicated that tremendous changes have occurred in the quality control procedures over the past few years. These are primarily caused by the advent of novel high-throughput SNP chip genotyping technologies. The technologies have been described in the previous Chapter 3, and the work flow of a GWA study is described in detail in Chapter 14 together with the corresponding statistical analyses.

Similar to quality control procedures and the analysis of gene expression microarray data [467, 601], several preparatory steps from the laboratory protocols can be quite important for the quality control of SNP chips. To give just one example, the age of the DNA can have an effect on the genotyping quality. It is therefore important

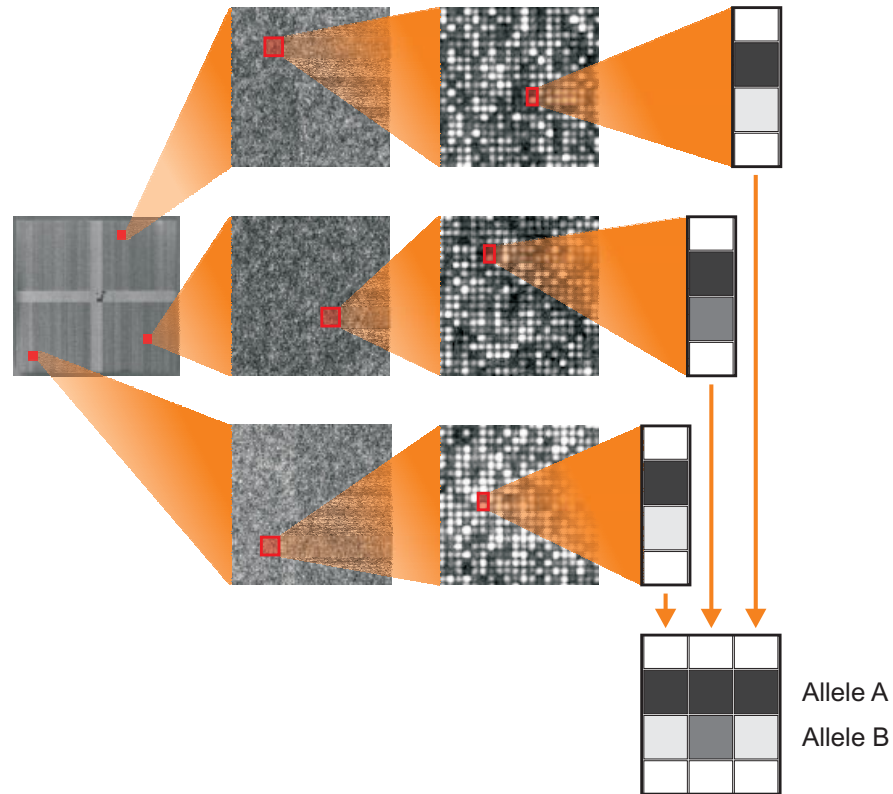


Fig. 4.8 Affymetrix Genome-Wide SNP 6.0 chip. The left part of the figure shows that the Affymetrix 6.0 chip consists of four arrays. The second and third parts of the figure are obtained by zooming in. Every SNP is represented either by three or by four probe pairs. Four probes were used by the manufacturer for SNPs with less clear clouds in the signal intensity plot (see Figure 4.9). Probe pairs were randomly distributed over the arrays subject to border effects from neighboring probes. The probes of a pair of alleles are located on adjacent positions on the chip and are pairwise right-most embedded. The rightmost part of the figure therefore represents the faked gray values for an SNP consisting in three probe pairs. The pixel values are used for calculating the signal intensities. Specifically, when the chip is scanned, signal intensities or fluorescences are read out for every probe. The larger the read-out, the more DNA on the probe, i.e., the more target DNA.

to note that producers give specific protocols how the DNA should be prepared and how the samples should be processed (see, e.g., the GeneChip Mapping 500K Assay Manual of AFFYMETRIX). In fact, they even give best practices for analysis work flow (see, e.g., Genotyping Console 3.0 User Manual of AFFYMETRIX, p. 257). Here, we describe the most important approaches for the quality control of SNP chips.

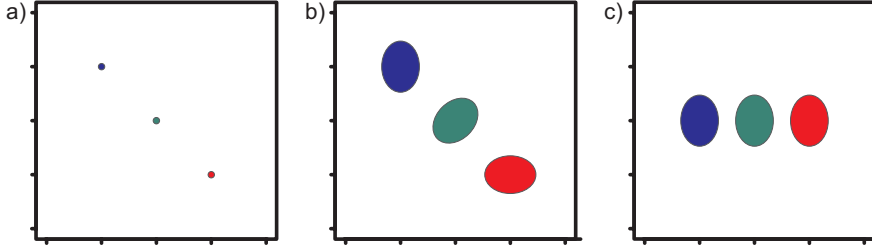


Fig. 4.9 Signal intensity plots. a) shows the idealized situation that all subjects can be assigned uniquely without measurement error to one of the three possible genotypes. b) displays the more realistic scenario of a high-quality SNP. Signal intensities of all subjects represent three clouds. The cloud of signal intensities for heterozygous subjects is not exactly between the clouds of the homozygous subjects. A slight shift is observed because of different probe affinities. c) displays the signal intensities I_A and I_B of alleles A and B in the sense of a Bland–Altman plot or an MA-plot. The x -axis shows the contrasts, i.e., the normalized difference of the signal intensities $(I_A - I_B)/(I_A + I_B)$ of both alleles, and the y -axis gives the sum of the intensities $I_A + I_B$ of both alleles. a) and b) are reprinted from Ref. [767].

Quality control typically starts after array scanning, i.e., when the image has been read. This process is described here for the Affymetrix Genome-Wide SNP 6.0 chip in Figure 4.8. At the end of the image scanning process, signal intensities or fluorescences are available for every probe. To reduce technical variation, i.e., variation which is not of biological origin, signal intensities are normalized before genotypes are called in some procedures [536], while they are not normalized in others (see, e.g., Ref. [318]). For overviews on normalization methods, see Refs. [70, 71, 335].

The genotype calling step is important, and here we describe the basic ideas using Figure 4.9. In a perfect world, all conditions in the experiment would be identical, all subjects with the same genotypes would have identical signal intensities, and there would be only three different signal intensities (Figure 4.9a). However, there are many factors affecting the signal intensity at the subject level, including DNA concentration or specific features during the hybridization process so that signal intensities of all subjects for the three genotypes form three clouds (Figure 4.9b). Signal intensities I_A and I_B for the first allele A and the second allele B , respectively, may be transformed and presented in the sense of a Bland–Altman plot [66] or an MA-plot [335] (Figure 4.9c). Specifically, the sum $sI = I_A + I_B$ of (normalized) intensities is plotted against the *contrast*

$$c = \frac{I_A - I_B}{I_A + I_B}, \quad (4.10)$$

i.e., the normalized difference. Several other contrast measures have been proposed, including $c_P = \sinh \frac{2I_A I_B}{I_A + I_B} / \sinh(2)$ and $\ln c$ [536].

In Figure 4.9 genotypes have already been assigned to the clouds. In practice, only signal intensities are available, and the genotype calling step consists in a cluster algorithm applied to the signal intensities (Figure 4.10). However, in contrast to

standard cluster algorithms, where no prior information is available, we already know that three clouds are present for an SNP. Furthermore, from previous experiments, a priori knowledge is or might be available about the position of the cluster in the two-dimensional signal intensity space. This biological knowledge is utilized in the genotype calling algorithms. A series of different genotype calling algorithms has been proposed, and overviews on available genotype calling algorithms have been given, e.g., in the reviews in Refs. [657, 767].

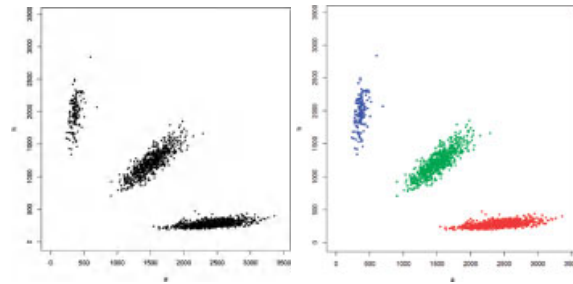


Fig. 4.10 Signal intensity plots pre- and post-genotype calling. The left figure shows the signal intensities prior to genotype calling and the right figure gives the signal intensities after genotypes have been assigned by a genotype calling algorithm.

After genotype calling, i.e., after the assignment of genotypes to subjects based on the signal intensities, standard quality control (sQC) is performed both on the subject level and on the SNP level. Summaries on this topic have been given in Refs. [657, 708, 767], and the most important sQC measures for autosomal markers are described in the following. Two fundamentally different approaches are distinguished, and they can be characterized by the terms sample quality control (Section 4.4.1) and SNP quality control (Section 4.4.2). The global sQC filters described here are helpful in removing SNPs with clustering problems. They thus reduce a large number of highly significant erroneous associations, and they lower the genomic control lambda (see Section 11.4.3) so that quantile-quantile plots (Q-Q plots) look as expected. In brief, Q-Q plots compare the obtained test statistics with what is expected under the null hypothesis of no association. Q-Q plots are therefore a tool for visualizing and assessing the extent of systematic deviation from the distribution under the null hypothesis; for a detailed description, see Chapter 14. However, the visual inspection of signal intensity plots is still the ultimate quality control approach when an association has been identified, and this will be described in Section 4.5.

4.4.1 How are samples checked in high-throughput genotyping studies?

On the sample level, three different criteria are usually employed for filtering subjects and they are described in this section. The *proportion of missing genotypes* is a useful first indicator of poor genotype quality. The proportion of missing genotypes on the

sample level is unfortunately and incorrectly often termed call rate. *Call fraction* is, however, a correct synonym. A high fraction of missing genotypes usually implies hybridization problems that can be caused by faulty arrays or poor DNA quality. A typical choice is to exclude probands from the analysis who have been successfully genotyped at less than 97% of the SNPs. Thus, the proportion of missing genotype per sample is calculated as the extent of null genotype calls. It is important to note that this proportion still includes monomorphic SNPs. Monomorphic SNPs are SNPs where all subjects in the study have the same genotype. In fact, a monomorphic SNP is not polymorphism. In practice, the call fraction is sometimes lowered because of conditions outlined in a genotyping contract. For example, it might be that individual data are considered as having sufficient quality in the laboratory if the call fraction exceeds 90%.

The second good filter is *heterozygosity* because a high mean heterozygosity can be an indicator of DNA contamination. If heterozygosity is too low, hybridization might have failed. The mean heterozygosity is calculated as the proportion of heterozygous genotypes over the available SNPs. For filtering, a typical approach is to estimate the mean (m) and standard deviation (SD) of the heterozygosity across all study subjects and to exclude all subjects outside of a $3 SD$ interval around the mean m . A common alternative is to identify and exclude outliers based on the histogram of heterozygosities. No general proportion for lower and upper thresholds can be given because the heterozygosity of a sample depends on several factors, and the most important one is ethnicity.

The first two filter criteria can be represented separately or in a single figure to decide on the filtering threshold (Figure 4.11).

The third criterion investigates the relatedness of the subjects. In large-scale studies, *cryptic relatedness*, i.e., correlation between genotyped samples caused by an accidental sample duplication or inadvertent use of related subjects cannot be excluded. Specifically, for a random pair of study subjects there is a 50% probability that a randomly chosen allele is IBS; for details on allele sharing measures, see Chapter 8. A substantial increase in this probability is an indicator of relatedness. The IBS value can be calculated as follows. Let X_{ij} denote the number of A alleles of subject i at an autosomal SNP j . Then the IBS value between subjects i and i' at SNP j is

$$\text{IBS}_{ii'j} = 2 - |X_{ij} - X_{i'j}|.$$

The mean IBS for subjects i and i' simply is the average over the J available SNPs, i.e.,

$$\text{IBS}_{ii'} = \frac{1}{J} \sum_{j=1}^J \text{IBS}_{ii'j}.$$

If there is a sample pair with an increased IBS value, at least one of them should be excluded from the study. An alternative to the IBS approach uses the probability that an allele is IBD, i.e., the allele has been inherited from the same founder [548]. If the relationship can be determined uniquely, appropriate adjustments for the cryptic relatedness might be possible. However, the power of an association study should be investigated in detail if correlated samples are present.

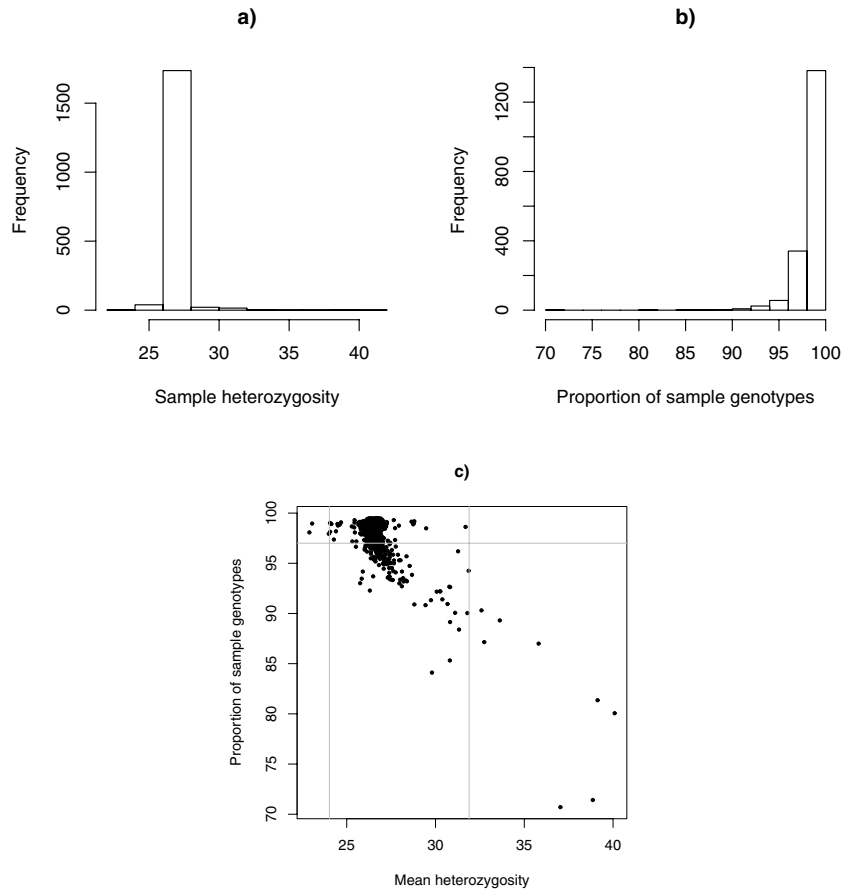


Fig. 4.11 Individual missing data and sample heterozygosity for each sample out of approximately 3000 control individuals from the Gutenberg Heart Study, assessed with the Affymetrix Genome-Wide SNP 6.0 chip across approximately 900,000 autosomal SNPs. Histogram a) shows the proportions of sample heterozygosities. Histogram b) displays the proportion of missing genotypes per sample. In scatterplot c), the proportion of missing genotypes is plotted against mean heterozygosity, and each point represents a subject. Samples not falling in the upper middle square are excluded from further considerations. Data have been kindly provided by Tanja Zeller.

The IBS distance can also be used to identify heterogeneity in the sample. Specifically, one may perform a multidimensional scaling (MDS) using the similarity between pairs of individuals using the IBS measure, averaged over all SNPs, as similarity measure. MDS has been nicely described, e.g., in Refs. [486, 611], and the interested reader may refer to the literature. One can display the first two com-

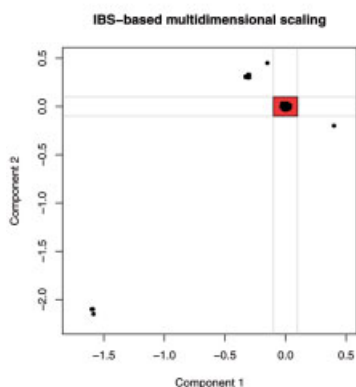


Fig. 4.12 Multidimensional scaling (MDS) for detecting genetic heterogeneity between samples, i.e., population stratification using the identity by state (IBS) measure. All subjects outside of the red square have values that are sufficiently far from 0 on one or both components in the IBS-based MDS. They should be excluded from the sample so that a genetically homogeneous sample is investigated. Otherwise, appropriate corrections for population stratification need to be performed.

ponents of the MDS in a scatterplot (Figure 4.12), and samples strongly deviating from 0 on at least one of the two components represent samples that may have a different genetic background. For example, in a population-based study, we have identified previously undetected samples of Asian origin and of African origin with this approach. We subsequently excluded these samples from further considerations because we restricted our attention to Caucasian samples.

4.4.2 How are single nucleotide polymorphisms checked in high-throughput genotyping studies?

Several quality procedures are available on the SNP level and three of them are described in detail in this section. The first important quality measure on the SNP-level is *missing frequency (MiF)*, i.e., the proportion of missing genotypes. It is complement to the call fraction, i.e., the subject-wise proportion of missing genotypes. An also used unfortunate term is the call rate that is $1 - \text{MiF}$. This criterion is the most important on the SNP level because it is an indicator of how well the clusters of an SNP are separated. The further they are apart and the more homogeneous the clusters are, the less likely it is that the genotype calling algorithm gives missing data. SNPs with a high proportion of missing genotypes should be removed. A standard filter criterion is to exclude SNPs with a $\text{MiF} > 2\%$ in cohort studies. For case-control studies, the MiF should be separately investigated in cases and in controls because differential missingness between cases and controls can result in spurious associations [131].

The second criterion is to check for deviation from HWE. SNPs are excluded if there is an excess heterozygosity or heterozygote deficiency. Although this can help in identifying SNPs with obvious genotyping errors [658], care should be taken when or even whether this filter is used. For example, for diseases where selection either of homozygotes or of heterozygotes might play an important role, a deviation from HWE can be expected. The use of the HWE criterion would probably filter out many if not all relevant association signals (also see Section 11.2.1). Furthermore, more

experience with genotype calling algorithms is required because the number of SNPs with heterozygote deficiency or excess heterozygosity substantially depends on the genotype calling algorithm used [250].

The third criterion is the MAF, i.e., the frequency of the rare allele. Most genotype calling algorithms tend to perform poorly for SNPs with low MAF. Furthermore, current single GWA studies do not have the power to detect associations with SNPs with a low MAF (see Section 11.3).

An additional set of filter criteria already exists (see, e.g., Ref. [422]). For example, we recommend the following gender-specific filters:

- Absolute difference in proportion of missing genotypes for males and females per SNP: if the difference is large, the SNP should be excluded.
- Proportion of heterozygotes in males and females in all samples: if heterozygosity varies substantially between males and females, the SNP should be excluded.
- Missing data by gender: SNPs with differential missingness by gender should be excluded.
- Test of allelic association by gender among controls or all subjects from a population-based cohort: if genotypes substantially vary between gender within control subjects or within all subjects from a population-based cohort, the SNP should be excluded.

4.5 HOW SHOULD CLUSTER PLOTS AND HOW CAN THE QUALITY OF CLUSTERS BE INVESTIGATED IN HIGH-THROUGHPUT GENOTYPING STUDIES?

Although the sQC approaches on the subject level and on the SNP level described above are helpful in identifying SNPs of low quality, the visual inspection of signal intensity plots is still the ultimate quality control approach when an association has been identified [767]. For example, Affymetrix clearly recommends the visual inspection of all candidate SNPs ([5], p. 257). As in diagnostic studies, we recommend that at least two experienced readers visually inspect the cluster plots in a blinded fashion. Affymetrix' recommendation to inspect only candidate SNPs is probably a consequence of the fact that systematic visual inspection of all cluster plots is impossible in a high throughput setting because of the high workload. And readers are fatigued after a short period. Therefore, approaches are helpful that allow the automated inspection of cluster plots or signal intensities.

At the beginning of Section 4.4, we have described that the genotype calling step can be compared with employing a clustering algorithm. However, there are two differences to standard clustering. First, there should be either three or two clouds in the signal intensity plot; two clouds only if the genotype frequency for one of the homozygotes is low. Second, there is a priori information about previous signal intensities, i.e., the position of the clusters. This a priori biological information has

an effect on the judgment whether the SNP can be used for further analysis. The two basic aspects of analyzing a clustering algorithm are

- Quality: How good is the clustering, i.e., the genotype calling produced?
- Speed: How fast can it be found?

In the following, we deal only with the first question, and for this we only consider approaches that assume all points are of equal importance in the determination of the clustering and in measuring the quality of the clustering. Many approaches have been proposed for evaluating the clustering quality, and excellent reviews have been given in the literature [274, 275, 284, 355, 373].

The major distinction is made between external validity and internal validity. Internal and external validity should not be confused with internal and external validation that has been discussed in detail, e.g., in Refs. [19]. External validity measures compare the actual genotyping result from the genotype calling algorithm with the true genotypes. They therefore investigate genotyping errors. Internal validity measures do not use any additional external information which often is also unavailable. Internal validity measures thus investigate how well the genotypes correspond to the natural cluster structure of the signal intensities. In this section, we consider only internal validity measures.

Intuitively, a genotype calling algorithm performs well for a specific SNP if neighboring points in a signal intensity plot are assigned to the same genotype and points that are dissimilar are assigned to different genotypes. Furthermore, the three clusters should be clearly separate, i.e., there should be a large distance between the different genotype groups. Phrased differently, a good SNP will have small distances within a genotype group and large distances between different genotypes.

Formally, the validity of a genotype calling can be measured by the following yardsticks [284, 355, 760]:

- *Compactness* or *homogeneity* measures closeness of genotypes. This concept is related to the intra-cluster variation and therefore a typical example for such a measure is the variance. Of course, the variance also indicates how different the subjects within a genotype group are. However, a low value of variance is an indicator of closeness.
- *Connectedness* attempts to assess how well a partitioning group's subjects are connected with their nearest neighbors. Representatives of such measures count violations of nearest neighbor relationships.
- *Separability*, also termed *separation*, indicates how distinct two genotype groups are. It computes the distance between two different clusters. An overall rating for a partitioning can be defined as the average weighted inter-cluster distance, where the distance between individual clusters can be computed as the distance between cluster centroids or as the minimum distance between data items belonging to different clusters. Alternatively, cluster separation in a partitioning may, for example, be assessed as the minimum separation observed between individual clusters in the partitioning.
- *Combinations*: A number of approaches combine measures of the above types, and several measures assess both intra-cluster homogeneity and inter-cluster

separation. The final score is computed as the linear or non-linear combination of the two measures. An example of a linear combination is the SD-validity Index (see below), and examples for non-linear combinations are the Dunn Index, Dunn-like Indices, the Davies–Bouldin Index, or the Silhouette Width (for all measures, see below).

- *Stability* is a special form of internal cluster validity. The question that is asked here is how sensitive a method is to perturbation of the data, i.e., how sensitive the genotypes are with respect to small changes in the signal intensity? These approaches for assessing cluster validity are clearly not external since they do not compare the obtained genotypes with other external information, i.e., other external data sets. Measures of this type repeatedly re-sample or perturb the original data set, and re-cluster the resulting data. The consistency of the corresponding results provides an estimate of the significance of the clusters obtained from the original data set.

Below, we describe approaches to the four different criteria. We do not consider all approaches that have been proposed in the literature in the context of cluster analysis in detail; for excellent reviews on this topic, see Refs. [274, 275, 284]. Instead, we focus on methods that have been used in the context of SNP genotype calling. First, we consider measures of cluster compactness (Section 4.5.1). Second, we consider a measure of connectedness (Section 4.5.2), followed by a measure of separability (Section 4.5.3). Before we finally discuss several combination measures in Section 4.5.5, we present a perturbation method for measuring stability of the genotypes (Section 4.5.4).

To simplify notation, thus to avoid a group index, we define these measures for a single study group, i.e., subjects from a cohort study or either the cases or the controls from a case-control study. In the following, the clusters are indexed by k , $k = 1, 2, 3$, the sample size per cluster is n_k and $n = n_1 + n_2 + n_3$. The distance measure we consider is the Euclidian distance between any two subjects i and i' . To this end, let $c(i)$ and $s(i)$ denote the contrast of Eq. (4.10) and sum, respectively, for subject i . Then, the Euclidian distance is given by

$$d(i, i') = \sqrt{(c(i) - c(i'))^2 + (s(i) - s(i'))^2}. \quad (4.11)$$

If necessary indices are used for indicating the cluster, e.g., $c_k(i)$ for a subject from cluster i or $d_{k,k'}(i, i')$ for the distance between i and i' , when i is from cluster k and i' belongs to cluster k' . The center of cluster k is denoted by

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^k \begin{pmatrix} c_k(i) \\ s_k(i) \end{pmatrix},$$

and the distance between subject i of cluster k and the center of cluster k is $d(i, \mu_k)$.

4.5.1 What are measures for cluster compactness?

Typical measures for cluster compactness are the *cluster-specific intra-cluster variance* $\mathbb{V}ar(\mathbf{IC}_k)$ and the *overall intra-cluster variance* $\mathbb{V}ar(\mathbf{IC})$, which are based on the squared Euclidian distance d^2 and given by

$$\sigma_k^2 = \mathbb{V}ar(\mathbf{IC}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} d_k^2(i, \boldsymbol{\mu}_k) \quad \text{and} \quad \sigma_w^2 = \mathbb{V}ar(\mathbf{IC}) = \frac{1}{n} \sum_{k=1}^3 \sum_{i=1}^{n_k} d_k^2(i, \boldsymbol{\mu}_k).$$

In applications, the *root mean square distance (RMSD)* is often considered that is just the square root of $\text{RMSD} = \sqrt{\mathbb{V}ar(\mathbf{IC})}$. The intra-cluster variance should be small, and SNPs should be excluded if RMSD is large.

4.5.2 What are measures for cluster connectedness?

To reflect cluster connectedness, Handl and Knowles [283] used the connectivity, which evaluates the degree to which neighboring data points have been placed in the same cluster. Consider subject i of cluster k , and let $\text{nn}_{i(j)}$ denote the j th nearest neighbor of i that is determined by using Eq. (4.11). If the j th nearest neighbor $\text{nn}_{i(j)}$ has the same genotype as subject i , subjects i and $\text{nn}_{i(j)}$ are connected, and the connectivity between them is 0. Otherwise, it is $1/j$. This may be summarized to

$$C_{i, \text{nn}_{i(j)}} = \begin{cases} 0 & \text{if } i \text{ and } \text{nn}_{i(j)} \text{ have the same genotype,} \\ \frac{1}{j} & \text{if } i \text{ and } \text{nn}_{i(j)} \text{ have different genotypes.} \end{cases}$$

The number of neighbors to be considered for c is denoted by J so that the *connectivity* finally is

$$\text{Conn} = \sum_{i=1}^n \sum_{j=1}^J C_{i, \text{nn}_{i(j)}}.$$

The connectedness is good if Conn is small, and SNPs should be excluded if Conn is too large.

Recently, Schillert et al. [591] proposed to count the number of samples that lie too close to the cluster of a different genotype. SNPs are excluded for which a certain threshold for overlapping samples per cluster is exceeded.

4.5.3 What are measures for cluster separation?

Several reasonable measures for cluster separation are available. In the context of genotype calling, the minimum distance between any two clusters seems to be meaningful. Alternatively, one may consider the average inter-cluster distance between the two homozygous genotype groups and the heterozygotes. In detail, the distance between any two clusters can be computed as the distance between cluster centers, i.e., $d(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'})$, where d is the Euclidian distance as above. Without loss of gen-

erality, assume that clusters are ordered from left to right for $k = 1, 2, 3$. Then the *minimum inter-cluster distance* (minD) between any two clusters is

$$\text{minD} = \min_{k \neq k'} d(\mu_k, \mu_{k'}),$$

and the *average inter-cluster distance* (meanD) is

$$\text{meanD} = \frac{d(\mu_1, \mu_2) + d(\mu_2, \mu_3)}{2}.$$

4.5.4 How can stability of genotypes be investigated by perturbation analysis?

Teo et al. [659] perturbed signal intensities by adding random noise. Specifically, normally distributed errors are added independently to both normalized intensities I_A and I_B . In detail, the perturbed intensities are $I'_A = I_A + \varepsilon_A$ and $I'_B = I_B + \varepsilon_B$ for independently identically normally distributed random variables $\varepsilon_a, \varepsilon_B \sim N(0, \sigma^2)$. The authors proposed as variances 2%, 10%, and 20% of the mean intensities, and SNPs should be filtered if 1%, 2%, 5%, or 10% of the genotypes change. The major disadvantage of this approach is its central processing unit (CPU) time. Specifically, genotypes need to be called anew after adding the perturbation. Therefore, this method is theoretically appealing but not ideally suited for use in practice.

Other standard approaches that have been proposed in the literature for evaluating cluster stability are even less well suited because they typically start with splitting the sample, e.g., in two halves. The first half is used for genotype calling, and the second half is used for prediction [284]. The first argument against the use of this idea in the context of genotype calling is that not all available subjects are used for genotype calling and this can be problematic in case of small sample sizes and low MAFs. The second argument is that it also requires several rounds of time-consuming genotype calling because the sample needs to be split in halves repeatedly.

4.5.5 What are reasonable combined measures of genotype validity?

A series of different measures for evaluating the genotype validity using combined measures have been proposed and overviews have been given, e.g., in Refs. [274, 275, 284, 355, 373]. The quality of a SNP depends on how well the clusters are separated. At the same time, they should be as compact as possible. As a measure of separateness, Plagnol et al. [536] considered the difference between the centers of adjacent clouds. They divided this difference by the sum of the standard deviation for these two clouds, i.e., by a measure of compactness. In their measure, they used one additional idea. When the contrast is plotted on the x -axis and the sum of the intensities on the y -axis, only the contrasts need to be considered because the variability on the y -axis should be similar and the clouds should be vertical. This has the advantage that long small clouds are not depreciated because of their lower compactness.

In detail, denote the cluster-specific contrast means and contrast standard deviations by $\bar{c}_k$ and s_k . Without loss of generality, assume that clusters are ordered from left to right for $k = 1, 2, 3$. Then the *cluster separation criterion* is given by

$$\text{CSC} = \min \left\{ \frac{\bar{c}_2 - \bar{c}_1}{s_1 + s_2}, \frac{\bar{c}_3 - \bar{c}_2}{s_2 + s_3} \right\}. \quad (4.12)$$

If more than one group is available as in case-control studies, the cluster separation criterion is the minimum over the group-specific CSCs. The CSC can be interpreted as a *Dunn Index (DI)* and belongs to the group of Dunn-like indices [275], where a measure of separateness is used in the numerator and a measure of compactness in the denominator. However, for DIs, the measure of compactness usually is the maximum of the standard deviations s_k . This is in contrast to Eq. (4.12), where the sum is used. A good SNP has maximal inter-cluster distances and minimal intra-cluster distances. Thus, large values of CSC and DI correspond to good SNPs. Although the DI is simple to understand and to calculate, it is unfortunately unstable in case of outliers.

As for the CSC, the *Davies–Bouldin (DB) Index* compares the inter-cluster distance with the corresponding intra-cluster distance. It is a function of the ratio of the sum of within-cluster variance to between-cluster separation. Specifically, we consider

$$\text{DB} = \frac{1}{3} \sum_{k=1}^3 \max_{k \neq k'} \left\{ \frac{\sigma_k^2 + \sigma_{k'}^2}{d^2(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'})} \right\}. \quad (4.13)$$

Thus, the DB Index measures the average of similarity between each cluster and its most similar one. As the clusters should be compact and separated, a low DB Index means better separated genotypes. The DB Index can be defined in different ways. For example, Dixon et al. [164] used the Euclidian for defining both the within-cluster variability and the between-cluster distance. The DB Index can be defined even more generally (see Refs. [355, 373]).

Another measure related to DI, DB, and CSC is the *Calinski–Harabasz (CH) Index*. It is a pseudo-F statistic and compares the between-cluster sum of squares with the within-cluster sum of squares. It is defined as

$$\text{CH} = \frac{\sum_{k=1}^3 d^2(\boldsymbol{\mu}_k, \boldsymbol{\mu}) / 2}{\sum_{k=1}^3 \sum_{i=1}^{n_k} d_k^2(i, \boldsymbol{\mu}_k) / (n - 2)},$$

where $\boldsymbol{\mu}$ is the center over all three clusters.

The Modified Hubert's Index (MHI) assesses the fit between the intensities and the genotypes. To this end, we consider the Euclidian distances $d(i, i')$ of two subjects, and $q(i, i')$ is the concordance indicator, i.e., whether i and i' have been assigned the same genotype. In detail,

$$q(i, i') = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ have been assigned the same genotype,} \\ 0 & \text{if } i \text{ and } i' \text{ have been assigned different genotypes.} \end{cases}$$

Then, the *Modified Hubert's Index* Γ is given by

$$\Gamma = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n d(i, i') q(i, i').$$

The MHI can also be used in a standardized form. This means that the Pearson correlation coefficient between d and q is used instead of the simple product [57]. To this end, let $\bar{d}$ and s_d denote the average distance and standard deviation over all $n(n-1)/2$ pairs of subjects, and $\bar{q}$ and s_q denote the average and standard deviation of the concordance indicator q . Then, the *standardized MHI* (Γ_s) is

$$\Gamma_s = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \frac{(d(i, i') - \bar{d})(q(i, i') - \bar{q})}{s_d s_q}.$$

As with the genotype calling itself, all these measures are more likely to be problematic for low MAFs.

The Silhouette Index, also termed Silhouette statistic or Silhouette score of Rousseeuw [572], is a popular approach for judging cluster separation. It has been used by Lovmar et al. [430] in the “pre-GWA studies era” for measuring the quality of TaqMan genotypes. In the Silhouette approach, the average distance is calculated for every point of a cluster to all other points of the same cluster. This average distance is then compared with the average distance of the point to the neighboring cluster. Specifically, let $a_k(i)$ denote the average Euclidian distance from subject i of cluster k to all data points in the same genotype cluster, i.e.,

$$a_k(i) = \frac{1}{n_k - 1} \sum_{i' \neq i, i'=1}^{n_k} d_{k,k}(i, i').$$

Similarly, we define $b_{k'}(i)$ as the average Euclidian distance from i to all data points in cluster k' , $k \neq k'$:

$$b_{k'}(i) = \frac{1}{n_{k'}} \sum_{i'=1}^{n_{k'}} d_{k,k'}(i, i').$$

$b_k(i)$ is the average distance of all data points in the cluster closest to the data point, i.e., the minimum of the $b_{k'}(i)$:

$$b_k(i) = \min_{k' \neq k} \{b_{k'}(i)\}.$$

The Silhouette $SI_k(i)$ measures how well subject i matches the clustering obtained by the genotype calling algorithm by looking at the difference between $b_k(i)$ and $a_k(i)$:

$$SI_k(i) = \frac{b_k(i) - a_k(i)}{\max\{a_k(i), b_k(i)\}} = \begin{cases} 1 - a_k(i)/b_k(i) & \text{if } a_k(i) < b_k(i), \\ 0 & \text{if } a_k(i) = b_k(i), \\ b_k(i)/a_k(i) - 1 & \text{if } a_k(i) > b_k(i). \end{cases}$$

For each cluster, the average width of a Silhouette, termed *cluster-specific Silhouette Index* (SI_k) is obtained as the mean of all Silhouettes of a cluster. The *Silhouette Index* (SI) is the arithmetic mean of the Silhouette widths so that

$$SI_k = \frac{1}{n_k} \sum_{i=1}^{n_k} SI_k(i) \quad \text{and} \quad SI = \frac{1}{n} \sum_{k=1}^3 SI_k.$$

When a genotype group contains only a single subject, the Silhouette of this subject is set to 0.

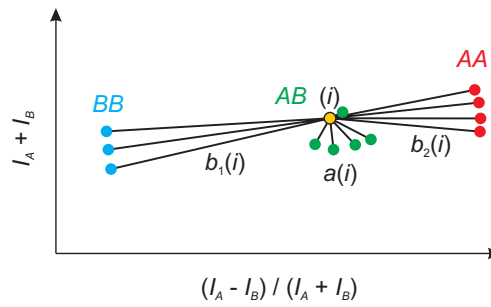


Fig. 4.13 Principle of Silhouette scores. Basic idea of Silhouette scores for quality assessment of clusters using Silhouette scores, illustrated for a single data point i from cluster $k = 2$. The sum of signal intensities $I_A + I_B$ is plotted against the contrast of signal intensities $(I_A - I_B)/(I_A + I_B)$. For each data point i in cluster k in the scatterplot, the Silhouette $s_k(i)$ is calculated by the formula in the figure. $a_k(i)$ is the average Euclidian distance from i of cluster k to all data points in the same genotype cluster (spotted green lines), and $b_k(i)$ is the average Euclidian distance from i to all data points in the cluster closest to the data point, i.e., either $b_1(i)$ (blue lines) or $b_3(i)$ (red lines) [430, 572]. The average Silhouette width $s(k)$ of a cluster $k = 1, 2, 3$ is the mean of all Silhouettes $s(i)$ for each genotype cluster. The overall average Silhouette width is the mean of the average Silhouette width over all clusters. The figure has been created in an analogy to Ref. [430], Fig. I.

Figure 4.13 depicts the basic idea of the Silhouette Index approach. The SI is restricted to the interval $[-1, 1]$. Clearly separating clusters have values close to 1, and -1 is the worst outcome in the separation. If $SI_k(i)$ is close to 0, the subject could well be assigned to another closest cluster as well, and the sample lies equally far away from two clusters. If $SI_k(i)$ is negative, it better fits in a different cluster.

One alternative to the standard SI would be to consider only the Silhouette width using the Euclidian distance of the contrasts instead of the Euclidian distance on both axes, contrast, and sum. Furthermore, a simple indicator of the quality of the SNP would be to count the number of positive Silhouettes, and a *modified Silhouette Index* (mSI) is the proportion of Silhouettes $SI_k(i)$ with a positive sign [164].

Example 4.4. In this example, we illustrate signal intensity plots generated by TaqMan probe technology. Figure 4.14 shows the standard and the transformed signal intensity plots for the SNP rs2241622 typed on a TaqMan platform (Applied

Biosystems). The SNP was typed using an ABI assay, 4 ng DNA in 10 μ l total volume, including ROX reference dye. PCR conditions were 67 °C annealing, 40 cycles. Genotypes of this SNP were automatically called using the proprietary sequence detection system (SDS) software of Applied Biosystems. We estimated some cluster validity indices for this SNP and obtained for the cluster separation criterion $CSC = 11.46$. The Davies–Bouldin Index was $DB = 0.0140$, the connectivity revealed $Conn = 3.2869$, and the Silhouette Index was $SI = 0.90$. All measures thus indicate that the SNP is of high quality.

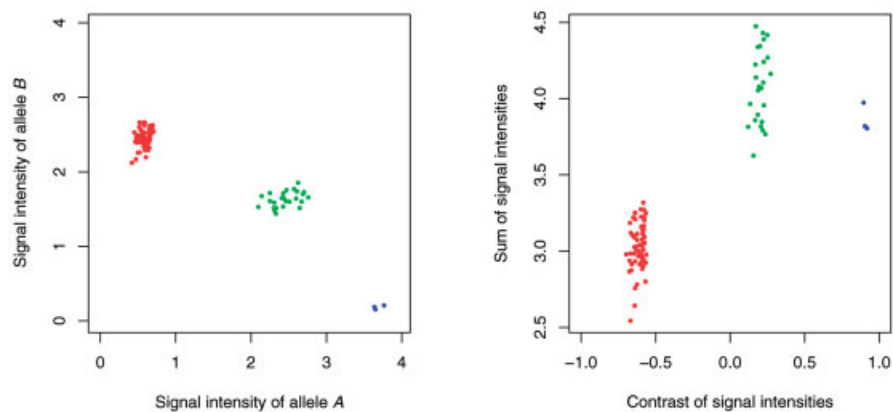


Fig. 4.14 Signal intensity plots for rs2241622 typed on the TaqMan platform. This single nucleotide polymorphism is of good quality, and the genotypes were called automatically using the Sequence Detection System software of Applied Biosystems. The untransformed signal intensities are displayed in the left figure, while the transformed ones are shown in the right figure. Data have been kindly provided by Tanja Zeller.

Figure 4.15 shows signal intensity plots for the manually called rs2015622. This SNP was typed with a self-designed assay using primer and probes from Sigma and Bioline TaqPolymerase. In addition, ROX reference dye was used. PCR conditions were 67 °C annealing and 40 cycles. Signal intensities were read-out using the TaqMan platform of Applied Biosystems. The upper panel shows the genotyping results when 10 ng DNA were used in 7 μ l total volume, while the lower panel gives the genotyping result when 10 ng DNA was used. The figure clearly indicates improved genotyping, when a higher amount of DNA was used. This can be seen from the better separating clusters and the cluster validity indices. Specifically, we obtained for the 10 ng DNA assay $CSC = 2.6470$, $DB = 0.3647$, $Conn = 5.9183$, and $SI = 0.64$. In contrast, for the 15 ng DNA assay we yielded $CSC = 3.0810$, $DB = 0.1365$, $Conn = 2.7385$, and $SI = 0.68$. Nevertheless, even with the higher amount of DNA, this SNP has a lower quality than rs2241622.

Example 4.5. In this example, we illustrate signal intensity plots from a GWA study that used the Affymetrix Genome-Wide SNP 6.0 Array. More than 3300 population-based samples from the Gutenberg Heart Study were processed with this array.

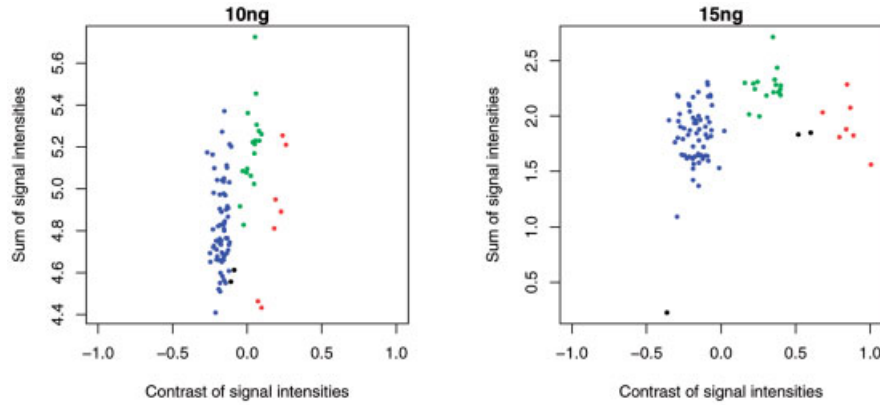


Fig. 4.15 Signal intensity plot for rs2015622 typed on the TaqMan platform. This assay to genotype rs2015622 is of bad quality, and the genotypes were called manually using the Sequence Detection System software of Applied Biosystems. The left panel shows the TaqMan results when 10ng DNA were used as template for polymerase chain reaction (PCR), and the right panel for 15ng DNA. The quality of the genotyping was better for the higher amount of DNA because the clouds separated more clearly. The transformed signal intensities are shown. Data have been kindly provided by Tanja Zeller.

Figure 4.16 shows a good quality SNP, and all other SNPs of this figure were of bad quality. We chose, however, these SNPs for illustrating the different values of several cluster validity indices that are provided in Table 4.3. The only SNP with an SI > 0.5 is the SNP of good quality. It is also the SNP with the lowest DB Index, highest cluster separation criterion, and lowest connectivity. The example thus illustrates the usefulness of these cluster validity measures.

Table 4.3 Cluster validity indices for the 6 single nucleotide polymorphisms (SNPs) for which the signal intensity plots are shown in Figure 4.16.

SNP	DB	CSC	Conn	SI
rs6950845	0.0419	14.47	0.00	0.83
rs10501461	0.3406	1.946	0.42	0.49
rs11806327	0.3893	1.566	1.86	0.68
rs12101060	1.5222	1.478	6.10	0.33
rs12407460	0.5602	1.806	0.13	0.36
rs16923480	0.4592	2.354	6.00	0.40

DB: Davies-Bouldin Index; CSC: cluster separation criterion; Conn: connectivity; SI: Silhouette Index.

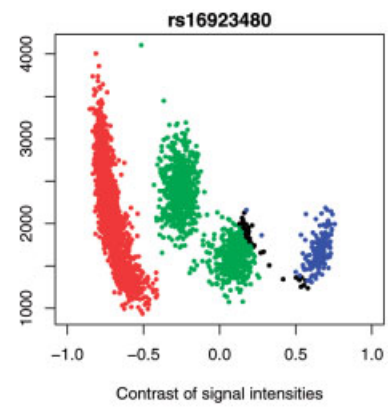
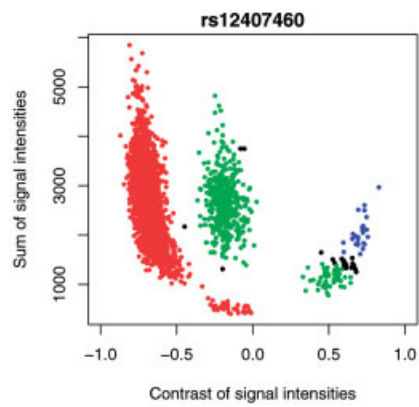
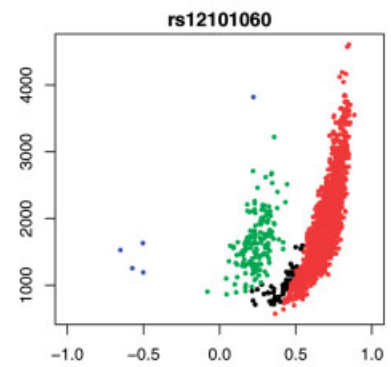
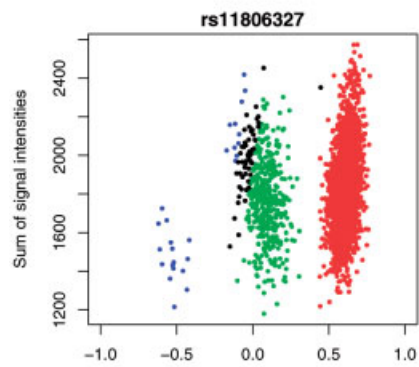
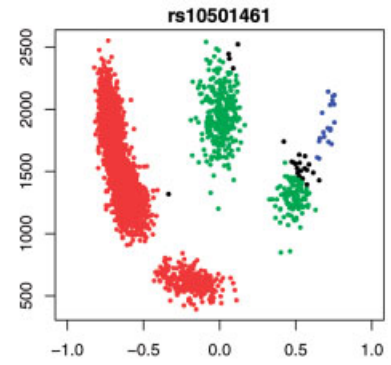
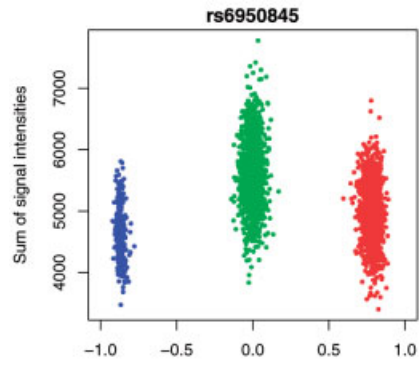


Fig. 4.16 Signal intensity plots for 6 single nucleotide polymorphisms (SNPs) typed on the Affymetrix Genome-Wide SNP 6.0 Array. The transformed signal intensities are displayed. The first SNP rs6950845 is of good quality, while the other 5 SNPs clearly show genotyping problems. For the rs10501461, more than three clouds can be distinguished, and one cloud shows almost no signal intensities for either allele. rs11806327 reveals three clouds, but the cloud of heterozygous subjects was cut in two by the genotype calling algorithm. While one part of the subjects was correctly assigned to be heterozygous (green part), another part of the same cloud were called homozygous by the genotype calling algorithm (blue part). The subjects between the green-colored and the blue-colored parts were called as missing genotypes, and they are displayed in black. rs12101060 does not clearly separate the green and the red clouds, and the subjects in-between are assigned the missing genotype. rs12407460 shows four clouds, and the cloud for the homozygous *A* subjects is split between homozygous *AA* and heterozygous subjects. For rs16923480, four clouds can be clearly distinguished, and one cloud is cut in two. Data have been kindly provided by Stefan Blankenberg.

4.6 PROBLEMS

Problem 4.1 The families in the following pedigrees have been genotyped at two independent STR markers.

Problem 4.1.1. Is the pedigree displayed in Figure 4.17 consistent with Mendelian inheritance?

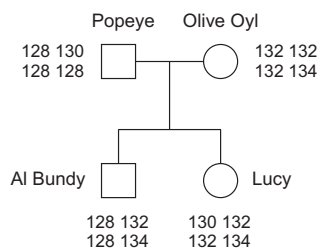


Fig. 4.17 Pedigree for Problem 4.1.1

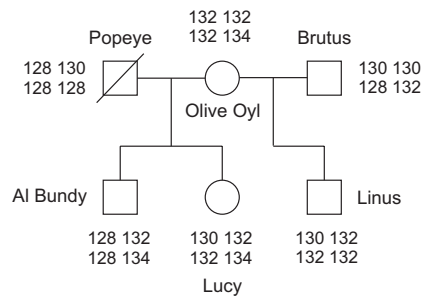


Fig. 4.18 Pedigree for Problem 4.1.2

Problem 4.1.2. Further investigations at a later time point showed that Popeye has meanwhile passed away. Olive Oyl married Brutus, and they have a son named Linus (see Figure 4.18). Does this new information offer a solution for the apparent genotyping error (Figure 4.17)?

Problem 4.2 Peggy and Al Bundy have two daughters (Figure 4.19). Are the genotypes compatible with Mendelian laws? If not, which error is most likely?

Problem 4.3 Part of the family from Problem 4.2 has been genotyped at three further markers (trinucleotide STRs); see Figure 4.20.

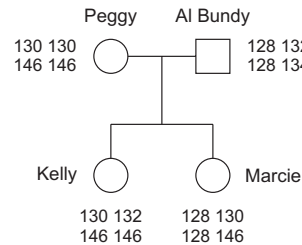


Fig. 4.19 Pedigree for Problem 4.2

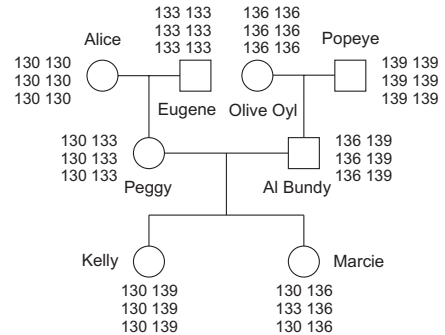


Fig. 4.20 Pedigree for Problems 4.3.1 and 4.3.2

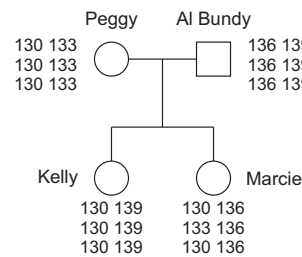


Fig. 4.21 Pedigree for Problem 4.3.3

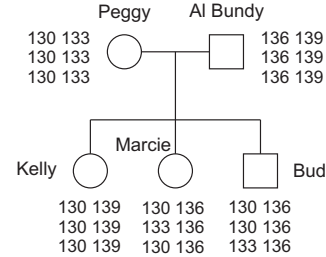


Fig. 4.22 Pedigree for Problem 4.3.4

Problem 4.3.1. Is the pedigree depicted in Figure 4.20 compatible with Mendelian laws?

Problem 4.3.2. If the pedigree is compatible with Mendelian laws, are the genotyping results plausible? If not, under which conditions are the genotypes plausible?

Problem 4.3.3. Consider that Alice, Olive Oyl, Eugene, and Popeye had not been genotyped (Figure 4.21). Are the observed genotypes plausible?

Problem 4.3.4. Consider that Peggy and Al Bundy have a third offspring, Bud (see Figure 4.22). Is this genotype distribution plausible?

Problem 4.4 Consider a diallelic marker with alleles A_1 and A_2 . In a sample of 140 individuals, 20 and 80 are homozygous for the A_1 and A_2 allele, respectively. Does the population deviate from HWE at the given marker?

Problem 4.5 Consider a diallelic marker with the alleles A_1 and A_2 . In a sample of 150 individuals, 32 and 52 are homozygous for the A_1 and A_2 allele, respectively. Does the population deviate from HWE at the given marker? Can compatibility with HWE be established?

Problem 4.6 How many simulations are required, approximately, to estimate a p -value with a precision of $e = 0.01$ at the 95% confidence level?

Problem 4.7 How many simulations are required, approximately, to estimate a p -value around 0.05 with a precision of $e = 0.01$ at the 99% confidence level?

Problem 4.8 In this problem, we consider differences in the variances of allele frequency estimates when there is HWD and when there is HWE.

Problem 4.8.1. Show that the variance in the estimated allele frequency $\hat{p}$ is equal to $\text{Var}(\hat{p}) = \frac{p(1-p)}{2n} + \frac{p_{11}-p^2}{2n} = \frac{p(1-p)}{2n} + \frac{D_{A_1}}{2n}$ in the general case, i.e., when there might be HWD.

Problem 4.8.2. Show that $\text{Var}(\hat{p})$ reduces to $\frac{p(1-p)}{2n}$ when HWE is fulfilled.

Problem 4.8.3. What happens to $\text{Var}(\hat{p})$ under HWD if $p_{11} = 0$?

Problem 4.9 Derive the asymptotic distribution of $\ln \hat{\omega}$.

Problem 4.10 Derive the sample size calculation formula given in Eq. (4.9).

URLs

ALTERTEST

see PREST

Affymetrix

http://www.affymetrix.com/support/downloads/manuals/500k_assay_manual.pdf

http://www.affymetrix.com/support/downloads/manuals/genotyping_console_manual.pdf

CLUSTERA

<http://www.medsci.uu.se/molmed/software.htm>

GENEHUNTER

<http://www.fhcrc.org/science/labs/kruglyak/Downloads/>

Genizon

http://www.genizon.com/english/services/snp_genotyping.html

GWA DATABASE

<http://www.genome.gov>

HWG-TEST

<http://www.imbs.uni-luebeck.de/software.html>

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO)

<http://www.iso.org>

MENDEL

<http://www.genetics.ucla.edu/software/mendel5>

PEDCHECK

<http://watson.hgen.pitt.edu/register/docs/pedcheck.html>

PREST

<http://galton.uchicago.edu/mcpeek/software/prest/>

GOFHWEX

http://zihost5.zi-mannheim.de/HWE_ConfLim/ExactConfBound_RelHetZyg.sas

SIMWALK2

<http://watson.hgen.pitt.edu/docs/simwalk2.html>

VASARELY

<http://s92417348.onlinehome.us/software/vasarely/index.html>

World Medical Association

<http://www.wma.net/e/policy/b3.htm>

5

Genetic Map Distances

In a genetic epidemiological study, from just a few to more than 1,000,000 different genetic markers may be used. Their order and distance are of great importance for subsequent statistical analyses. In this chapter, we discuss distances between genetic markers. We do not consider the problem of locus ordering using meiotic mapping in this book. The interested reader may refer to Jürg Ott's excellent text book [516] (see also Problem 7.1).

Several definitions for the term *genetic distance* have been given. For example, Chakraborty [111] relates the concept of genetic distance to the measurement of genetic difference between populations. This definition is thus designed to answer how dissimilar populations are with respect to their genetic compositions. As an alternative, genetic distance may denote *map distance* between genetic markers. This is the understanding we will follow in this chapter. At the beginning, we discuss physical distance (Section 5.1). In Section 5.2, we introduce the concept of genetic mapping, discuss various map functions, and consider the property of multilocus feasibility. The last section focuses on the newest approach for measuring genetic distance by linkage disequilibrium units.

5.1 WHAT IS PHYSICAL DISTANCE?

Physical distance is the most natural measure of genetic distance between two genetic loci. It is based on the molecular identification of the loci and a sequencing of the DNA strand between them. The unit for the physical distance is the basepair (*bp*) given in kilobases (1Kb = 1000 bases) or megabases (1Mb = 1000Kb). Because

DNA still contains introns, although they are deleted in transcribed RNA, physical distance usually differs between different transcription stages.

Physical distances are not used in genetic mapping, e.g., in linkage analysis (Chapter 7), because they do not fit in a statistical framework where probability theory is required. Therefore, such measures of genetic distance are preferable that adequately reflect either the probability of a crossing over in an interval between two loci or the probability of observing a recombination between the markers at the two loci. This leads to the genetic or meiotic map distance, which will be considered in the next section.

5.2 WHAT IS MAP DISTANCE?

In their early work on gene mapping, Sturtevant [646] and Morgan [475] assumed that genetic map distance is equivalent to the recombination fraction. There are, however, confounding factors in the analysis of the recombination fraction: multiple crossing overs and interference. Therefore, the concept has to be weakened. The basic assumption underlying the concept of *genetic map distance* is that the probability of a crossing over is proportional to the length of the chromosomal region. The recombination fraction naturally reflects this assumption because it was defined as the probability of observing a recombination, i.e., of having an odd number of crossing overs between two loci. This probability should increase with the underlying physical distance.

5.2.1 What is a distance?

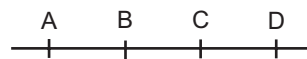


Fig. 5.1 Four loci, resulting in three chromosomal segments.

However, probabilities are not additive, and thus this measure is not a true distance measure. For illustration, consider Figure 5.1, where four loci are given, resulting in three segments. The recombination fractions for each segment are $\theta_{AB} = 0.3$, $\theta_{BC} = 0.2$, and $\theta_{CD} = 0.1$. The length of the whole segment from *A* to *D* cannot be the sum of the separate segment lengths because this would exceed the natural maximum of a recombination fraction, which is $\frac{1}{2}$. In contrast to probabilities, expected values of random variables are additive even if the random variables are not independent. Therefore, expected values are true distance measures as they can be constructed to fulfill the following assumptions:

1. Non-negativity: A distance cannot be negative.
2. Symmetry: The genetic distance between *A* and *B* is identical to the distance between *B* and *A*.
3. Triangle constraint: The sum of the distances *AB* and *BC* cannot be smaller than the distance *AC*.

The *map length* of a chromosomal segment—the genetic distance of its end points—is defined as the expected number of crossing overs taking place in this segment.

The dimension of *map distance* is Morgan (1 M = 100 centiMorgan = 100 cM). A segment of length 1 M exhibits on average one crossing over per meiosis.

5.2.2 What is the Haldane map function? What are other important map functions?

The question now is how a 1:1 relation between the map distance in Morgan and the recombination fraction θ can be established. For simplicity, we first assume the absence of interference (see Section 4.2.4). As an example, let us consider an interval of length $x = \frac{1}{3}\text{M} = 0.333\text{M}$ so that one crossing over is observed in every third meiosis on average. The probability of having $k = 0, 1, 2, \dots$ crossing overs in a single meiosis in the interval can be modeled by a Poisson distribution so that $P(X = k) = \frac{x^k}{k!} e^{-x}$. The probability of having no crossing over is $P(X = 0) = \frac{0.3^0}{0!} e^{-0.3} \approx 0.7165$. Similarly, one obtains $P(X = 1) \approx 0.2388$, $P(X = 2) \approx 0.0398$, $P(X = 3) \approx 0.0044$, and $P(X = 4) \approx 0.0004$. The probability for more than four crossing overs in this interval is approximately 0.000026 and therefore negligible. The recombination fraction θ is the probability of an odd number of crossing overs, thus $\theta = \sum_{j=0}^{\infty} P(X = 2j + 1)$. In this example, θ is approximately

$$\theta \approx P(X = 1) + P(X = 3) \approx 0.2388 + 0.0044 = 0.2433.$$

Correspondingly, the probability of observing no recombination is

$$1 - \theta \approx P(X = 0) + P(X = 2) + P(X = 4) \approx 0.7165 + 0.0398 + 0.0004 = 0.7567.$$

For the general case, the map distance x measured in Morgan M can be related to the recombination fraction θ by the *Haldane map function* [273]:

$$\theta = \frac{1}{2}(1 - e^{-2x}) \quad \text{with inverse} \quad x = -\frac{1}{2} \ln(1 - 2\theta). \quad (5.1)$$

With this map function, it is possible to transform non-additive recombination fractions to the additive Morgan scale. Distances from different intervals may then be added and transformed back to recombination fractions. This is demonstrated in the following example.

Example 5.1. Assume the presence of four genetic markers in the order $A-B-C-D$ and with distances $\theta_{AB} = 0.3$, $\theta_{BC} = 0.1$, and $\theta_{BD} = 0.15$. The map distance between markers A and B is

$$x_{AB} = -\frac{1}{2} \ln(1 - 2\theta) = -\frac{1}{2} \ln(1 - 2 \cdot 0.3) \approx 0.4581$$

using Eq. (5.1). Similarly, one obtains $x_{BC} \approx 0.1116$ and $x_{BD} \approx 0.1783$. The map distance between A and C then is $x_{AC} = x_{AB} + x_{BC} \approx 0.4581 + 0.1116 = 0.5697$, which can be converted to a recombination fraction using Eq. (5.1):

$$\theta_{AC} = \frac{1}{2}(1 - 2e^{-2x_{AC}}) \approx \frac{1}{2}(1 - 2e^{-2 \cdot 0.5697}) \approx 0.3400$$

Similarly, $x_{CD} = x_{BD} - x_{BC} \approx 0.1783 - 0.1116 = 0.0667$, yielding $\theta_{CD} \approx 0.0624$.

As noted above, Haldane's map function assumes the absence of interference. In the presence of positive interference, the probability of a double crossing over is reduced. This directly leads to a higher recombination fraction θ for a given number of expected crossing overs compared to the situation of no interference, as two crossing overs cannot be detected. The other extreme, complete interference is reflected through *Morgan's map function* $\theta = x$ [474]. Complete interference implies that one chiasma completely suppresses any other crossing over on the chromosome so that the probability of a crossing over is identical to the recombination fraction. In practice, the Morgan map function is useful only for small genetic distances, as a rule of thumb for $\theta \leq 0.1$.

Moderate interference is assumed by the *Kosambi map function* [372], which is given by

$$x = \frac{1}{4} \ln \frac{1 + 2\theta}{1 - 2\theta} \quad \text{with inverse} \quad \theta = \frac{1}{2} \frac{e^{4x} - 1}{e^{4x} + 1}.$$

It is probably the most popular and widely used map function in genetics since it seems to adequately reflect the level of interference observed in the human and other mammals.

A map function that includes both the Haldane map function and the Kosambi map function as special cases has been derived by Felsenstein [210]. It allows varying degrees of interference, with the strongest if the map function is identical to the Kosambi map function. The *Felsenstein map function* is given by

$$x = \frac{1}{2(2-k)} \ln \left(1 + \frac{2\theta(1-k)}{1-2\theta} \right), \quad 0 \leq k < 2, \quad \text{with inverse} \quad \theta = \frac{1}{2} \frac{e^{2x(2-k)} - 1}{e^{2x(2-k)} + 1 - k}.$$

With $k = 1$ the Felsenstein map function is identical to the Haldane map function, and for $k = 0$ it is equal to Kosambi's map function. Many other map functions have been proposed in the literature (see, e.g., Ref. [516]). However, they are of little importance in genetic mapping.

Figure 5.2 displays the conversion from map distances to recombination fractions for the three map functions by Haldane, Kosambi, and Morgan. It is easily seen that differences between map functions are negligible as long as the map distance is ≤ 0.1 M.

5.2.3 Is there a correspondence between physical and map distance?

On average, 1 cM corresponds to 0.88 Mb. However, the actual correspondence widely varies for different chromosomal regions. The reason for this is that the occurrence of crossing overs is not equally distributed across the genome. There are recombination hot spots and cold spots that show greatly increased and decreased recombination activities, respectively. In addition, chiasmata are more frequent in female than in male meiosis. Hence, the total map length is different between genders, with the most extreme deviations occurring in the pseudoautosomal region PAR1, i.e., the short arm of the X and the Y chromosomes. Males have an obligatory crossing over in this 2.6 Mb region, resulting in a length of 50 cM. The correspondence

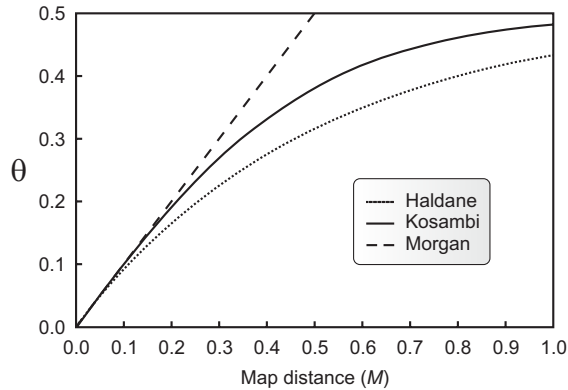


Fig. 5.2 Graphs of the conversion from map distance in Morgan M to the recombination fraction θ with the map functions by Haldane, Kosambi, and Morgan.

therefore is $1 \text{ Mb} \approx 19 \text{ cM}$. In females, however, the correspondence is $1 \text{ Mb} \approx 2.7 \text{ cM}$ for this region. The other extreme is that no crossing over can be observed in normal meiosis for the Y chromosome outside of the pseudoautosomal region. Here, $0 \text{ cM} \approx 56 \text{ Mb}$.

Though these extreme deviations can be observed for the sex chromosomes, useful rules of thumb regarding the autosomal genome are that 1 male cM averages 1.05 Mb, and 1 female cM averages 0.70 Mb, or even rougher, 1 cM equals approximately 1 Mb [636]. Furthermore, the total length of the human genome can be assumed to be approximately 3300 cM.

5.2.4 What is multilocus feasibility?

Almost all map functions have been derived for two adjacent chromosomal intervals, corresponding to three genetic markers. If, however, the recombination fraction is derived for more than two intervals, negative probabilities can turn out for some map functions and some recombination constellations. This will be illustrated below.

Assume the presence of four genetic markers in the order $A-B-C-D$ with distances $\theta_{AB} = 0.1$, $\theta_{BC} = 0.05$, and $\theta_{CD} = 0.2$. If recombination fractions are converted to map distances using the Kosambi map function, one obtains $x_{AB} = 0.1014$, $x_{BC} = 0.0502$, and $x_{CD} = 0.2118$. From these, all recombination fractions of interest can be computed.

One may be interested in the probability that recombinations occur in all three intervals, i.e., between A and B , between B and C , and between C and D . It is straightforward but quite cumbersome to compute this probability (see Problem 5.2 for the case of two intervals with three loci). In fact, a general formula to compute these probabilities for the case of n markers on a chromosome has been derived (see Problem 5.3 and Ref. [346]).

It turns out that the probability of a triple recombination is -0.0012 (see Problem 5.3.3), thus negative in the constellation described above. Obviously, a probability may not be negative, and therefore the use of the Kosambi map function results in an undefined probability in this example. If such a negative probability cannot

occur for any recombination constellation and an arbitrary number of markers given a specific map function, the map function is called *multilocus feasible*. This property has been investigated in detail e.g., by Liberman and Karlin [417], who have derived necessary and sufficient conditions for a map function to be multilocus feasible. These are summarized as follows.

Consider a map function $M(x)$ that is n times differentiable $M^{(n)}(x)$ for all non-negative distances x .

1. *Necessary condition for multilocus feasibility:* Suppose that $M(x)$ is multilocus feasible. Then

$$(-1)^n M^{(n)}(0) \leq 0 \quad \text{for all } n.$$

2. *Sufficient condition for multilocus feasibility:* A map function $M(x)$ is multilocus feasible if their derivatives fulfill the inequality

$$(-1)^n M^{(n)}(x) \leq 0 \quad \text{for all } n.$$

Example 5.2. The Haldane map function $\theta = \frac{1}{2}(1 - e^{-2x})$ has derivatives

$$M^{(n)}(x) = (-2)^{n-1} e^{-2x}.$$

Because $(-1)^n M^{(n)}(x) = -2^{n-1} e^{-2x}$, the sufficient condition for multilocus feasibility is fulfilled.

Example 5.3. The Kosambi map function is not multilocus feasible (see Problem 5.4).

Though multilocus feasible map functions seem to be preferable at first glance, they have been criticized for being generally unrealistic [194]. In applications involving multiple markers, loci are usually analyzed assuming an absence of interference [623]. Therefore, estimated recombination fractions are corrected at the end using an appropriate map function. It is important to note that this obviously illogical approach results in estimated map distances that are similar to those that would be obtained assuming more appropriate stochastic processes.

5.3 WHAT ARE LINKAGE DISEQUILIBRIUM UNITS?

The localization and identification of disease genes—a process termed *positional mapping*—by linkage analysis (see Chapter 7) was and is very successful for Mendelian diseases that have a strong genetic effect and are relatively rare. For the identification of complex disease genes (see Chapter 2), however, greater effort is generally required. After an initial linkage finding, fine mapping is carried out to narrow down the putative disease locus. This is usually done by association analyses, in the context of fine mapping often termed *linkage disequilibrium (LD) mapping*, involving single nucleotide polymorphisms (SNPs). For statistical analysis, haplotypes frequently are

Table 5.1 Summary of haplotype frequencies at two single nucleotide polymorphisms (SNPs).

	SNP marker		Total
	b	B	
a	$p_{11} = (1 - p_A)(1 - p_B) + \mathcal{D}$	$p_{12} = (1 - p_A)p_B - \mathcal{D}$	$1 - p_A$
A	$p_{21} = p_A(1 - p_B) - \mathcal{D}$	$p_{22} = p_A p_B + \mathcal{D}$	p_A
Total	$1 - p_B$	p_B	1

of primary interest rather than single markers (for details, see Chapter 13). Therefore, a new kind of genome map, a metric LD map, has been established [382, 482].

Traditionally, LD refers to the non-random association between two alleles at two loci on a chromosome in a natural breeding population, as depicted in Table 5.1. Consider an SNP marker with alleles A and a , where p_A denotes the allele frequency of A . If this SNP marker is not in LD with a second marker having alleles B and b with allele frequency p_B for allele B , the frequency p_{11} of the haplotype AB equals $p_A p_B$. If there is LD, haplotype frequencies are distorted by $\mathcal{D}$ that is the covariance between the two SNP markers. Below we need the *LD probability*

$$\eta = \frac{\mathcal{D}}{p_A(1 - p_B)},$$

which is identical to the usual Lewontin $\mathcal{D}'$ measure if $p_A \leq p_B$ and $p_A \leq 1 - p_A$ (see Chapter 10 for a detailed discussion on $\mathcal{D}'$ and a different interpretation of LD in terms of genotypes).

Before η can be interpreted, we have to note that the four haplotype probabilities p_{ij} from Table 5.1 may be written as functions of η :

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \eta \begin{bmatrix} 1 - p_B & p_B - p_A \\ 0 & p_A \end{bmatrix} + (1 - \eta) \begin{bmatrix} (1 - p_A)(1 - p_B) & (1 - p_A)p_B \\ p_A(1 - p_B) & p_A p_B \end{bmatrix} \quad (5.2)$$

Equation (5.2) shows that, under certain conditions, η is the probability that a random haplotype has descended without recombination from a founder population in which p_{21} was 0, whereas the complementary class with frequency $1 - \eta$ has undergone at least one crossing over between the loci and thus the alleles are independent (for a detailed discussion, we refer the reader to Refs. [482]). Furthermore, Eq. (5.2) shows that $\eta = 0$ in the absence of LD and $\eta = 1$ in the case of complete LD.

LD maps are based on recombination events exactly as genetic map functions are. The LD η is, however, influenced not only by recombination but also by processes including mutation and/or selection. Furthermore, instead of investigating recombination from one generation to another to establish the map, many recombination events have accumulated because chromosomes have been transmitted across many, say n , generations (see Figure 5.3).

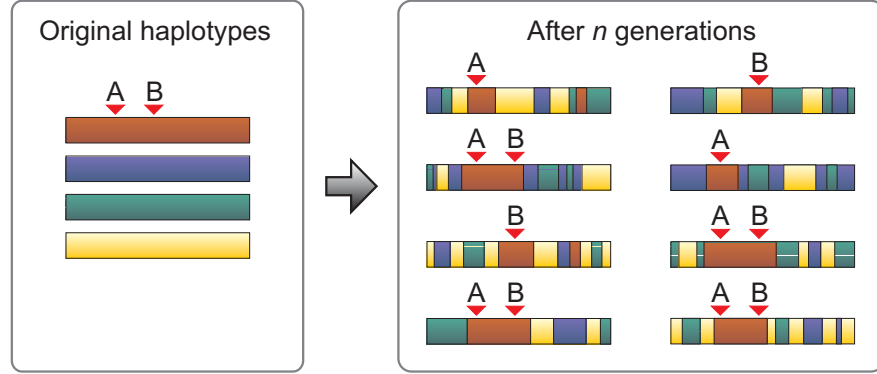


Fig. 5.3 Original haplotypes and haplotypes after n generations.

Under reasonable assumptions, Morton and colleagues [482] have shown that the LD probability η between two loci can be expressed by the *Malecot model* ([434], p. 84)

$$\eta_t = (1 - L)W e^{-\theta t} + L. \quad (5.3)$$

Here, θ is a small recombination fraction, and t is the number of generations, as the population frequency of the rarest two-marker haplotype was minimal. L is the residual LD at large distance, and W reflects a monophyletic or polyphyletic origin [165] of susceptible haplotypes and is 1 if there is a unique susceptible haplotype and marker mutation is negligible, and less than 1 otherwise. In general, t exceeds 100 generations, and therefore $e^{-\theta t}$ is negligible unless θ is so small that θt is proportional to the map distance x in cM. In this case θt and εx are interchangeable [482], where ε is the exponential decline of the LD η_x with map distance x . Therefore, Eq. (5.3) may be rewritten as

$$\eta_x = (1 - L)W e^{-\varepsilon x} + L. \quad (5.4)$$

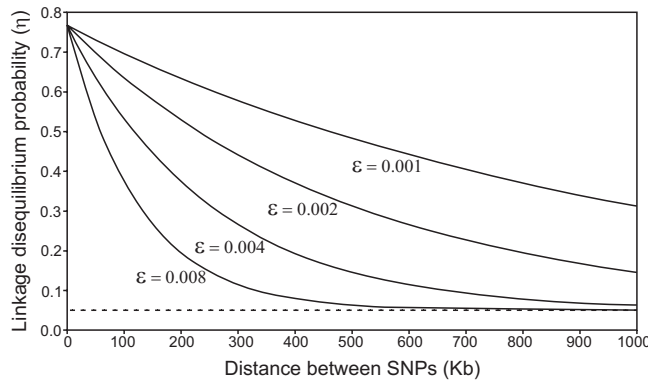


Fig. 5.4 The linkage disequilibrium probability η is described as a function of distance x in Kilo bases (Kb) between loci for varying swept radii ε . Graphs illustrate the decline of linkage disequilibrium probability η with distance Kb assuming $W = 0.75$ and $L = 0.05$.

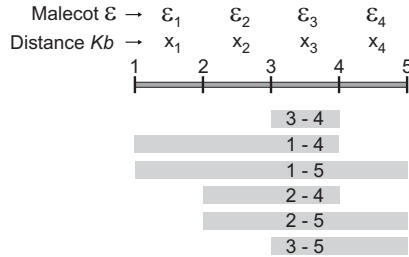


Fig. 5.5 To obtain an estimator for ε_3 for the interval 3–4, one can use the pairwise information for the interval 3–4 as well as for all intervals including the interval 3–4.

This *Malecot equation* helps in explaining differences in LD between populations and chromosomal regions. Indeed, εx is a useful metric for LD studies. Because it is more accurately known than θt , it should be preferred in estimation. Whether the Malecot equation is expressed in terms of t , cM, or Kb—which can be used as an alternative distance measure for cM—the L parameter still predicts LD between unlinked loci.

The construction of an LD map under the Malecot model requires estimation of ε in each map interval and uses this to construct an LD scale, as it is a family of exponential curves as shown in Figure 5.4 in regions of high LD [382]. Given n markers $i = 1, 2, \dots, n$ on the LD map, let the length of the i th interval be $\varepsilon_i x_i$ LD units (LDU), where ε_i estimates the Malecot parameter and x_i is the length of the interval on the

Table 5.2 Linkage disequilibrium parameters of a genome scan in humans.

Mean ε	LDU ^a for the genome	Swept radius	
		Kb^b	LDU ^a
0.001	3300	1000	1
0.002	6600	500	1
0.005	16,500	200	1
0.01	33,000	100	1
0.1	330,000	10	1

The swept radius in Kilobases (Kb) is the inverse of the mean ε .

^a Linkage disequilibrium unit, ^b kilobases

physical map in Kb . The Malecot model is fitted for every interval using all of the pairwise data that are informative for the interval. To obtain an estimator for ε_3 for the interval 3–4 in Figure 5.5, one can use the pairwise information for the interval 3–4 as well as for all intervals including the interval 3–4, i.e., 1–4, 1–5, 2–4, 2–5, and 3–5. Map distances in LDU are established iteratively as the product of the length of an interval x_i in Kb and the ε_i in this interval, thus $\varepsilon_i x_i$ for the i th interval with a region having $\sum_i \varepsilon_i x_i$ LDU [382]. The mean ε for the region is $\sum_i \varepsilon_i x_i / \sum_i x_i$, and its inverse is termed the *swept radius* $1/\varepsilon$, which gives the extent of useful LD, with $0 < \varepsilon < 1$ on the Kb scale and $\varepsilon = 1$ for a standard LD map. One of the elegant properties of the model is that recombination affects only ε but not W , whereas mutation and drift systematically affect only W . The asymptote L can be estimated or predicted (L_p) from the information about $\hat{\eta}$ and is proportional to sample size [482].

Table 5.2 displays LD parameters in a human genome-wide scan that has a length of approximately 3300 Mb, equaling $3.3 \cdot 10^6$ Kb. With a mean ϵ of 0.005, one can expect 16,500 LDU for the whole genome.

Example 5.4. In this example, we reanalyze published data [341, 740]. Jeffreys et al. [341] genotyped 296 SNPs in a 216 Kb segment of the class II region of the major histocompatibility complex (MHC) on chromosome 6p21.3 using the DNA from 50 unrelated north European semen donors. They determined the pattern of LD using Lewontin's D' (see Section 5.3 and Chapter 10) and the sperm crossing over activity, while Zhang et al. [740] estimated the LD map of chromosome 6p21.3.

Figure 5.6 depicts the heat map, as generated with GOLD, based on D' for the region on 6p21.3. The three largest blocks show good agreement with the recombination hot spots (Figure 5.7b) observed by Jeffreys and colleagues. The LD map (Figure 5.7a) corresponds well with the heat map. Furthermore, the three major steps in the LD map coincide with the empirically determined recombination hot spots. The reader should note that the advantage of the LD map in contrast to haplotype block definition approaches is that the LD map is additive and relies not on specific but rather mostly arbitrary assumptions for block definitions (see Chapters 10 and 13).

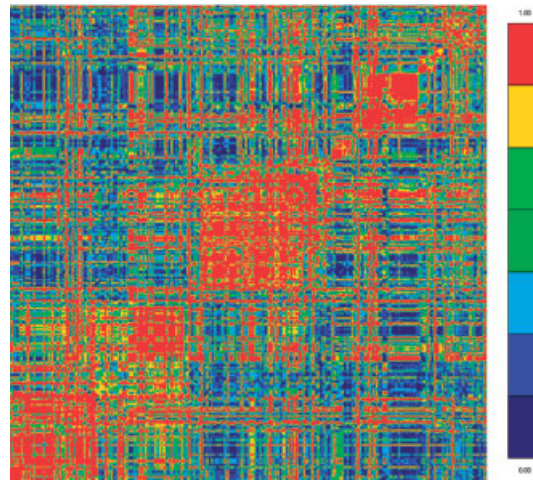


Fig. 5.6 Heat map for the class II region of the major histocompatibility complex (MHC) in 50 north Europeans based on Lewontin's D' . The higher the linkage disequilibrium between two loci, the more red the color of the spot in the heat map.

In summary, we have illustrated that the LD map shows steep increases at chromosomal regions with empirically determined recombination hot spots.

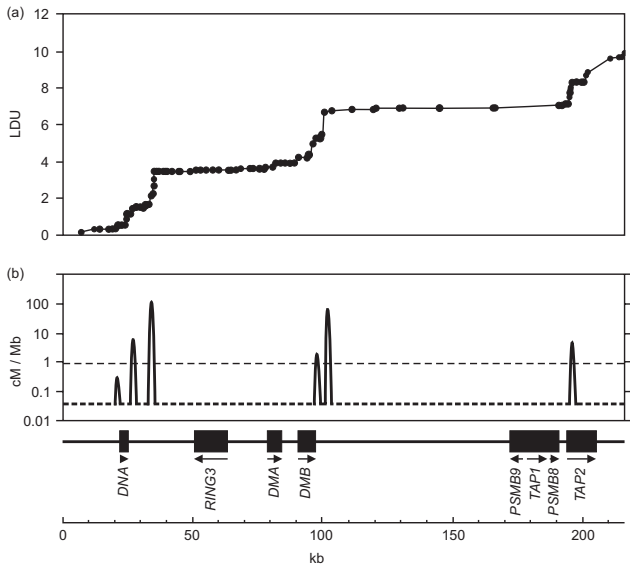


Fig. 5.7 Linkage disequilibrium map of chromosome 6p21.3 reported by Zhang et al. [740] (a) and meiotic recombination (b) reported by Jeffreys et al. [341] oriented from *p*ter to *q*ter. The dotted line is a rough estimate of recombination within the major blocks defined by the recombination hot spots as $\text{cM/Mb} = 0.04$. The data were kindly provided by Sir Alec J. Jeffreys and Andrew Collins.

5.4 PROBLEMS

Problem 5.1 Use the Haldane map function for conversions.

- Problem 5.1.1.** $\theta_{AB} = 0.05$, $x_{AB} = ?$
- Problem 5.1.2.** $x_{AB} = 2cM$, $\theta_{AB} = ?$
- Problem 5.1.3.** $\theta_{AB} = 0.05$, $\theta_{BC} = 0.085$, $\theta_{AC} = ?$
- Problem 5.1.4.** $\theta_{AB} = 0.05$, $\theta_{AC} = 0.085$, $\theta_{AB} = ?$

Problem 5.2 Consider three loci *A*, *B*, and *C*. The two-point recombination fractions are denoted by θ_{AB} and θ_{BC} . Furthermore, g_{ij} , $i, j = 0, 1$ denote the joint probabilities for recombination *R* and non-recombination *NR*. Thus, g_{11} is the probability that a recombination occurs in both intervals *A*–*B* and *B*–*C* (Table 5.3). The probability of a recombination between *A* and *B*, irrespective of a recombination event between *B* and *C*, is $\theta_{AB} = g_{11} + g_{10}$. Similar conditions can be formulated for the intervals *B*–*C* and *A*–*C*. Invert the relations from the g_{ij} 's to the θ 's to obtain joint recombination probabilities in terms of two-point recombination fractions. What are the restrictions for the recombination probabilities to ensure that the g_{ij} are non-negative?

Table 5.3 Joint probabilities g_{ij} and recombination fractions θ for recombination *R* and non-recombination *NR* among loci *A*, *B*, and *C*.

Loci <i>A</i> and <i>B</i>	Loci <i>B</i> and <i>C</i>		Total
	<i>R</i>	<i>NR</i>	
<i>R</i>	g_{11}	g_{10}	θ_{AB}
<i>NR</i>	g_{01}	g_{00}	$1 - \theta_{AB}$
Total	θ_{BC}	$1 - \theta_{BC}$	1

Problem 5.3 Suppose there are n markers, and the recombination event of interest is denoted by the column vector $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_{n-1})'$, where

$$\epsilon_i = \begin{cases} 1, & \text{recombination in interval } i \\ 0, & \text{no recombination in interval } i \end{cases}$$

for the $n-1$ possible recombination intervals. There are 2^{n-1} possible combinations of recombination events that can be summarized in a $2^{n-1} \times (n-1)$ matrix Δ

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1(n-1)} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{2^{n-1},1} & \delta_{2^{n-1},2} & \dots & \delta_{2^{n-1},n-1} \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{2^{n-1}} \end{pmatrix}$$

with the row vector $\delta_i = (\delta_{i1}, \delta_{i2}, \dots, \delta_{i(n-1)})$ and

$$\delta_{ij} = \begin{cases} 1, & \text{recombination,} \\ 0, & \text{no recombination,} \end{cases} \quad i = 1, \dots, 2^{n-1}.$$

Let r_{δ_i} be the recombination fraction for the recombination event δ_i . The probability for the recombination event ϵ is then given by

$$P(\epsilon) = \frac{1}{s^{n-1}} \cdot \sum_{i=1}^{2^{n-1}} (-1)^{s_i} (1 - 2r_{\delta_i}), \quad \text{where } s_i = \sum_{j=1}^{n-1} \delta_{ij} \epsilon_j = \delta'_i \epsilon. \quad (5.5)$$

Problem 5.3.1. Consider the case of three loci as in Problem 5.2. Show that $P(\epsilon) = P((1, 1)') = \frac{1}{2}(\theta_{AB} + \theta_{BC} - \theta_{AC})$ using Eq. (5.5). Note that this probability is denoted g_{11} in Problem 5.2.

Problem 5.3.2. Consider the case of four loci and show $P(\epsilon) = P((1, 1, 1)') = \frac{1}{4}(\theta_{AB} + \theta_{BC} + \theta_{CD} - \theta_{AC} - \theta_{BD} - \theta_{AB,CD} + \theta_{AD})$ using Eq. (5.5). Here, $\theta_{AB,CD}$ denotes the probability that recombinations occur in the intervals $A-B$ and $C-D$. It may thus be computed by $\theta_{AB,CD} = M(x_{AB} + x_{CD}) = M(M^{-1}(\theta_{AB}) + M^{-1}(\theta_{CD}))$.

Problem 5.3.3. Use the formula from Problem 5.3.2 to show that $P((1, 1, 1)) = -0.0012$ if $\theta_{AB} = 0.1$, $\theta_{BC} = 0.05$, and $\theta_{CD} = 0.2$ when the Kosambi map function is used.

Problem 5.4 Show that the Kosambi map function is not multilocus feasible. Hint: Consider $M^{(3)}(0)$.

URLs

GOLD

<http://www.sph.umich.edu/csg/abecasis/GOLD/>

6

Familiality, Heritability, and Segregation Analysis

Genetic epidemiological research for a specific disease or quantitative trait followed a systematic approach for identifying the genetic basis of simple or complex genetic disease until quite recently; see, e.g., the recommendation of Ref. [185]. To this end, the following 7 points had to be addressed one by one:

1. Familial correlations and/or recurrence risk ratios: Has the disease or phenotype of interest a familial component, or, phrased differently: Is there familial aggregation?
2. Twin or adoption studies or family studies using extended pedigrees: What are the reasons for familial aggregation? Can the familial aggregation be explained by shared environment or by genetic factors? How big is the proportion of variance attributable to genes?
3. Segregation analysis: Is there evidence for a major gene? Is the segregation pattern following Mendelian inheritance for this major gene?
4. Linkage analysis (Chapters 7–9): Where are the genes or, phrased differently, in which chromosomal region is the functionally relevant variant?
5. Association analysis (Chapters 10–14): What are the genes or, phrased differently, which specific variant or specific haplotype is associated with the trait of interest?
6. Risk estimation, i.e., relative risks or genotype relative risks (Chapters 11 and 14): How big is the risk associated with the risk factor?
7. Functional studies: Is the identified associated variant or haplotype functionally relevant? What functional relevance it has?

For a complex genetic disease, inconsistent or unclear results from a segregation analysis have not hindered further linkage and/or association analyses. However, evidence for a genetic basis for the trait of interest was required, i.e., positive answers to questions 1 and 2. Meanwhile, this flow of genetic epidemiological studies has changed substantially with the advent of a genome-wide association (GWA) studies. Quite often steps 2 and 3, and in some cases, the first four steps are skipped. As a consequence, the classical twin studies, adoption studies or family studies in nuclear or extended pedigrees not using genetic markers are not in vogue, anymore. And there are several arguments why researchers start with association analysis in unrelated individuals straight away.

- Steps 1 to 4 can only be investigated in family studies. The ascertainment of families is substantially more expensive and time consuming than the conduct of non-family-based cohort or case-control studies (see Chapter 10). The substantial decrease in genotyping cost is an important aspect.
- Many population-based studies are available as biobanks. As a consequence, on the molecular genetic level not only a single disease but also a huge number of phenotypes are investigated.
- It will be shown below that family studies for steps 2 and 3, i.e., investigating familiarity and segregation patterns, rely on strong assumptions that are usually not met in applications.

Although family studies suffer from being expensive and time consuming, they still are of great importance, and the two main arguments are [132]:

- Evidence that a disease is inherited, i.e., that it is transmitted from generation to generation, requires family data.
- Genetic particularities like genomic imprinting (see Chapter 2), anticipation (see Chapter 2) and epigenetic phenomena can only be detected in family studies. Here, *epigenetics* means changes in a trait, e.g., gene expression, which is caused by mechanisms other than changes in the underlying DNA sequence.

In this chapter, we discuss study approaches to answer the first three questions in some detail because of their historical importance. It is important to distinguish between studies that use reported family history and other studies in which information has been obtained personally from every family member. We therefore start by comparing the family history method with the family study method in Section 6.1. Approaches for investigating familial resemblance are described in Section 6.2. These are either based on recurrence risk ratios or the analysis correlation coefficients between relative pairs. Although approaches for investigating familial resemblance give a basic impression whether familial correlations are high or low, they do not allow separating the effect of genes and fridges. In Section 6.3, we therefore discuss the concept of heritability which has been developed for estimating the proportion of variance attributable to genetic effects. This lays the basis for the subsequent Section 6.4 in which we consider different twin studies and adoption studies for estimating heritability. In Section 6.5, we critically discuss the limitations of the

different study designs and the resulting estimators. Finally, we consider segregation analysis which has been and still is used for estimating major gene effects.

6.1 WHAT IS THE DIFFERENCE BETWEEN THE FAMILY HISTORY METHOD AND THE FAMILY STUDY METHOD?

Two types of family studies need to be distinguished, the *family history method* and the *family study method*, which have been discussed in detail by Khoury et al. [353]. In the family history method, investigators obtain information from individual study participants concerning the presence or the absence of disease in his/her family members. When study participants are unavailable, this information is usually obtained from family members such as spouses, parents or other close relatives. In children, family history information is usually obtained from their parents.

The family history method is prone to a series of errors. First, information about presence or absence of the disease in relatives is not validated.

Second, in case-control studies (see Chapter 10), there might be differential recall between cases and controls. If cases remember more affected relatives than do controls, estimates are biased. This is of special importance in second- or third-degree relatives because of the concernment of the index case. Differences in the increased recognition and reporting of a disease in the family may, however, depend on the disease. Some diseases are distressing to the patient and his/her close relatives so that they do not want to communicate about the disease at all, including more distant relatives.

While the family history method provides a quick overview on the familiarity of a disease, a positive family history can be used merely as an exposure variable in a case-control study. In most studies, a family history is defined to be positive if at least one first degree relative to the proband is affected. In general, a test is carried out whether the positive family history has an effect on the disease risk in the proband. However, family history is not a simple to measure attribute. Especially, the validity has to be questioned [488]. Finally, the probability of a positive family history depends on factors, including family size, and even the entire family structure. As illustrated by Khoury et al. [353], if cases tend to come from larger families than controls, the probability of finding at least one affected relative will be larger in case families than in control families. Finally, we note that this information is not always collected and/or analyzed.

An alternative to the family history method is the family study approach. Here, the proband forms the basis, and relatives to the proband are enrolled in the study. These additional family members are interviewed, and their medical records are obtained. Of course, this study is expensive, but it has several major advantages compared to the family history method [353]:

1. The disease can be directly evaluated in family members, and a diagnosis can be made at an early stage of a disease using a common medical examination protocol.

2. Environmental risk factors can be directly recorded, biochemical laboratory parameters can be assessed. This leads to a simpler dissociation of other factors.
3. Tissue, including blood, can be collected making linkage and association analysis possible.

The difference in estimates between the family history method and the family study approach is illustrated in the following example.

Example 6.1. *There is great variability in the estimates of familial cases for pancreatic cancer (PC), and the proportion of familial cases ranges from 2.7 to 16% [46]. One major reason for this variability and the relatively high point estimates may be that the majority of studies relied on reported cancers rather than cancers confirmed by medical records in the family, which may entail considerable false reporting.*

Table 6.1 Prevalence of familial pancreatic cancer (FPC) in German pancreatic cancer patients ($n = 479$).

	No. of FPC cases	Percentage	95% CI
Confirmed by histological records	5	1.05	0.34–2.42
Confirmed by medical or histological records	9	1.88	0.86–43.54
Confirmed by telephone interview, medical, or histological records	17	3.55	2.09–5.63
Positive family history reported by the case proband	23	4.81	3.07–7.13

Reported are absolute numbers, percentages and the corresponding 95% confidence interval (CI). Table from Ref. [46].

From July 1999 to December 2001, 479 patients with confirmed PC, who were treated in 18 German medical or surgical departments, were enrolled in a prospective observational study to estimate the prevalence of familial PC [46]. All patients were interviewed using a standardized questionnaire with regard to family structure and family history, and the histology was confirmed. The family history was considered positive for PC if at least two first-degree relatives were affected with ductal adenocarcinoma of the pancreas. A reported family history for PC was verified by consulting the original pathology specimen and medical records whenever possible. The family history of patients without retrievable medical records of affected relatives was confirmed by a standardized telephone interview of at least two first-degree relatives. All PC patients who reported a positive family history were counseled by an experienced geneticist, and a 3-generation pedigree was prepared. A full medical history was taken, and all tumors within each family including age of onset were documented. Seventy-three patients had to be excluded from further analysis because of several reasons including pancreatic tumors other than ductal adenocarcinoma.

Detailed findings of the study are displayed in Table 6.1. Twenty-three of 479 (prevalence 4.8%, 95% confidence interval (CI) 3.1–7.1) patients reported at least one first-degree relative with PC. The familial aggregation could be confirmed by histology in 5 of 23 patients (1.1%, 95% CI 0.3–2.4), by medical records in 9 of 23 patients (1.9%, 95% CI 0.9–3.5), and by standardized interviews of first-degree relatives in 17 of 23 patients (3.5%, 95% CI 2.1–5.6).

The prevalence found in this study is much lower than reported by all previous studies; for a detailed discussion, see Ref. [46]. It is important to note that the gold standard is to validate a reported history of cancer by review of medical records and death certificates. However, despite extensive efforts, medical documentation of potentially affected family members could be obtained in the German multicenter study in only 60% of the cases as reported by other groups. The prevalence of familial PC among all PCs is therefore probably underestimated when only cases confirmed by medical records are considered because true cases will be excluded. The prevalence is certainly overestimated by considering all cases with a positive family history reported by the index case because the reported history may be wrong—as it was for one potential familial case in the study. However, the accuracy of reporting cancers in first-degree relatives has been found to be generally correct, in particular if the cancer diagnosis was independently confirmed by several family members. Thus, the best approximation to the true prevalence of familial PC might be to include those cases which could be confirmed by standardized interviews of first-degree relatives. Based on this approach, the prevalence of familial PC in Germany would be 3.5% (95% CI 2.1–5.6%), which is much lower than the postulated prevalence in the literature.

6.2 HAS THE PHENOTYPE OF INTEREST A FAMILIAL COMPONENT? WHAT ARE RECURRENCE RISK RATIOS?

The first step in genetic research is the investigation of familial resemblance, and case reports of families often are the first evidence for a genetic component of a disease. For example, the first reports on familial aggregation of dyslexia appeared in 1905, a decade after the first single cases had been described [215]. The first large systematic family study for dyslexia was published as early as in 1950, and a series of other family studies followed [215, 278]. Meanwhile, several genes have been identified, and for a review the reader may refer to Ref. [244].

6.2.1 Has the phenotype of interest a familial component?

Several study designs have been proposed for estimating the degree of familiarity; for overviews, see, e.g., Refs. [414, 559]. *Familial resemblance* describes the situation in which relatives are phenotypically more similar than unrelated individuals. For example, Fig. 6.1 shows that relatively small variability is observed for human sizes



Fig. 6.1 Variation in human sizes. Shown are on the left-hand side two types of little people, together with a giant, a super obese lady, and a man of average size. In the first row on the left are the “Dolls”, three sisters and a brother, born in Germany of normal-sized parents whose four other children were also normal-sized. Next to the midgets are two unrelated achondroplastic dwarfs born in the United States. The tall person (7 feet, 4 inches) was born in Germany, and the tallness ran in his family. The super obese lady weighed 540 pounds and was born in the United States. The normal size person is Amram Scheinfeld (5 feet, 9 inches). He is author of the book from which the photo is reprinted, see Ref. [588]. On the right-hand side two monozygotic twin-brothers are shown married to monozygotic twin sisters, with one couple parents of monozygotic twin daughters, the other of a son born four days earlier. The twin couples shared the same home in New York, and raised their children as brothers and sisters, which genetically they are. For, while legally the girls and the boy were double first cousins, the fact that their parents are monozygotic twins make them biologically as much sisters and brothers as any children of the same family. Interestingly, the children until they were 8 years old did not differentiate between the two sets of parents and, while knowing which were their true parents, continued to refer impartially to either woman as “mother” and either man as “father”. Figure preprinted from Ref. [588].

among genetically identically or genetically similar individuals. At the same time, genetically unrelated individuals may vary substantially in human sizes.

Familial resemblance of quantitative traits can be assessed more formally by comparing pairwise correlation coefficients. For example, Fig. 6.2 displays the correlation coefficients for intelligence test scores for different relative pairs from family studies. The figure clearly shows that the correlation increases with the degree of relatedness. Thus, there is familial resemblance for intelligence test scores.

For diseases, the comparison of correlation coefficients between different relative pairs also serves as an indicator whether genetic factors might play a role. For example, to investigate familial aggregation we estimated the correlation between sib-pairs, parent–offspring pairs, and spouses using data from an epidemiological study of esophageal cancer in a high-incidence area in China [762]. If a trait has an important genetic component, the parent–offspring correlation should be high as

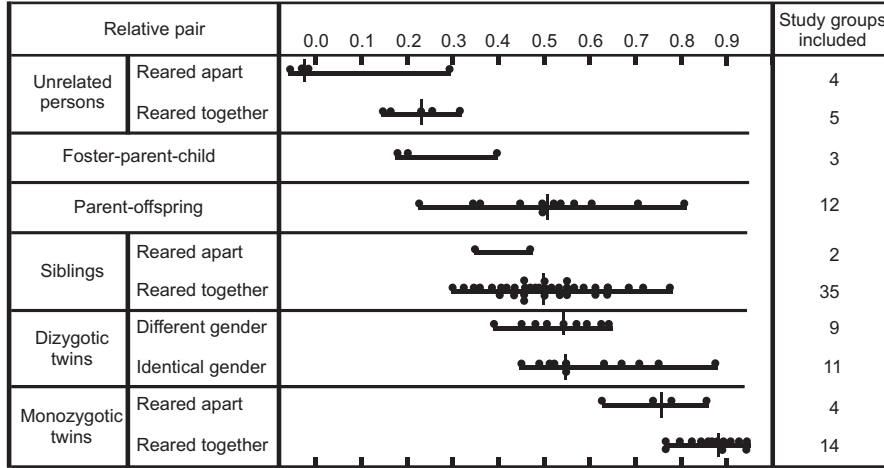


Fig. 6.2 Correlation coefficients for intelligence test scores from 52 different studies. Some studies reported data from more than one pair of relatives, giving a total of 99 groups. Correlation coefficients are indicated by circles for each study; medians are shown by vertical lines; horizontal lines represent the ranges. Figure created according to Ref. [108].

should be the sib-pair correlation because these pairs share 50% of their genetic material, at least on average. Our analysis revealed low correlations between parent-offspring pairs of 0.05 as well as between sib-pairs of 0.06 compared to the relatively high correlation between the genetically unrelated parents of 0.32. These results suggest a more important role of environmental factors for the majority of esophageal cancer cases in this study population.

Of course, this similarity can arise not only through shared genes but also through shared environment. However, the aim of all these approaches is to estimate the proportion of variance attributable to genes, i.e., the heritability.

6.2.2 What are recurrence risk ratios?

An alternative to the correlation coefficient approach is the recurrence risk ratio approach that is commonly used for investigating the familiarity of diseases. Recurrence risk ratios have been introduced by Risch [561], and they are therefore also known as *Risch's λ -values*. Recurrence risk ratios can be defined for different types of relatives of a proband, and it is a very general method. For example, it allows to consider several loci simultaneously. However, in the first step, we restrict the model to a single disease susceptibility locus.

Let the random variables y_1 and y_2 denote the affection status of two individuals with relationship type R . Now define the recurrence risk $K_R = P(y_2 = \text{affected} | y_1 = \text{affected})$. This gives the probability that a type R relative of an affected individual is also affected. With $K = E(y_1) = P(y_1 = \text{affected})$ being

the population prevalence, the *recurrence risk ratio* λ_R for a type R relative of an affected individual is given as the risk ratio

$$\lambda_R = \frac{K_R}{K}.$$

Though the definition of λ_R does not coincide with a classical definition of a risk ratio from epidemiology where a time component is explicitly included (note that K is the population prevalence and not an incidence), it is extremely useful for applications [14, 766]. For this, the concept needs to be extended to the multilocus case.

The second step therefore is an extension to two-locus models. The penetrance now becomes a function of two possibly interacting loci L_1 and L_2 . Two types of two-locus models have received specific attention because these can be described as functions of the locus-specific recurrence risk ratios. The first model is a multiplicative model, representing epistasis, i.e., interaction among loci (see Chapters 2 and 11). The second model is additive in nature, similar to a model with genetic heterogeneity (see Chapter 2) and therefore characterized by no locus interaction. The penetrances according to these models are

$$\text{Multiplicative: } f(g_1, g_2) = P(\text{aff} | g_1, g_2) = f_1(g_1) \cdot f_2(g_2).$$

$$\text{Additive: } f(g_1, g_2) = P(\text{aff} | g_1, g_2) = f_1(g_1) + f_2(g_2).$$

Here, G_1 and G_2 denote the genotypes at the disease loci 1 and 2, respectively. For example, if the two loci are diallelic with alleles D_l, N_l for $l = 1, 2$ and if they act in a dominant fashion, one obtains $f_1(D_1, D_1) = f_1(D_1, N_1) = 1$ and $f_1(N_1, N_1) = 0$ for the first locus. If the two loci act in a recessive manner, $f_1(D_1, D_1) = 1$ and $f_1(D_1, N_1) = f_1(N_1, N_1) = 0$. The terms for the second locus are defined analogously to those of the first locus.

With

$$K = \sum_{k=0}^2 \sum_{l=0}^2 p_{1k} p_{2l} f(g_1 = k, g_2 = l)$$

denoting the population prevalence, where p_{1k} and p_{2l} are the genotype frequencies at the first and second locus for genotypes k and l , respectively, the multiplicative model becomes

$$K = \sum_{k=0}^2 \sum_{l=0}^2 p_{1k} p_{2l} f_1(k) f_2(l) = \left(\sum_{k=0}^2 p_{1k} f_1(k) \right) \left(\sum_{l=0}^2 p_{2l} f_2(l) \right) =: K_1 \cdot K_2.$$

K_1 and K_2 have been termed *prevalence factors*, as they are defined in terms of the penetrance factors $f_1(k)$ and $f_2(k)$. Risch [561] has shown that $K_R = K_{1R} \cdot K_{2R}$ for the multiplicative model, where K_{1R} and K_{2R} are defined analogously to the single locus case but now using $f_i(G_l)$ instead of the single-locus penetrances f_j . Thus, the recurrence risk ratio λ_R can be written in terms of the locus-specific recurrence risk ratios λ_{1R} and λ_{2R} for the multiplicative genetic model. They are given by

$$\lambda_R = \lambda_{R1} \cdot \lambda_{R2}.$$

Here, the recurrence risk ratio λ_R is the product of the *recurrence risk ratio factors* $\lambda_{R1} = K_{R1}/K_1$ and $\lambda_{R2} = K_{R2}/K_2$.

For the additive genetic model, one similarly obtains

$$\lambda_R = \left(\frac{K_1}{K}\right)^2 (\lambda_{R1} - 1) + \left(\frac{K_2}{K}\right)^2 (\lambda_{R2} - 1).$$

Thus, for an additive genetic model the recurrence risk ratio λ_R is a weighted sum of terms from each contributing locus.

With these results, it is straightforward to extend the two-locus case to the multi-locus model. For example, if L loci determine the disease in a multiplicative fashion, the recurrence risk ratio is given by $\lambda_R = \lambda_{R1} \cdot \lambda_{R2} \cdot \dots \cdot \lambda_{RL}$, where the λ_l are analogously defined as above. The reader should note that both the multiplicative and the additive model are considered only for their mathematical elegance. In fact, a multilocus disease model will most likely be neither strictly multiplicative nor additive [561]. Nevertheless, the overall and the locus-specific recurrence risk ratios are the key to sample size calculations. Because λ are multiplicative in the multiplicative model, a common assumption is that either a single-locus model holds or multiple loci act in a multiplicative fashion.

Table 6.2 summarizes λ_S estimates from a broad variety of diseases. This gives the reader an impression of the magnitude that can be expected in genetic studies. It can clearly be seen that the magnitude of the λ values substantially varies.

Table 6.2 Sibling recurrence risk ratio λ_S estimates as reported for dichotomous traits.

Phenotype	λ_S	Reference	Phenotype	λ_S	Reference
AITD ^a	16.9	[678]	Hypertension	4	[20]
Alcohol dependence	4	[20]	IDDM ^e	15	[560]
Asthma	3	[20]	Multiple sclerosis	20–30	[20, 560]
Autism	75	[20]	Open-angle glaucoma	8	[20]
Bipolar ^b	15	[20]	Osteoarthritis	23	[20]
Crohn disease	25–35	[20]	Prostate cancer	2.3–3	[20, 736]
Hemochromatosis ^c	41	[560]	Psoriasis	7	[20]
Hemochromatosis ^d	65	[560]	Rheumatoid arthritis	5–8	[20]
High myopia	4.9	[202]	Schizophrenia	9–10	[20, 561]
Hodgkin's disease	7	[560]	Tuberculoid leprosy	2.4	[560]

^a Autoimmune thyroid disease, ^b Bipolar affective disorder, ^c Recurrence risk ratio for idiopathic hemochromatosis in males, ^d Recurrence risk ratio for idiopathic hemochromatosis in females, ^e Insulin-dependent diabetes mellitus.

The concept of recurrence risk ratios can be extended to quantitative traits if these are dichotomized. Because quantitative traits are dichotomized using specific pre-defined thresholds, they offer the flexibility that different thresholds can be used for the two relatives. For example, one may consider the 95th age- and gender-specific body mass index (BMI; kg/m^2) percentile for one sibling and the 90th percentile for

the other. For quantitative traits, the term recurrence risk ratio may thus be used in a generalized way, and the natural term therefore is *generalized recurrence risk ratio*.

Table 6.3 shows the generalized recurrence risk ratios for obesity and dyslexia. Unfortunately, point estimates for λ_M and λ_O are not available for dyslexia. The λ values increase with the more extreme threshold for defining the disease. Furthermore, $\lambda_M \geq \lambda_S \geq \lambda_O$ for obesity (cp. Section 8.3.1.3).

Table 6.3 Generalized recurrence risk ratio estimates as reported for dichotomized quantitative traits.

Phenotype	λ_S	λ_M	λ_O	Reference	Phenotype	λ_S	Reference
Obesity ^a	1.8	3.5	1.4	[768]	Dyslexia ^c	4.9	[766]
Obesity ^b	4.2	10.3	2.6	[768]	Dyslexia ^d	6.6	[766]
Obesity ^a	2.6	5.7	1.7	[14]			
Obesity ^b	4.2	11.7	2.6	[14]			

^a Using 90th percentile of the distribution of the body mass index (BMI), i.e., the 10% most obese.

^b Using 95th BMI percentile. ^c The extreme 5% of the population for spelling is considered affected.

^d The extreme 2.5% of the population for spelling is considered affected.

The recurrence risk ratio approach can be used to investigate whether genetic or non-genetic factors play a role. For example, a higher concordance in pairs of brothers compared to father–son pairs points to an X-chromosomal or a recessive disease [471]. Recurrence risk ratios can also demonstrate that a disease is complex and cannot be explained by a simple Mendelian single-locus model [561]. However, when these approaches are used, the investigator should be aware that point estimates do not reflect sample variability. The uncertainty inherent to the estimation procedure, i.e., the variability of the estimates, should therefore be taken into account when conclusions are drawn. Furthermore, any possible bias in the point estimates, e.g., through the sampling design, needs to be accounted for. For example, sibling recurrence risk ratios have sometimes been reported even if the ascertainment of families was through an offspring, not through a parent. In detail, researchers claimed to report $P(\text{offspring affected}|\text{parent affected})$, while the ascertainment scheme only allowed estimation of $P(\text{parent affected}|\text{offspring affected})$.

Although this approach gives a basic impression whether familial aggregation is high or low, it does not allow to separate the effect of genes and fridges. In the next section, we therefore discuss the concept of heritability that has been developed for estimating the proportion of variance attributable to genetic effects.

6.3 WHAT IS THE CONCEPT OF HERITABILITY?

To investigate the proportion of variance attributable to genetic effects, we need to introduce an appropriate statistical model. In this statistical model, we need to distinguish so-called additive genetic effects from dominance genetic effects. For this reason, we start with the simplest variance analytic model possible in Section 6.3.1. It

includes a genetic effect that is due to a specific chromosomal locus. In Section 6.3.2, we generalize and modify this simple model in several ways.

6.3.1 What is the variance analytic Falconer model at a trait locus?

The starting point is the variance analytic *Falconer model*, where the phenotype y of a subject is additively decomposed into a general mean μ , a genetic effect g that is due to a specific chromosomal locus, termed *trait locus*, and an error term ε :

$$y = \mu + g + \varepsilon \quad (6.1)$$

We assume that the residual has no systematic effect, i.e., $\mathbb{E}(\varepsilon) = 0$, and the variance of the residual is denoted by $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$. Additionally, we surmise a diallelic trait locus with alleles A_1 and A_2 and allele frequencies p and q for alleles A_1 and A_2 , respectively, random mating, Hardy–Weinberg equilibrium (HWE), and no major gene by residual interaction.

The genetic effect is modeled as follows:

$$g = \begin{cases} a & \text{if the individual has genotype } A_1A_1 \\ d & \text{if the individual has genotype } A_1A_2 \\ -a & \text{if the individual has genotype } A_2A_2 \end{cases} \quad (6.2)$$

The genetic effect is illustrated in Figure 6.3 with the assumption of a normally distributed error term ε and a general mean of $\mu = 0$. Although the value of the

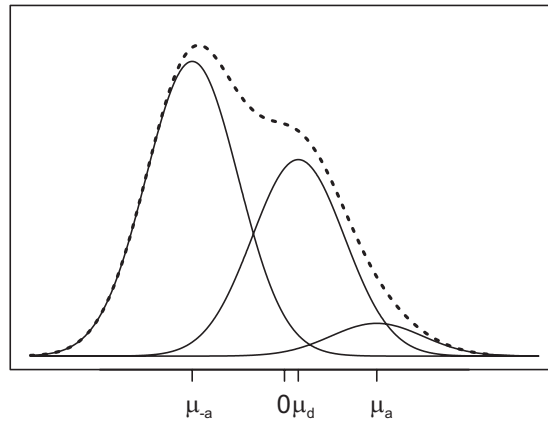


Fig. 6.3 The genetic effect using the Falconer model from Eq. (6.2) with the assumption of a normally distributed error e and $\mu = 0$ in Eq. (6.1). Individuals with genotype A_1A_1 have mean a , while individuals with genotypes A_1A_2 and A_2A_2 have means d and $-a$, respectively. The three displayed distributions are already weighted by their respective probabilities p^2 , $2pq$, and q^2 . The joint distribution in the population generally is not a normal distribution but a mixture of the three normal distributions. Solid lines indicate genotype-specific distributions, while the dotted line shows the mixture of the three single distributions in the population.

general mean is 0, the overall mean can be different from 0 because the genotype-specific means are $\mu_a = \mu + a$, $\mu_d = \mu + d$, and $\mu_{-a} = \mu - a$ for genotypes A_1A_1 , A_1A_2 , and A_2A_2 , respectively. Therefore, the overall mean is the mixture of the genotype-specific means multiplied by their respective occurrence probabilities, i.e., by $\mu + p^2a + 2pqd + q^2(-a)$. The variability in the phenotypic distribution depends on three factors: 1) the *additive genetic effect* a , 2) the *dominance genetic effect* d , and 3) the allele frequency p of the *high allele* A_1 . Specific genetic models can be obtained as follows:

1. $a = d$: dominant genetic model	$\begin{array}{ccc} & & \\ \mu_{-a} & 0 & \mu_a = \mu_d \end{array}$
2. $-a = d$: recessive genetic model	$\begin{array}{ccc} & & \\ \mu_{-a} = \mu_d & 0 & \mu_a \end{array}$
3. $d = 0$: additive genetic model	$\begin{array}{ccc} & & \\ \mu_{-a} & 0 = \mu_d & \mu_a \end{array}$
4. $d > a$: model with overdominance	$\begin{array}{ccc} & & \\ \mu_{-a} & 0 & \mu_a < \mu_d \end{array}$
5. $d < -a$: model with underdominance	$\begin{array}{ccc} & & \\ \mu_d < \mu_{-a} & 0 & \mu_a \end{array}$

Here, the straight line on the right illustrates the means of the underlying genetic models.

If $d = 0$, the variability of the mixture distribution of the phenotype is explained by the allele frequency p and the size of the genetic effect a . The mean and variance of the genetic effect for the whole population are given by $\mu_g = (p - q)a$ and $\sigma_g^2 = 2pqa^2$ in this case (see Problem 6.1). If $d \neq 0$, the variance is also influenced by both the additive genetic effect a and the dominance effect d . It can be shown that the genetic variance caused by the trait locus can be additively decomposed into the *additive genetic variance* σ_a^2 and the *dominance genetic variance* σ_d^2 , which are given by [198]

$$\sigma_a^2 = 2pq(a - d(p - q))^2 \quad \text{and} \quad \sigma_d^2 = (2pqd)^2. \quad (6.3)$$

The sum of the *genetic variance* is $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$, and the error variance σ_ϵ^2 is the *total variance* $\sigma^2 = \sigma_g^2 + \sigma_\epsilon^2$. The proportion of the total variance that can be explained by the trait locus is termed *heritability in the broad sense due to the trait locus* or *broad sense heritability due to the trait locus*

$$h_B^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}.$$

h^2 corresponds to the coefficient of determination R^2 from standard regression models. In some genetic models, the dominance effect is of minor interest, and thus the proportion of genetic variance that can be explained by the additive genetic effect is considered. The resulting heritability is the *heritability in the narrow sense due to the trait locus* or *narrow sense heritability due to the trait locus*

$$h_N^2 = \frac{\sigma_a^2}{\sigma_g^2 + \sigma_\epsilon^2}.$$

If $d = 0$, the model is said to have *no dominance variance*, and $h_B^2 = h_N^2$.

6.3.2 How do more general Falconer models look like?

In this section, we consider various generalizations of the simple Falconer model from the previous section. The first natural extension is to consider not only a single trait locus model but an *oligogenic model* with J trait loci:

$$y = \mu + \sum_{j=1}^J g_j + \varepsilon$$

If all J loci are diallelic with locus-specific additive genetic effects a_j and dominance genetic effects d_j , locus-specific additive and dominance variances can be considered. Several assumptions underly this model, and two important are: 1) The equation indicates that the J loci are independent because no interaction is included. 2) The model is identifiable only if genetic markers represent the trait loci.

In this chapter, we restrict our attention to models that do not make use of genetic markers. As a result, the effect of the different loci cannot be distinguished, and the effects of all genes are summarized in a genetic component G :

$$y = \mu + G + \varepsilon \quad (6.4)$$

G is commonly termed *polygenic component*, and it measures the genetic contribution of all genes in Eq. (6.4). G is usually decomposed in a *polygenic additive effect* A and a polygenic dominance effect D , and the genetic variance σ_G^2 is decomposed in the additive genetic variance σ_A^2 and the dominance genetic variance σ_D^2 . In the absence of a major gene effect, the trait variance thus becomes

$$\text{Var}(y) = \sigma^2 = \sigma_A^2 + \sigma_D^2 + \sigma_\varepsilon^2.$$

Now, we can define the *broad sense heritability* as

$$H_B^2 = \frac{\sigma_G^2}{\sigma^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\varepsilon^2}.$$

It is the proportion of variance attributable to *all* genetic effects and therefore denoted by a big H^2 . This is in contrast to the last Section, where the broad sense heritability *due to the trait locus* was considered, denoted by a little h^2 .

We can also define the *narrow sense heritability* by

$$H_N^2 = \frac{\sigma_A^2}{\sigma^2} = \frac{\sigma_A^2}{\sigma_G^2 + \sigma_\varepsilon^2}.$$

It is the proportion of variance attributable to *all additive* genetic effects.

In Eq. (6.4), all genetic effects have been summarized in a single genetic component G . In segregation analysis, a distinction is made between a single random effect major locus g and all other genetic components G , and the basic model is

$$y = \mu + g + G + \varepsilon.$$

This model is identifiable only if g and G are uncorrelated so that $\text{Cov}(g, G) = 0$. In addition, both g and ε and G and ε need to be pairwise uncorrelated.

The models considered so far do not allow environmental effects. These have been implicitly absorbed in the error term ε . The next extension therefore is to include environmental effects. Here, *shared environment* may be distinguished from *non-shared environment*. In most models, shared environment, i.e., environmental factors that are common to all family members, is considered as random effect E . Non-shared environment, i.e., subject-specific environmental factors, is included by some measured variables x and an appropriate function, say, $x'\beta$. Here, β is the parameter vector to be estimated, and x is the corresponding vector of environmental variables. In many classical papers on family studies, environmental factors were not measured so that the non-shared environment was absorbed in the error term ε .

For the model to be identifiable we assume that the polygenic effect, and the error term are pairwise uncorrelated. Without loss of generality, we furthermore assume $\mathbb{E}(G) = \mathbb{E}(E) = \mathbb{E}(\varepsilon) = 0$. As a consequence, Eq. (6.4) extends to

$$y = \mu + A + D + E + x'\beta + \varepsilon, \quad (6.5)$$

with mean and variance of an individual given by

$$\begin{aligned} \mathbb{E}(y) &= \mu + x'\beta, \quad \text{and} \\ \mathbb{V}\text{ar}(y) &= \sigma^2 = \sigma_A^2 + \sigma_D^2 + \sigma_E^2 + \sigma_\varepsilon^2. \end{aligned} \quad (6.6)$$

Equations (6.5) and (6.6) point to two questionable assumptions. The variance decomposition of Eq. (6.6) is only valid if a) there is no correlation between genes and environment and b) there are no non-linear functional relationships between G and E . If G and E are correlated, the variance of $\mathbb{V}\text{ar}(y)$ becomes

$$\mathbb{V}\text{ar}(y) = \sigma^2 = \sigma_G^2 + \sigma_E^2 + 2\sigma_{G,E} + \sigma_\varepsilon^2,$$

making the model non-identifiable because of the gene environment covariance $\sigma_{G,E}$. An unpleasant property of observational studies is, however, that genes and environment cannot be forced to be independent through experimental conditions. To make the model identifiable, $\sigma_{G,E} = 0$ has to be assumed, which leads to an inflated estimate of σ_G^2 if $\sigma_{G,E} \neq 0$. In this case, the estimate of H_B^2 is inflated, too.

6.3.3 What is the covariance between pairs of relatives? What are the kinship coefficient and Jacquard's Δ_7 coefficient?

In most applications, entire families are available for analysis, not only a single subject or a single relative pair. This requires adjustment for the non-independence of subjects in the statistical analyses. Specifically, for a nuclear family consisting in both parents and two offspring, one full sib-pair, two parent-offspring pairs, and one pair of spouses can be formed. Although statistical models for correlated data are more complicated than those for independent observations, the non-independence has one advantage: It allows to model the covariance or the correlation between

subjects. Specifically, the covariance between subjects y_t and $y_{t'}$ from a family is given by

$$\mathbb{C}\text{ov}(y_t, y_{t'}) = 2c_{t,t'}\sigma_A^2 + \Delta_{7;t,t'}\sigma_D^2 + \gamma_{t,t'}\sigma_E^2, \quad (6.7)$$

where $c_{t,t'}$ is the *kinship coefficient* between subjects t and t' , and it measures the relationship between two individuals. More formally, for two individuals t and t' , it is the probability that an allele selected randomly from individual t and an allele selected randomly from the same autosomal locus of subject t' are identical by descent. An allele is said to be identical by descent (IBD) for a pair of relatives if it has been inherited from the same common ancestor; for details, see Chapter 8. A few examples for values of c are shown in Fig. 6.4 (also see Table 6.4), and c equals 0 for unrelated individuals.

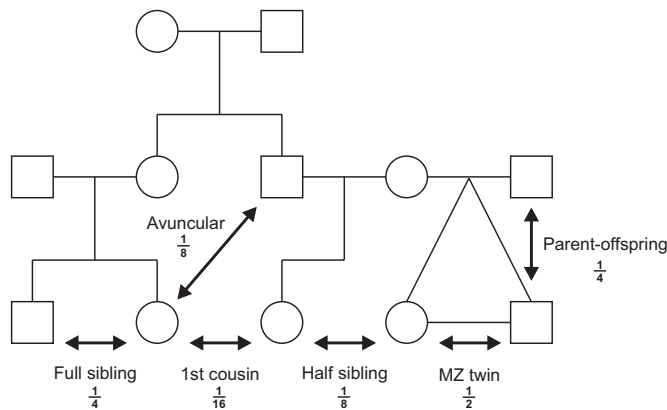


Fig. 6.4 Kinship coefficients for common familial relationships.

$\Delta_{7;t,t'}$ is *Jacquard's Δ_7 coefficient* reporting the probability of a pair sharing two alleles identical by descent (IBD). Finally, $\gamma_{t,t'}$ is an indicator variable for shared environment, i.e., $\gamma_{t,t'} = 1$ if subjects live together or were reared together, and 0 otherwise (see Table 6.4).

The variance–covariance model of Eqs. (6.6) and (6.7) can be estimated in a structural equation model, i.e., a LISREL-type approach; see, e.g., Refs. [92, 493]. The general structural equation model approach can also be used for constructing path models, especially when a vector of different traits or repeated measurements are available; for details, see Refs. [353, 493]. For estimation of parameters and their standard errors, most investigators used the assumption of multivariate normally distributed traits in families and estimated the model by maximum likelihood. However, if the data are skewed and/or kurtotic, this can lead to a substantial bias as explained nicely in Ref. [734]. An alternative is to allow deviations from the multivariate normal by using pseudo maximum likelihood estimation together with the robust variance estimator as introduced in Ref. [33] for mean and covariance structure models. Another alternative to the structural equation model are mixed models or generalized estimating equations (GEE) for the mean and association structure

Table 6.4 Kinship coefficient c , Jacquard's Δ_7 coefficient, the indicator variable γ for shared environment, and the covariance for pairs of unilineal relatives.

Relationship	c	Δ_7	$\gamma_{t,t'}$	$\text{Cov}(y_t, y_{t'})$
Monozygotic twin (reared together)	$\frac{1}{2}$	1	1	$\sigma_A^2 + \sigma_D^2 + \sigma_E^2$
Monozygotic twin (reared apart)	$\frac{1}{2}$	1	0	$\sigma_A^2 + \sigma_D^2$
First degree				
Dizygotic twin (reared together)	$\frac{1}{4}$	$\frac{1}{4}$	1	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_E^2$
Dizygotic twin (reared apart)	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2$
Full sibling (reared together)	$\frac{1}{4}$	$\frac{1}{4}$	1	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_E^2$
Full sibling (reared apart)	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2$
Parent–offspring (living together)	$\frac{1}{4}$	0	1	$\frac{1}{2}\sigma_A^2 + \sigma_E^2$
Second degree				
Half sibling (reared together)	$\frac{1}{8}$	0	1	$\frac{1}{4}\sigma_A^2 + \sigma_E^2$
Half sibling (reared apart)	$\frac{1}{8}$	0	0	$\frac{1}{4}\sigma_A^2$
Avuncular (living apart)	$\frac{1}{8}$	0	0	$\frac{1}{4}\sigma_A^2$
Grandparent–grandchild (living apart)	$\frac{1}{8}$	0	0	$\frac{1}{4}\sigma_A^2$
Third degree				
First cousin (living apart)	$\frac{1}{16}$	0	0	$\frac{1}{8}\sigma_A^2$
Great-grandparent–offspring (living apart)	$\frac{1}{16}$	0	0	$\frac{1}{8}\sigma_A^2$
Unrelated				
Spouse (living together)	0	0	1	σ_E^2

The kinship coefficient equals the probability of sharing an allele identical by descent. Jacquard's Δ_7 coefficient reports the probability of a pair sharing two alleles identical by descent. γ is an indicator whether shared environment is present or absent. The covariance $\text{Cov}(y_t, y_{t'}) = 2c_{t,t'}\sigma_A^2 + \Delta_{7;t,t'}\sigma_D^2 + \gamma_{t,t'}\sigma_E^2$ for a pair of relatives t and t' with phenotypes y and y' reflects the additive genetic variance and the dominance genetic variance. The variance component for shared environment σ_E^2 is present in this covariance only if subjects live together or were reared together.

which follow the same ideas as the mean and covariance structure models in this setting [32]. For diseases, the same models can be used if appropriately applied to non-metric dependent variables; see, e.g., Ref. [92].

A challenging aspect is the ascertainment correction that is necessary in most family studies. Specifically, the methods outlined above are applicable to randomly sampled families. In practice, families are often included in a study through one or more probands who are affected or have an extreme value for the quantitative trait of interest. To this end, the likelihood function conditioned on the proband's trait value can be used when the assumption of a correctly specified multivariate distribution is utilized. For dichotomous phenotypes, solutions also exist; see, e.g., Refs. [413, 666, 718, 747].

An alternative to the variance–covariance structure approach sketched in this section is a regression approach. The trait value of an offspring can be regressed on the parental phenotypes, and the so-called empirical heritability can be estimated from the resulting regression coefficient; see Problem 6.2 and Refs. [198, 353]. This approach will be briefly discussed in the context of adoption studies below.

6.4 WHAT ARE TWIN STUDIES? WHAT ARE ADOPTION STUDIES?

The collection of extended pedigrees and the phenotyping of family members for estimating familial resemblance is time consuming and expensive. Therefore, twin studies and adoption studies have received great attention as alternative study designs. They are much simpler to conduct, especially when representative twin registries or complete adoption registries are available. Like family studies, both twin and adoption studies aim to estimate heritability. Environmental factors can be investigated by studying discordant twin pairs.

6.4.1 What are twin studies?

The following fundamental idea forms the basis of twin studies. Monozygotic (MZ) twins share all genes and dizygotic (DZ) twins share about half their genes. Environmental factors that have an effect on the trait or disease of interest are shared by both pairs of twins to the same degree. Differences in the concordance between monozygotic twins and dizygotic twins should therefore be attributable to genetic factors. Ideally, the genetic effect can be quantified. For example, if a disease is caused by a gene following an autosomal dominant mode of inheritance with complete penetrance and no phenocopies, the concordance rate between monozygotic twins should be 100%, while it should be 50% in dizygotic twins.

This can be formalized as follows: The covariance between MZ twins reared together is $\sigma_A^2 + \sigma_D^2 + \sigma_E^2$, and it is $\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_E^2$ for DZ twins. If we assume the phenotypic variance to be homoscedastic σ^2 for all subjects as in Eq. (6.6), the phenotypic correlation between MZ twins ϱ_{MZT} , where the subscript T denotes reared together, and DZ twins reared together ϱ_{DZT} becomes

$$\varrho_{MZT} = \frac{\sigma_A^2 + \sigma_D^2 + \sigma_E^2}{\sigma^2} \quad \text{and} \quad \varrho_{DZT} = \frac{\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_E^2}{\sigma^2}, \quad (6.8)$$

respectively. The difference $\varrho_{MZT} - \varrho_{DZT}$ yields

$$\varrho_{MZT} - \varrho_{DZT} = \frac{\frac{1}{2}\sigma_A^2 + \frac{3}{4}\sigma_D^2}{\sigma^2}.$$

Therefore, $2(\varrho_{MZT} - \varrho_{DZT}) = H_B^2 + \frac{1}{2}\frac{\sigma_D^2}{\sigma^2} \geq H_B^2$ is often used for estimating the broad sense heritability although the estimator is slightly overestimated if $\sigma_D^2 \neq 0$. If $\sigma_D^2 \approx 0$, $2(\varrho_{MZT} - \varrho_{DZT}) = H_B^2 = H_N^2$. Many different alternative estimators have been proposed; for details, see, e.g., Refs. [198, 267].

For estimating the proportion of variance explained by shared environment,

$$2\varrho_{DZT} - \varrho_{MZT} = \frac{\sigma_E^2 - \frac{1}{2}\sigma_D^2}{\sigma^2} \leq \frac{\sigma_E^2}{\sigma^2}$$

is often used. Analogous to the heritability estimator, $2\hat{\varrho}_{DZT} - \hat{\varrho}_{MZT}$ is unbiased only if there is no dominance genetic variance, i.e., if $\sigma_D^2 = 0$.

There are three fundamental assumptions underlying this estimation approach:

- Genetic and environmental factors need to be uncorrelated. This has already been considered in the previous section.
- Mating between individuals is at random.
- Environmental factors need to be equal in MZ and DZ twins, termed the *equal environment assumption in twins (EEA)*.

To avoid shared environment in twins reared together, researchers have proposed to investigate twins reared apart; see, e.g., Ref. [80]. With this study design, H_B^2 is estimated by $2(\hat{\varrho}_{MZA} - \hat{\varrho}_{DZA})$, where the subscript *A* denotes reared apart. If both twins reared apart and reared together are available, the shared environment can also be estimated by contrasting DZT and DZA twins or MZT and MZA twins. An argument against this idea is that twins reared apart typically grow up in families with the same cultural and social background so that they live in similar environments [344, 353]. Some of the twins reared apart do have close contacts with each other, too.

6.4.2 What are adoption studies?

Several researchers have argued that the EEA is generally not fulfilled. Specifically, the EEA assumption means that both prenatal and postnatal environments have to be identical. However, prenatal differences in environment between MZ and DZ twins are caused by their early embryological development that affects subsequent fetal growth and survival. In addition to these fundamental concerns, there are other methodological problems related to twin studies, including lack of representativity [344]. Because of the many crucial assumptions and methodological problems, some authors state that heritability estimates obtained from twin studies do not serve any scientific purpose. Although adjustments for some of these issues have been proposed (see, e.g., Ref. [176]), the major concerns against the use of twin studies remain. These also apply to twin studies for diseases instead of quantitative traits. The only difference from considering quantitative traits is that concordance rates are compared instead of correlation coefficients.

An alternative to twin studies are adoption studies. The fundamental idea underlying adoption studies is identical to that of twin studies with twins reared apart. If an offspring is adopted immediately after birth, familial resemblance between a parent and an offspring is caused only by genetic factors. At the same time, an adopted child and the adoptive parent have no genes in common so that their similarity can be explained only by the shared environment. Therefore, the coefficient from a simple

linear regression of the offspring's phenotype on the parental phenotype reveals the heritability in this case (see Problem 6.2).

Similar to twin studies, results from adoption studies are valid only under several important assumptions which are rarely met in practice:

- Parent and offspring have no shared environment, which is questionable at least for mother–offspring pairs because of the prenatal time.
- The adoptive families are representative.
- The adopted child is representative.
- There is no age effect.
- There is no gender effect.
- There is no selective placement.

Although twin and adoption studies underlie these crucial assumptions that are rarely met in applications, several important findings have been published by using these study results; for an overview, see Ref. [75].

6.5 WHAT ARE CRITICAL ASPECTS WHEN INVESTIGATING FAMILIAL RESEMBLANCE?

In Section 6.3, we have already discussed the two critical assumptions inherent to any observational study that aims to estimate heritability. The variance decomposition of Eq. (6.6) is valid only if a) there is no correlation between genes and environment and b) there are no non-linear functional relationships between G and E . Otherwise, the model is not identifiable.

The critical assumption of equal environments underlying twin studies has also been discussed. An advantage of twin studies is, however, that differences in age do not play a role. Furthermore, gender effects can be overcome by using twin pairs of identical gender. This is different from family studies where age- and gender effects can play an important role. As a result, heritability estimates from family studies may be quite different from those from twin studies if no appropriate adjustments are performed for the variables. This is illustrated for the BMI.

Twin studies have produced consistent and high heritability estimates for the BMI in the range between 0.6 and 0.9 [79]. These high estimates apply to twins reared both together and apart. Except for the newborn period for which a lower heritability of 0.4 has been estimated, age does not affect heritability estimates to a substantial degree. However, heritability estimates from adoption studies and family studies are considerably lower, and the lowest estimate being in the magnitude of 0.1 for BMI. Interestingly, one large family study has estimated the heritability to be 0.67, which is in the same range as those derived from twin studies. It is probably the adjustment for age effects that makes this difference; for details see, e.g., Ref. [289].

Irrespective of the study design, *secular trends* may invalidate the results for the trait of interest. This problem has been illustrated nicely using the quantitative trait height, see Ref. [679], for which high heritabilities have been reported.

In the past 150 years, average height in males has increased in many countries (Tables 6.5 and 6.6), while it remained fairly constant from the Stone Age to the middle of the 19th century. As explained in detail in Ref. [679], the reduction of inbreeding cannot explain the observed increase in total. Instead, better nutrition during childhood and reduction of infectious diseases, especially of the gastrointestinal tract, most likely cause this increase in height. Table 6.6 specifically shows substantial differences in the increase by occupational group. If this confounding aspect between genetic factors and occupational group is not adequately taken into account, a bias in heritability estimates may occur. Because these secular trends may also be active in current time, the analyses of different relative pairs who vary substantially in age, such as grandparent–grandchild pairs and sib-pairs, can be heavily affected. Therefore, the problem of age effects again comes into play.

Table 6.5 Average height (in cm) of adult males.

	Sweden	Norway	Denmark
Stone Age	169.5	164	170.0
Bronze Age	166.5	—	166.5
Iron Age	167.0	167.0	168.0
Middle Age	167.5	168.0	165.5
1855	167.5	168.0	165.5
1939	174.5	174.5	171.5
2008	181.5	179.7	180.6

Data taken from Ref. [679] and Wikipedia.

6.6 HOW CAN EVIDENCE FOR A MAJOR GENE EFFECT BE ESTABLISHED? HOW CAN A SEGREGATION PATTERN FOLLOWING MENDELIAN INHERITANCE BE DETERMINED?

When twin studies, adoption studies, and/or family studies using extended pedigrees have provided support for the hypothesis that the disease or trait of interest has a genetic basis, segregation analysis often is the next step. With segregation analysis, the researcher tries to identify the mode of inheritance of the disease or of the quantitative trait. Using maximum likelihood estimation the most plausible segregation pattern is estimated from pedigrees, preferably from multigenerational family data. This approach was quite successful for monogenic diseases. These commonly follow a simple Mendelian mode of inheritance, and a single major gene can be expected to have a strong genetic effect. Furthermore, only two important allelic states, i.e., functional and non-functional, can be expected.

Again, we stress that no genetic marker information is utilized in segregation analysis. Therefore, the existence of a latent, i.e., unobserved variable with the three states A_1A_1 , A_1A_2 , and A_2A_2 , is assumed. This latent variable is usually termed an individual's *type* [247]. Genotypes are a special case of types that transmit to offspring in Mendelian fashion. Thus, the term *type* is used as the more general term and allows many kinds of discrete transmission, including non-Mendelian. Genotype probabilities are therefore extended to type probabilities, but they are still denoted

Table 6.6 Height (in centimeters) of Swiss conscripts.

Kanton Luzern	1897–1902	1927–1932	Increase
Merchants and students	166.6	171.2	4.4
Factory laborers	161.8	167.0	5.2
Farmers	163.1	166.1	3.0
Kanton Schwyz	1887	1935	
Intellectuals	167.0	170.6	3.6
Heavy physical labor laborers	164.0	169.4	4.4
Light physical labor laborers	163.2	168.0	4.8
Farmers	162.9	168.7	5.8
Factory laborers	155.9	169.6	13.7
City of Zürich	1910	1930	
Merchants and students	169.6	172.7	3.1
Tailors	166.5	169.5	3.0
Factory workers	166.4	170.5	4.1
Farmers	165.8	168.4	2.4
Blacksmiths	165.7	168.8	3.1

Data taken from Ref. [404]; see Ref. [679].

by $p_{A_1 A_1}$, $p_{A_1 A_2}$, and $p_{A_2 A_2}$. Under Hardy–Weinberg equilibrium (Chapter 2), $p_{A_1 A_1} = p_{A_1}^2 = p^2$, where $p_{A_1} = p$ is the frequency of the *component*, also named *component allele* A .

The elements of the genetic model typically estimated in segregation analysis for diseases, i.e., dichotomous traits are

- the component allele frequency, more specifically, the disease allele frequency,
- the transmission probabilities, i.e., the probability that a parental genotype transmits a particular component to an offspring, and
- the penetrances, i.e., the affection probabilities given the type.

For quantitative traits, the estimated elements are

- the component allele frequency of the high allele,
- the transmission probabilities,
- the means given the type, and
- the standard deviation given the type.

The model includes the *transition probabilities*, $P(g_O | g_F, g_M)$, i.e., the probability for an offspring type g_O given paternal type g_F and maternal type g_M . The

transition probabilities are determined by the *transmission probabilities*. Specifically, τ_{g_F} denotes the probability that allele A_1 is transmitted to an offspring from the father given he has type g_F . τ_{g_M} is defined analogously. The transition probabilities can be expressed by the transmission probabilities:

$$\begin{aligned} P(g_O = A_1 A_1 | g_F, g_M) &= \tau_{g_F} \tau_{g_M}, \\ P(g_O = A_1 A_2 | g_F, g_M) &= \tau_{g_F} (1 - \tau_{g_M}) + (1 - \tau_{g_F}) \tau_{g_M}, \\ P(g_O = A_2 A_2 | g_F, g_M) &= (1 - \tau_{g_F}) (1 - \tau_{g_M}). \end{aligned}$$

Several transmission models are considered to be standard, and these are given in Table 6.7.

Table 6.7 Standard transmission models for segregation analysis.

Transmission model	Parameters
Homogeneous, no transmission (HNT)	$\tau_{A_1 A_1} = \tau_{A_1 A_2} = \tau_{A_2 A_2} = p_{A_1}$
Homogeneous, Mendelian (HM)	$\tau_{A_1 A_1} = 1, \tau_{A_1 A_2} = 0.5, \tau_{A_2 A_2} = 0$
General homogeneous (HG)	$0 \leq \tau_{A_1 A_1}, \tau_{A_2 A_2} \leq 1,$ $\tau_{A_1 A_2} = \frac{(p - p^2 \tau_{A_1 A_1} - (1 - p)^2 \tau_{A_2 A_2})}{2p(1 - p)}$
Free $\tau_{A_1 A_2}$ (FT)	$\tau_{A_1 A_1} = 1, 0 \leq \tau_{A_1 A_2} \leq 1, \tau_{A_2 A_2} = 0$
General (G)	$0 \leq \tau_{A_1 A_1}, \tau_{A_1 A_2}, \tau_{A_2 A_2} \leq 1$

So far, we have defined type frequencies, transition, and transmission probabilities. Now, we need to establish the function between type and phenotype. To this end, let y_i denote the phenotype of subject i in a family. For a disease or any dichotomous trait, the penetrance function is denoted by $\pi_{g_i} = P(y_i | g_i)$ (see Chapter 2). For a quantitative trait, the functional relation between the type and the phenotype is established by the density of the normal distribution. There are two important differences when comparing dichotomous and quantitative traits for segregation analysis. First, three type groups, hence three penetrances have to be estimated under a general penetrance model. In contrast, three means and three variances need to be estimated from the sample when considering a quantitative trait. To reduce the complexity of the model, type independent variances $\sigma_{g_i}^2 = \sigma^2$ are usually assumed.

Second, residual correlation often plays a role for quantitative traits. To this end, reconsider the simple Falconer model of Eq. (6.5)

$$y = \mu + g + \varepsilon = \mu + a + d + \varepsilon,$$

where a and d represent the latent type effects, and ε the residual. For specific relative pairs, e.g., sib-pairs, this residual correlation is assumed to be identical over generations and between different parts of the family. We have, however, illustrated in the previous section that secular trends might be present which, in turn, can affect these correlations. One further aspect related to the residual family correlation needs to be discussed. In the second half of the 1980s, the *regressive models* of Bonney

[73] became popular because they allowed type-specific covariate effects in the segregation analysis. Two residual correlation models have been used in applications quite often. These were the *class A model* and the *class D model*. In the class A model, the residual sib-pair correlation is restricted. Specifically, let ϱ_{SS} denote the sib-pair residual correlation, ϱ_{PO} the residual parent-offspring correlation, and ϱ_{FM} the spouse correlation, then the restriction is $\varrho_{SS} = 2\varrho_{PO}^2/(1 + \varrho_{FM})$. When there is no residual spouse correlation, the residual sib-pair correlation is twice the residual parent-offspring correlation. In some applications, a substantially higher ϱ_{SS} was observed compared to ϱ_{PO} , making the class A model assumption too restrictive. As a result, the more general class D model that does not impose this restriction was preferred in applications. However, the implementation of the class D model was computationally daunting [76], and it only allowed the inclusion of type-specific covariates in the class A model, not in the class D model, for some time. Today, efficient implementations of both models are available several packages, including the user-friendly package S.A.G.E.

This comment on the implementation already points to important questions for segregation analysis. First, how can the likelihood be formulated and, second, what algorithm is appropriate for estimating model parameters? It is shown in the Appendix that the joint distribution of all dichotomous phenotypes in a randomly sampled family can be decomposed into the penetrances, the latent type probabilities of the founders of the pedigree, and the latent transition probabilities in non-founders. Specifically, for f founders and $n - f$ non-founders, one obtains

$$P(y_1, \dots, y_n) = \sum_{g_1, \dots, g_n} \prod_{i=1}^n P(y_i | g_i) \prod_{i=1}^f P(g_i) \prod_{i=f+1}^n P(g_i | g_{i,F}, g_{i,M}). \quad (6.9)$$

The parameters in this joint distribution are the transmission probabilities, the penetrances, and the type frequencies. In Eq. (6.9), the penetrances $P(y_i | g_i)$ depend only on the type, not on the phenotype of other subjects such as parents. Furthermore, the type g_i of an offspring is completely determined by the parental types. Thus, no information from other relatives is required. Equation (6.9) is written as a probability but can also be interpreted as a likelihood function L , where the underlying parameters have to be estimated.

For quantitative traits, the joint likelihood is formulated in a similar way. Here, a multivariate normal distribution is commonly assumed for the error term. Some colleagues have proposed to Box-Cox [82] transform the phenotypes for establishing a better fit of the phenotype to the normal distribution for quantitative traits. Because the phenotypes given the types are assumed to be normal, it is questionable whether the phenotypes should be transformed prior to model fitting.

The Elston-Stewart algorithm [187] is the commonly employed algorithm for estimating the parameters and is described in detail in the Appendix, Section A.1. Different implementations exist, and commonly used programs for segregation analysis are/were PAP, POINTER [390], and S.A.G.E. The parameterizations used in these programs differ, and a conversion between the different parameterizations is given in Table 6.8. Using maximum likelihood estimation, the parameters of Eq.

Table 6.8 Relationship between parameters for segregation analysis used by the programs PAP and POINTER and the standard notation used in Section 6.3.1.

PAP	POINTER	Conversion from POINTER to PAP	Conversion from PAP to standard notation
q frequency of A_1	p frequency of A_2	$q = 1 - p$	not required
μ_1 mean of A_1A_1	μ overall mean	$\mu_1 = \mu - 2pqdt - p^2t$	$\mu = \mu_1 + t/2$
μ_2 mean of A_1A_2	V overall variance	$V = \sigma_g^2 + \sigma_{ML}^2$	$a = t/2$
μ_3 mean of A_2A_2	t displacement between heterozygotes	$\mu_3 = t - \mu_1$	$d = t \cdot (d_{\text{POINTER}} - 0.5)$
$\mu_1 \leq \mu_2 \leq \mu_3$			
σ_g^2 within genotype variance	d relative position of the heterozygous mean	$\mu_2 = \mu_1 + dt$	$\sigma_E^2 = V - \sigma_{ML}^2$
	$d_{\text{POINTER}} = \frac{\mu_2 - \mu_1}{\mu_3 - \mu_1}$		
H proportion of σ_g^2 attributed to additive polygenes $= \sigma_a^2 / \sigma_g^2$	h^2 narrow sense heritability $= \sigma_a^2 / V$	$H = h^2 \frac{V}{\sigma_g^2}$	$\varrho_{FO} = \varrho_{MO} = H/2$
$\varrho_{E-spouse}$ environmental spouse correlation			$\varrho_{FM} = \varrho_{E-spouse}$
ϱ_{E-sib} environmental sibling correlation			$\varrho_{SS} = \varrho_{E-sib} + H/2$
$\sigma_{ML}^2 = q^2(\mu_1 - \mu)^2 + 2pq(\mu_2 - \mu)^2 + p^2(\mu_3 - \mu)^2$ variance due to major locus			

ϱ_{FO} residual father-offspring correlation; ϱ_{MO} residual mother offspring correlation. Table adapted from Ref. [764]. The PAP to POINTER conversions are similar to those given in Ref. [353], p. 273. The relevant S.A.G.E. parameters are identical to PAP parameters.

Table 6.9 Distribution of test statistics used in segregation analysis for transmission models.

	HNT	HM	HG	FT
HM	—			
HG	χ^2_2	$\frac{1}{4} + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$		
FT	—	$\frac{1}{2} + \frac{1}{2}\chi^2_1$	—	
G	χ^2_3	$\frac{1}{4}\chi^2_1 + \frac{1}{2}\chi^2_2 + \frac{1}{4}\chi^2_3$	χ^2_1	χ^2_2

HNT: homogeneous, no transmission; HM: homogeneous, Mendelian; HG: general homogeneous; FT: free $\tau_{A_1 A_2}$; G: general transmission model. Adapted from the S.A.G.E. User Manual.

(6.9) can be estimated, and the value of the log likelihood can be determined. An alternative approach is to use GEE [402, 638, 745, 746].

Some of the models are nested, and statistical tests can be carried out to investigate whether there are significant differences between a more general and a restricted model. It is important to note that some of the tests are not standard so that the results of Self and Liang [603] need to be applied. In the non-standard case, the limiting distribution is a mixture of χ^2 distributions, and the distributions of some test statistics for transmission models are displayed in Table 6.9. Similarly, Table 6.10 shows reasonable models for type effects in segregation analysis. Not all of these models are nested so that statistical testing using asymptotic distributions is not always possible. Furthermore, asymptotic distributions for order restrictions are hard to derive. This specifically applies to the decreasing and increasing penetrance models. The parameters in Table 6.10 are formulated in terms of penetrances. An extension to segregation analysis using the logistic regression function for penetrances and a parameter definition in this model or the formulation of mean parameters for quantitative traits is straightforward.

Table 6.10 Commonly used type effects in segregation analysis.

Model	Parameter
One parameter, fixed	$\pi = \pi_{A_1 A_1} = \pi_{A_1 A_2} = \pi_{A_2 A_2}$
Two parameters, fixed	$\pi_1 = \pi_{A_1 A_1} = \pi_{A_1 A_2}, \pi_2 = \pi_{A_2 A_2}$
	$\pi_1 = \pi_{A_1 A_1}, \pi_2 = \pi_{A_1 A_2} = \pi_{A_2 A_2}$
Three parameters	$\pi_{A_1 A_1}, \pi_{A_1 A_2}, \pi_{A_2 A_2}$
Dominant	$\pi_{A_1 A_1} = \pi_{A_1 A_2}, \pi_{A_2 A_2}$
Recessive	$\pi_{A_1 A_1}, \pi_{A_1 A_2} = \pi_{A_2 A_2}$
Additive	$\pi_{A_1 A_1}, \pi_{A_1 A_2} = \frac{1}{2}(\pi_{A_1 A_1} + \pi_{A_2 A_2}), \pi_{A_2 A_2}$
Decreasing	$\pi_{A_1 A_1} \geq \pi_{A_1 A_2} \geq \pi_{A_2 A_2}$
Increasing	$\pi_{A_1 A_1} \leq \pi_{A_1 A_2} \leq \pi_{A_2 A_2}$

A type probability without an index or a number as index indicates a fixed value. This value needs to be specified by the user.

For quantitative traits, one additional idea is often used that has been discussed before (Section 6.3). Given various relative pairs, the residual correlation, i.e., the correlation not attributable to the latent types, can be estimated. This correlation is then attributed to polygenic effects. Excess residual correlation in specific relative pairs, e.g., sib-pairs or spouses, can be estimated as well.

Because not all models can be evaluated using standard statistical theory, the *Akaike information criterion (AIC)* (see Section 11.5.5) is often used for model comparison. In brief, it is given by $AIC = -2 \log L + 2r$, where L denotes the likelihood function and r the number of parameters of the model, and the user chooses the model with the lowest AIC.

Although Tables 6.9 and 6.10 indicate a wide variety of possible models to be estimated in segregation analysis, only a small set of different models is commonly employed. This is best illustrated using a published real data example.

Example 6.2. *To illustrate the use of the different models, we consider the study of Moll et al. [469]. They investigated the role of genetic and environmental factors for the quantitative trait BMI in 1302 relatives identified through 284 schoolchildren from Muscatine (IA), in the United States of America. Prior to segregation analysis, they adjusted BMI levels for variability in age, gender, and relative type. They found that a mixture of two normal distributions fit the adjusted BMI data better than did a single normal distribution.*

In the next step, they conducted a segregation analysis. The estimated models are displayed in Table 6.11. The lowest log likelihood was obtained for model 3, the single-locus model, and the second lowest for the recessive model. These models are nested, and the value of the χ^2 statistic with 1 degree of freedom is $2(-6583.21 + 6584.51) = 2.60$ that is not significant at the 5% test level because $p = 0.107$. If the AIC had been used for model selection, the single-locus model would have been favored because it reveals the smallest AIC among all models displayed in Table 6.11. In detail, the log likelihood difference between models 3 and 5 is 2.60. The AIC of model 5 thus is 0.60 higher than the AIC of model 3.

In the final step of their analyses, Moll et al. analyzed further restrictions (Table 6.12). Transmission probabilities were assumed to conform with Mendelian inheritance. Models that did not include the effects of a single recessive locus (models 7 and 12) or that did not include the effects of polygenic loci (models 6, 9, and 12) revealed substantially worse fit. Models not including shared environment (models 6, 8, and 10) were also rejected.

Moll et al. concluded that the data show evidence for a recessive single-locus model and an additional polygenic component. Furthermore, the distribution of adjusted BMI is influenced to a small extent both by environmental factors shared by siblings living in the same house and by environmental factors shared by spouses living together.

Segregation analysis approach seems appealing to determine whether a segregation pattern conforms with classical Mendelian inheritance, i.e., autosomal, sex-linked, recessive, additive or dominant inheritance, or non-classical inheritance, i.e., mitochondrial diseases, genomic imprinting, parent of origin effect, anticipation, or,

Table 6.11 Results of segregation analysis of Ref. [469], Table 4.

Parameter	Maximum likelihood estimate $\pm$ standard error				
	1	2	3	4	5
	General transmission	Non-transmitted factor	Single locus	Dominant single-locus	Recessive single-locus
p	0.737 ± 0.03	0.770 ± 0.02	0.754 ± 0.02	0.979 ± 0.01	0.754 ± 0.02
μ_1	23.37 ± 0.32	23.97 ± 0.27	23.37 ± 0.30	24.01 ± 0.18	23.76 ± 0.17
μ_2	24.26 ± 0.40	23.98 ± 0.40	24.25 ± 0.37	37.21 ± 1.2	23.76 ± 0.17
μ_3	34.57 ± 0.64	34.93 ± 0.71	35.04 ± 0.67	37.21 ± 1.2	34.80 ± 0.63
σ	3.55 ± 0.11	3.68 ± 0.11	3.57 ± 0.11	3.92 ± 0.11	3.62 ± 0.10
h^2	0.644 ± 0.10	0.654 ± 0.09	0.615 ± 0.07	0.513 ± 0.06	0.645 ± 0.09
Spouse E	0.217 ± 0.10	0.235 ± 0.09	0.189 ± 0.10	0.227 ± 0.08	0.185 ± 0.10
Sib E	0.162 ± 0.06	0.171 ± 0.05	0.161 ± 0.06	0.176 ± 0.05	0.157 ± 0.06
τ_1	[1.0]	0.770 ± 0.02	(1.0)	(1.0)	(1.0)
τ_2	0.441 ± 0.05	0.770 ± 0.02	(0.5)	(0.5)	(0.5)
τ_3	0.514 ± 0.18	0.770 ± 0.02	(0)	(0)	(0)
$\log L$	-6581.00	-6596.71	-6583.21	-6617.91	-6584.51
χ^2					
model 1		31.42	4.42		
d.f.		3	3		
p -value		$6.9 \cdot 10^{-7}$	0.220		
model 3				69.40	2.60
d.f.				1	1
p -value				$8.0 \cdot 10^{-15}$	0.107

Displayed are maximum likelihood parameter estimates plus/minus standard errors ($\pm SE$), and the χ^2 comparison with model 1 and 5 together with the degrees of freedom (d.f.). Parentheses denote that value is fixed in model; brackets denote that value is estimated at the boundary of the parameter space.

finally, non-Mendelian inheritance, i.e., no pattern. And some investigators are convinced that segregation analysis is a prerequisite for linkage analyses. We agree that segregation analysis is helpful when diseases with a single major locus following Mendelian inheritance are considered. However, there are several assumptions inherent to segregation analysis that make them not suitable for complex diseases or quantitative traits studied nowadays. Specifically, the following assumptions are usually made for quantitative traits (see Ref. [764]):

Table 6.12 Results of segregation analysis of Ref. [469], Table 5, for testing hypotheses about polygenic and shared environmental effects (E) between spouses and siblings.

Parameter	Maximum likelihood estimate $\pm$ standard error										
	5	6	7	8	9	10	11	12			
	Recessive single-locus + polygenes + E	Recessive single-locus	Polygenes + E	Recessive single-locus + polygenes	Recessive single-locus + polygenes	Recessive single-locus + polygenes + spouse E	Recessive single-locus + polygenes + sib E	No single locus, no polygenes, no E			
p	0.754 $\pm$ 0.02	0.706 $\pm$ 0.02	(1.0)	0.755 $\pm$ 0.02	0.726 $\pm$ 0.02	0.755 $\pm$ 0.02	0.753 $\pm$ 0.02	(1.0)			
μ_1	23.76 $\pm$ 0.17	23.28 $\pm$ 0.17	24.41 $\pm$ 1.9	23.44 $\pm$ 0.13	23.71 $\pm$ 0.16	23.73 $\pm$ 0.17	23.74 $\pm$ 0.16	24.23 $\pm$ 0.12			
μ_2	23.76 $\pm$ 0.17	23.28 $\pm$ 0.17	24.41 $\pm$ 1.9	23.44 $\pm$ 0.13	23.71 $\pm$ 0.16	23.73 $\pm$ 0.17	23.74 $\pm$ 0.16	24.23 $\pm$ 0.12			
μ_3	34.80 $\pm$ 0.63	33.64 $\pm$ 0.61	24.41 $\pm$ 1.9	34.89 $\pm$ 0.62	34.07 $\pm$ 0.64	34.85 $\pm$ 0.64	34.83 $\pm$ 0.61	24.23 $\pm$ 0.12			
σ	3.62 $\pm$ 0.10	3.37 $\pm$ 0.10	4.56 $\pm$ 0.11	3.56 $\pm$ 0.09	3.48 $\pm$ 0.09	3.59 $\pm$ 0.10	3.59 $\pm$ 0.10	4.50 $\pm$ 0.09			
h^2	0.645 $\pm$ 0.09	(0)	0.582 $\pm$ 0.06	0.598 $\pm$ 0.05	(0)	0.659 $\pm$ 0.06	0.594 $\pm$ 0.06	(0)			
Spouse E	0.185 $\pm$ 0.10	(0)	0.211 $\pm$ 0.07	(0)	0.160 $\pm$ 0.11	0.202 $\pm$ 0.09	(0)	(0)			
Sib E	0.157 $\pm$ 0.06	(0)	0.216 $\pm$ 0.05	(0)	0.443 $\pm$ 0.06	(0)	0.172 $\pm$ 0.06	(0)			
$\log L$	-6584.51	-6649.30	-6688.53	-6590.56	-6632.53	-6588.03	-6586.45	-6773.05			
χ^2		129.58	208.04	12.10	96.04	7.04	3.88	377.08			
d.f.		3	2	2	1	1	1	5			
p -value		$6.6 \cdot 10^{-28}$	$5.6 \cdot 10^{-46}$	0.002	$1.1 \cdot 10^{-22}$	0.008	0.049	$2.6 \cdot 10^{-79}$			

Displayed are maximum likelihood parameter estimates plus/minus standard errors ($\pm SE$), the χ^2 comparison with model 5 together with the degrees of freedom (d.f.) and the p -value. Parentheses denote that value is fixed in model; brackets denote that value is estimated at the boundary of the parameter space.

- Hardy–Weinberg equilibrium
- No epistasis
- No pleiotropic effects
- No genetic heterogeneity
- No interaction between the major gene and the residual effect
- Identical variances for different types
- The trait given the type is normally distributed.

The last two assumptions do not need to be made. However, corrections for deviations from the first 5 assumptions are not straightforward. Another complicating factor is that the likelihood can be fairly flat because of the limited number of pedigrees and pedigree members available. Checking stability of parameter estimates is therefore important. If the likelihood is flat, a thorough comparison of the different models is difficult. Even more, detection of genetic factors deviating from classical Mendelian inheritance patterns is hard.

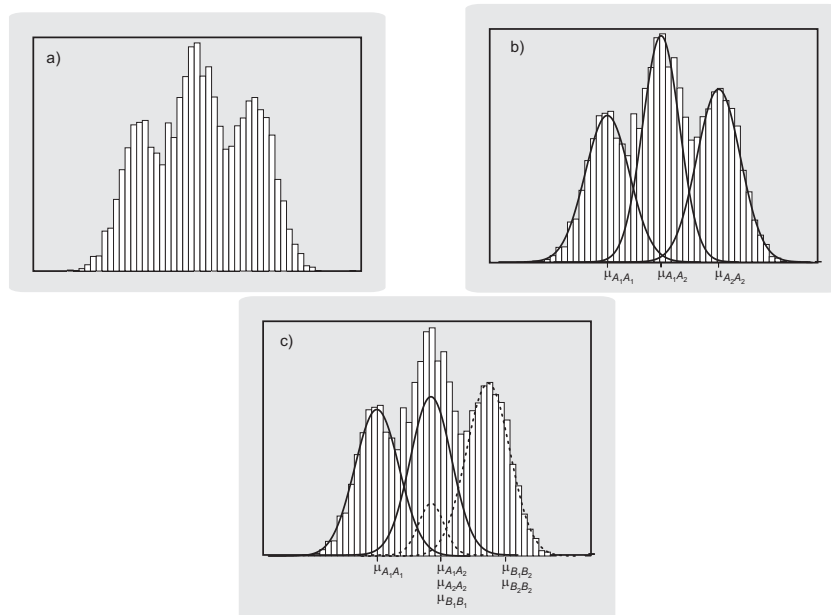


Fig. 6.5 Limits of segregation analysis. a) shows a trimodal distribution which can be fit by a single-locus model (b). In reality, the data have been generated using a two-locus model (c).

To illustrate the natural limit of segregation analysis, consider the standard pedigrees in Figure 6.5 using a simple model. Figure 6.5a shows the histogram of the traits in a population. One can easily recognize the tri-modality of the distribution. Therefore, a natural assumption is the existence of a single autosomal locus with three different means for types A_1A_1 , A_1A_2 , and A_2A_2 (Figure 6.5b). However,

the data have been simulated using a two-locus model with autosomal loci A and B, both of them acting in a dominant fashion for the high components A_2 and B_2 (Figure 6.5c). The genetic model from this artificial example cannot be usually detected in segregation analysis. Only if a large number of pedigrees having a structure that is rarely ascertained, if no secular trends are present, and if this model is considered in the segregation analysis, it might be identified.

For complex genetic disorders, results from segregation analysis are expected to be quite heterogeneous. For example, in segregation analysis for dyslexia no common genetic model was identified. For the BMI all studies revealed a major effect. This was, however, not always Mendelian. Even those studies that identified a major gene showed substantial variation in their results. These heterogeneous findings are typical for segregation analysis of complex traits. We therefore conclude that segregation analysis for complex traits is of very limited value for further studies, and results should be interpreted with great care.

6.7 PROBLEMS

Problem 6.1 Consider an additive genetic model based on Eqs. (6.1) and (6.2), i.e., $d = 0$. Show that the population mean and the variance that is due to the genetic locus are $(p - q)a$ and $2pqa^2$.

Problem 6.2 Consider a quantitative trait and a linear relationship between an offspring and the parents. Regress the offspring's trait value y_O on a single parental phenotype y_P using a linear regression. Show that the regression coefficient can be used for estimating familial resemblance by establishing a functional relation between the coefficient of determination and the regression coefficient. Next, consider the trait value of both parents y_F and y_M using the *midparent value* $(y_M + y_F)/2$. Show that the regression coefficient can be used for estimating familial resemblance.

URLs

JPAP

Pedigree analysis package for Java

<http://hasstedt.genetics.utah.edu/jpap/>

PAP

Pedigree analysis package

<http://hasstedt.genetics.utah.edu/pap/>

S.A.G.E.

<http://darwin.cwru.edu/sage/>

WIKIPEDIA ENTRY ON HUMAN HEIGHT

http://en.wikipedia.org/wiki/Human_height/

7

Model-Based Linkage Analysis

The recombination fraction θ has already been defined in previous chapters. In this chapter, however, we show how the recombination fraction can be estimated. For its definition, θ requires two loci that are called genetically linked—or linked for short—if θ is less than 0.5. In Section 7.1, we start by investigating linkage between two genetic markers using the maximum likelihood method. Section 7.1.1 deals with the easiest case of phase-known pedigrees, where recombinants and non-recombinants can be counted. In Sections 7.1.2 and 7.1.3, this approach is generalized to phase-unknown pedigrees and missing genotypes, respectively.

In many applications the aim is to establish a relationship between a disease and a chromosomal region, i.e., to *map* a disease to a genetic locus. Here, the genetic locus is represented by one or several genetic markers. In the second part of this chapter, we therefore describe the classical approach for linkage mapping. Here, linkage between a disease and a genetic marker is established by using an explicit genetic model of the trait and the marker locus in terms of penetrances and allele frequencies. In Section 7.2.1, we start off by considering phase-known pedigrees. Section 7.2.2 subsequently deals with the general case. The gain in informativity by typing additional individuals in a family is briefly discussed in Section 7.2.3. Standard likelihood ratio tests will be used throughout this chapter. However, the choice of appropriate significance levels for the tests will be deferred until the last section.

7.1 HOW CAN THE RECOMBINATION FRACTION BE ESTIMATED BETWEEN TWO GENETIC MARKERS?

The term *linkage* is of Old English and German origin and is defined as “the existence or establishment of connections between two things” [179]. This connection specifically is between two loci on the same chromosome that are close enough that their alleles cosegregate. Though the term linkage is of German origin, the German word for linkage is “Kopplung.” This word traces back to the Latin term “copula,” which is still used in English. In contrast, *association* is solely of Latin origin. Its roots are in the medieval Latin verb “associare,” which means “to connect,” and in the old Latin combination of words “ad socius.” “Ad” can be best translated as “close to” and “socius” as “together.” In genetics, both terms deal with connections between loci on the same chromosome. Furthermore, in both approaches the researcher tries to detect deviations from the Mendelian law of independent assortment (see Chapter 2), but in different ways! Linkage measures deviations from the Mendelian law of independent assortment, i.e., from independent segregation. The alleles causing this independent assortment may be different for different families. Therefore, linkage runs within families.

In contrast, association can be defined as the non-independence between an allele at a genetic marker and a disease (see Chapter 11). In genetic epidemiology, investigators search for associations that are caused by non-independent segregation. In the best case, association can be observed in families where the specific allele is transmitted more frequently than expected across the whole sample. Here, association is due to linkage. However, there are other reasons that can lead to the non-independence between a specific allele and a genetic marker. These will be studied in detail in Chapter 11.

Because the concepts linkage and association have a joint core, the following classical mnemonic usually helps in remembering their differences:

Linkage is with Loci—Association is with Alleles.

7.1.1 How can the recombination fraction be estimated in phase-known pedigrees?

In this section, the aim is to introduce the fundamental ideas of how the recombination fraction can be estimated. Basically, we will show that linkage analysis can be carried out by counting recombinants and non-recombinants. Let us start with the simplest case, where all individuals are genotyped and the parental phases can be reconstructed from the genotypes in grandparents. Consider Figure 7.1, displaying a three-generation pedigree with two genetic markers, labeled *A* and *B*. Both genotyped grandparents are homozygous at the marker loci. While the grandmother is homozygous with the 1 allele, the grandfather is homozygous with the 2 allele at both loci. According to Mendel’s law of uniformity, the father is heterozygous with genotypes 1 2 at both marker loci. Genotyping of the mother has revealed homozygosity with 1 1 at both loci.

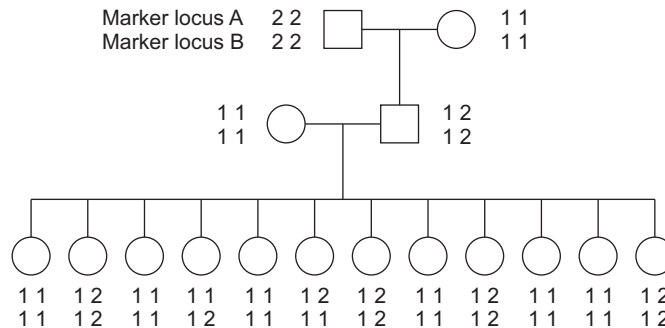


Fig. 7.1 Three-generation pedigree to illustrate linkage analysis between two genetic markers in phase-known pedigrees. All individuals are genotyped at the two marker loci *A* and *B*.

The first important observation is that the *phase*, i.e., the two haplotypes, can be determined in the father. Specifically, both 1 alleles of the father have been transmitted from the grandmother, while both 2 alleles have been inherited from the grandfather. If both markers are located on a single chromosome, the two 1 alleles form a haplotype, as do the two 2 alleles. Haplotypes are usually separated by boxes or vertical lines, as shown in Figure 7.2.

So far, we know that the father has haplotypes $\frac{1}{1}$ and $\frac{2}{2}$. Because the mother is homozygous at both marker loci, we cannot determine whether recombinations have occurred maternally in children. The mother thus is not informative for linkage. According to Mendel's law of uniformity, the mother has transmitted a 1 allele at both marker loci to her children. The gray-shaded alleles and haplotypes displayed in Figure 7.2 are therefore not directly used for estimating the recombination fraction.

The father is, however, heterozygous at both loci. Furthermore, we were able to determine his haplotypes using the grandparental genotypes. We can thus easily see paternal recombinations in the offspring. The alleles that have been inherited from the father to the 12 offspring are displayed in Figure 7.2. Recombinations are observed in offspring 4 and 6, with paternal haplotypes $\frac{1}{2}$ and $\frac{2}{1}$, respectively. In total, there are $n = 12$ informative meioses from the father to the offspring. $k = 2$ recombinations occurred out of these $n = 12$ meioses, and thus the recombination fraction is estimated as $\hat{\theta} = \frac{2}{12} \approx 0.1667$.

Obviously, we used the maximum likelihood (ML) approach for estimating the recombination fraction θ . We assumed the number of observable crossing overs to be binomial distributed with parameter θ . For estimation, we simply counted the number of recombinations among the number of informative meiosis in this pedigree. The approach can also be used to estimate θ across families.

The properties of $\hat{\theta}$ are easily deduced because θ is estimated by the ML method. Because the parameter space of recombination fractions is, however, restricted to $0 \leq \theta \leq \frac{1}{2}$, one should be aware that the ML estimator is biased, though it can be shown to be asymptotically unbiased.

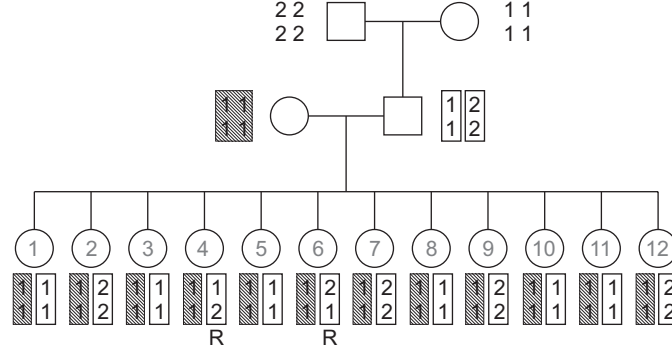


Fig. 7.2 Three-generation pedigree to illustrate linkage analysis between two genetic markers in phase-known pedigrees. All individuals are genotyped at the two marker loci A and B . The mother is not informative for linkage. Thus, the maternal genotypes and her transmissions to the third generation are marked gray. The father's phase could be determined; his haplotypes are given in boxes. In the numbered offspring, two recombinations can be observed, marked by R .

The next step is to establish a test for linkage. Two markers are linked if

- a haplotype is transmitted more often than expected by chance,
- the two alleles of a haplotype are not recombined with a 50% probability, or
- the recombination fraction $\theta < 0.5$.

The hypotheses for this one-sided test problem are therefore given by

$$H_0: \theta = \frac{1}{2} \quad \text{versus} \quad H_1: \theta < \frac{1}{2}.$$

θ is estimated by the ML method. We therefore utilize a maximum likelihood ratio test (MLRT) to test for linkage. As stated above, the number of recombinants can be assumed to follow a binomial distribution with parameter θ . The part of the likelihood that is relevant for maximization, informally termed *kernel of the likelihood*, therefore is

$$L(\theta) = \theta^k (1 - \theta)^{n-k},$$

and we maximize the likelihood ratio

$$\Lambda(\theta) = \frac{L(\theta)}{L(\theta)|_{\theta=0.5}}. \quad (7.1)$$

$\Lambda(\theta)$ describes the odds for linkage at θ . The decadic logarithm of $\Lambda(\theta)$ can therefore be interpreted as the logarithm of the odds score for linkage, termed *LOD score at θ* , for short. For historical reasons, the *LOD score function*

$$LOD(\theta) = \log \Lambda(\theta)$$

is preferred in genetic epidemiological applications over

$$LR(\theta) = 2 \ln \Lambda(\theta), \quad (7.2)$$

which is statistically more appealing: $LRT = \max_{\theta: \theta \in [0; 0.5]} LR(\theta)$, i.e., the maximum of Eq. (7.2) is asymptotically distributed as a 50:50 mixture of a χ_1^2 and a point mass at 0. Note that the mixture distribution is obtained because the alternative hypothesis $H_1: \theta < 0.5$ is one-sided.

In practice, p -values can be determined easily by using the normal distribution, as illustrated below in Example 7.1. One takes the positive square root of the test statistic, i.e., $+\sqrt{LRT}$. The p -value then is the corresponding upper quantile of the standard normal distribution: $p\text{-value} = 1 - \Phi(+\sqrt{LRT})$, where Φ denotes the distribution function of the standard normal distribution.

The maximum of the LOD score, i.e., $\max_{\theta: \theta \in [0; 0.5]} LOD(\theta)$, is sometimes called *the LOD score* in applications. This term is somewhat imprecise because several maxima can be estimated if linkage is considered between one marker and several others, possibly on different chromosomes. There is a simple relation between the LOD score and the LRT:

$$LOD = \max(\log \Lambda(\theta)) = \max \frac{\ln \Lambda(\theta)}{\ln 10} = \frac{\max(2 \ln \Lambda(\theta))}{2 \ln 10} = \frac{LRT}{2 \ln 10}$$

Thus, one only needs to divide the LRT by $2 \ln(10) \approx 4.605$ to obtain an LOD score. Analogously, the LOD score needs to be multiplied by $2 \ln(10)$ for obtaining the value of the test statistic. In summary,

$$LOD = \frac{LRT}{2 \ln(10)} \approx \frac{LRT}{4.605} \quad LRT = 2 \ln(10) \cdot LOD \approx 4.605 \cdot LOD. \quad (7.3)$$

Example 7.1. The pedigree displayed in Figure 7.1 has $n = 12$ informative meioses and $k = 2$ recombinations. The kernel of the likelihood is therefore given by $L(\theta) = \theta^2(1 - \theta)^{10}$. The kernel of the likelihood under no linkage, i.e., $\theta = 0.5$, therefore is $L(0.5) = 0.5^{12}$, and the likelihood ratio thus is $\Lambda(\theta) = 2^{12}\theta^2(1 - \theta)^{10}$. The maximum likelihood estimator was $\hat{\theta} = \frac{2}{12} \approx 0.1667$. To test for linkage, we first compute the maximum likelihood ratio test statistic $T = \max_{\theta} LR(\theta) = LR(\hat{\theta})$, which is given by

$$T = \max_{\theta \in [0; 0.5]} LR(\theta) = LR(\theta)|_{\theta=\hat{\theta}} = 2 \ln \left(2^{12} \left(\frac{2}{12} \right)^2 \left(1 - \frac{2}{12} \right)^{10} \right) \approx 5.8221.$$

The p -value can now be obtained by using the normal distribution: $\Phi(\sqrt{T}) \approx \Phi(2.4129) \approx 0.9921$, resulting in $p = 1 - 0.9921 = 0.0079$.

As an alternative, one could employ any “test for a single proportion” in this situation. For example, the exact binomial test for the hypothesis $H_0: \theta = 0.5$ versus $H_1: \theta < 0.5$ at $n = 12$ and $k = 2$ gives a p -value of 0.0161.

The maximum LOD score for the pedigree is

$$LOD = \max_{\theta} LOD(\theta) = LOD(\theta)|_{\hat{\theta}} = \log \left(2^{12} \left(\frac{2}{12} \right)^2 \left(1 - \frac{2}{12} \right)^{10} \right) \approx 1.2642.$$

One important property of the LOD score is its additivity across families, because LODs are logs of ratios of two likelihoods, and the contributions from different families are independent:

$$\begin{aligned} LOD(\theta) &= \log \frac{L_1(\theta)}{L_1(\theta)|_{\theta=\frac{1}{2}}} \cdot \frac{L_2(\theta)}{L_2(\theta)|_{\theta=\frac{1}{2}}} \cdots \frac{L_N(\theta)}{L_N(\theta)|_{\theta=\frac{1}{2}}} \\ &= LOD_1(\theta) + LOD_2(\theta) + \cdots + LOD_N(\theta) \end{aligned}$$

where $L_1, L_2, \dots, L_N$ denote the kernels of the likelihoods of families 1, 2, $\dots$, N .

We conclude this section by giving a slightly more theoretical formulation of the likelihood in phase-known nuclear pedigrees. It depends on two components: (1) the genotypes g in the parents (par) and (2) the genotypes in the offspring (off), which they have inherited from their parents. This transmission depends on the true recombination fraction θ , indicated by $||\theta$ in the formula below. Therefore, the kernel of the likelihood is given by

$$L(\theta) = P(g_{\text{par}}) \cdot P(g_{\text{off}}|g_{\text{par}}||\theta).$$

One question remains after these calculations: Is the p -value obtained from testing for linkage in this example pedigree so small that we can reject the null hypothesis of no linkage? This will be discussed in more detail in Section 7.2.

7.1.2 How can the recombination fraction be estimated in phase-unknown pedigrees?

Often there is ambiguity about the parental phase. In this section, we consider the same pedigree as before, but we assume that the grandparents were unavailable for genotyping (Figure 7.3). Consequently, the paternal haplotypes cannot be uniquely determined. They can be either $\frac{1}{1}$ and $\frac{2}{2}$ (case 1) or $\frac{1}{2}$ and $\frac{2}{1}$ (case 2). The phase thus is unknown and, according to Laplace probabilities, both paternal haplotype combinations are equally likely. The kernel of the likelihood for the first haplotype combination is $L_1(\theta) = \theta^2(1 - \theta)^{10}$ because two recombinations occur. The second haplotype combination has 10 recombinations, yielding a likelihood kernel of $L_2(\theta) = \theta^{10}(1 - \theta)^2$. Each kernel is equally likely, and therefore the joint kernel of the likelihood is the 1:1 combination of the two kernels, yielding

$$L(\theta) = 0.5 L_1(\theta) + 0.5 L_2(\theta) = 0.5 \theta^2(1 - \theta)^{10} + 0.5 \theta^{10}(1 - \theta)^2. \quad (7.4)$$

The corresponding LOD score function is

$$LOD(\theta) = \log \frac{0.5 \theta^2(1 - \theta)^{10} + 0.5 \theta^{10}(1 - \theta)^2}{0.5^{12}} = \log \frac{L_1(\theta) + L_2(\theta)}{0.5^{11}}.$$

It can be easily shown that Eq. (7.4) takes maxima at $\hat{\theta} = 0.1667$ and at $\hat{\theta} = 0.8333$. The latter exceeds the theoretical maximum of 0.5, and therefore $\hat{\theta}$ is restricted to 0.5 in this case, resulting in $LOD = 0$. The former maximum $\hat{\theta} = 0.1667$ gives an LOD score of 0.9632.

Because the phase is unknown in the father (F), a more general formulation of the kernel of the likelihood for nuclear families is

$$\begin{aligned} L(\theta) &= P(g_{\text{par}}) \cdot P(g_{\text{off}}|g_{\text{par}}||\theta) \\ &= \sum_{\text{phases}_F} P(\text{phase}_F) \cdot P(g_{\text{off}}|\text{phase}_F||\theta). \end{aligned}$$

7.1.3 How can the recombination fraction be estimated in pedigrees where some individuals are not genotyped?

In many applications, genotypes of some individuals are unknown. While missing genotypes in the last generation of a pedigree usually can be ignored, parental genotypes may not be neglected. Again, we consider the pedigree from Section 7.1.1. However, we now assume that the mother has not been genotyped (Figure 7.4). Obviously, the phase information depends on the haplotypes of the mother. We therefore first determine the possible haplotype combinations in the mother. Second, we derive the kernels of the likelihood for each possible haplotype combination. Third, we calculate the probability of each haplotype combination, and, finally we combine all three parts with the joint likelihood.

7.1.3.1 What are the possible haplotype combinations in the mother?

Because there are offspring homozygous for the 1 allele at both marker loci, the mother must have at least one 1 allele at both marker loci. If we had at least one offspring homozygous for the 2 allele at both marker loci, we could have reconstructed the maternal genotypes, leading to two possible haplotype combinations. However,

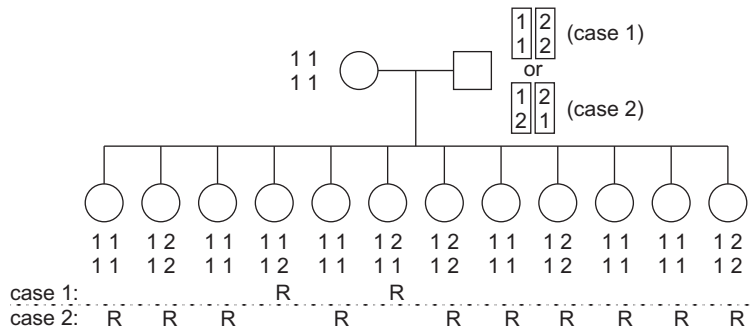


Fig. 7.3 Two-generation pedigree to illustrate linkage analysis between two genetic markers in phase-unknown pedigrees. All individuals are genotyped at the two marker loci *A* and *B*. In contrast to Figure 7.1, grandparental genotypes are unavailable. Two phases are possible in the father, denoted by cases 1 and 2. The segregation pattern subsequently reveals either 2 or 10 recombinations.

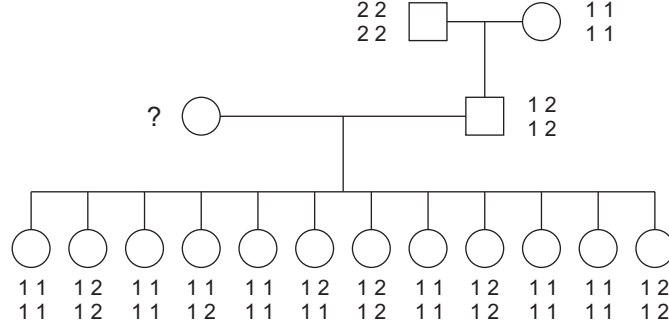


Fig. 7.4 Three-generation pedigree to illustrate linkage analysis between two genetic markers in pedigrees with unknown, i.e., missing, genotypes. All individuals except the mother are genotyped at both marker loci *A* and *B*. The phase in the father can be determined from grandparental genotypes.

in this pedigree we have not observed such an offspring. Therefore, the mother can have one of the following five haplotype combinations according to Mendelian laws:

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad H_3 = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad H_5 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

7.1.3.2 What is the kernel of the likelihood for each possible maternal haplotype combination?

If the mother is homozygous at one or both markers, one cannot determine whether a maternal recombination occurs. As a result, the mother is not informative for linkage for haplotype combinations H_1 , H_2 , and H_3 . Therefore, the situation is reduced to the setting considered before. The mother is heterozygous at both loci for haplotype combinations H_4 and H_5 , and thus informative for linkage. For H_1 , there are $n = 12$ informative paternal meioses, and the number of recombinations is $k = 2$, as determined in Section 7.1.1. For H_2 and H_3 , there are only 7 informative paternal meioses. Specifically, for both H_2 and H_3 , offspring being homozygous 1 at both marker loci do not exhibit a recombination, while those being heterozygous at both marker loci are not informative for linkage. The offspring that is heterozygous at the first locus and homozygous at the second is informative for H_2 and shows a recombination, but it is not informative for H_3 . In contrast, the offspring that is homozygous at the first but heterozygous at the second locus is recombinant and informative for H_3 but not informative for H_2 .

It is slightly more complicated to count the number of recombinants and non-recombinants for H_4 and H_5 . We therefore display the relevant parts of the corresponding pedigrees in Figure 7.4. Six offspring are homozygous for the 1 allele at both loci. Their haplotypes therefore are $\frac{1}{1}$. For H_4 , there are 12 informative meioses for these 6 offspring showing a total of 12 non-recombinants, while they have a total of 6 recombinants and 6 non-recombinants for H_5 .

Four offspring are heterozygous at both loci. Because for H_4 and H_5 both parents are also heterozygous at both marker loci, the phase cannot be determined

in the offspring. Both phases are equally likely if we assume linkage equilibrium between loci. For H_4 each of these offspring has either two non-recombinants or two recombinants, and exactly one recombinant and one non-recombinant for H_5 .

Finally, the two remaining offspring who are heterozygous 1 2 at exactly one locus both have one recombinant and one non-recombinant for H_4 . For H_5 they are either recombinant or non-recombinant for both meioses.

The kernels of the likelihoods corresponding to H_1 and H_5 thus are

$$\begin{aligned}
 L_1(\theta) &= \theta^2(1-\theta)^{10}, \\
 L_2(\theta) &= \underbrace{(1-\theta)^6}_{6 \times \text{hom at both}} \cdot \underbrace{\theta}_{1 \times \text{het at first}}, \\
 L_3(\theta) &= \underbrace{(1-\theta)^6}_{6 \times \text{hom at both}} \cdot \underbrace{\theta}_{1 \times \text{het at second}}, \\
 L_4(\theta) &= \underbrace{\left((1-\theta)^2\right)^6}_{6 \times \text{hom at both}} \cdot \underbrace{\left(\theta^2 + (1-\theta)^2\right)^4}_{4 \times \text{het at both}} \cdot \underbrace{\left(\theta(1-\theta)\right)^2}_{2 \times \text{het at one}}, \\
 L_5(\theta) &= \underbrace{\left(\theta(1-\theta)\right)^6}_{6 \times \text{hom at both}} \cdot \underbrace{\left(\theta(1-\theta)\right)^4}_{4 \times \text{het at both}} \cdot \underbrace{\left(\theta^2 + (1-\theta)^2\right)^2}_{2 \times \text{het at one}}.
 \end{aligned} \tag{7.5}$$

7.1.3.3 What is the probability for each possible maternal haplotype combination? To derive the kernel of the joint likelihood, the probabilities of the five possible haplotype combinations need to be determined and combined. This can be achieved by making two assumptions. First, if the markers are in linkage equilibrium (see Section 5.3), we can use the genotype frequencies of the single loci for estimating the frequencies of both haplotypes. Second, if we assume Hardy–Weinberg equilibrium at both marker loci, the frequencies of the haplotype combinations can be estimated from the marker allele frequencies. In any case, the linkage result will depend on the assumed allele frequencies, which will be illustrated below. In applications, allele frequencies are not known with certainty but have to be estimated from the data. In any case, it is more appropriate to estimate the allele frequencies from the data rather than assuming equally frequent alleles.

In the following, we denote the allele frequencies of the 1 alleles at the marker loci by p_1 and p_2 . Basically, 10 different haplotype combinations are possible for the mother if the offspring genotypes are disregarded. This means that compatibility with Mendelian inheritance is ignored in this step. The haplotype combinations and their probabilities are

$$\begin{aligned}
P(H_1) &= P\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \middle| \begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right) = p_1^2 p_2^2, \\
P(H_2) &= P\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \middle| \begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right) = 2p_1^2 p_2 (1 - p_2), \\
P(H_3) &= P\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \middle| \begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right) = 2p_1 (1 - p_1) p_2^2, \\
P(H_4) &= P\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \middle| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right) = 2p_1 (1 - p_1) p_2 (1 - p_2), \\
P(H_5) &= P\left(\begin{smallmatrix} 1 \\ 2 \end{smallmatrix} \middle| \begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right) = 2p_1 (1 - p_1) p_2 (1 - p_2), \\
P\left(\begin{smallmatrix} 1 \\ 2 \end{smallmatrix} \middle| \begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right) &= p_1^2 (1 - p_2)^2, \quad P\left(\begin{smallmatrix} 1 \\ 2 \end{smallmatrix} \middle| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right) = 2p_1 (1 - p_1) (1 - p_2)^2, \\
P\left(\begin{smallmatrix} 2 \\ 1 \end{smallmatrix} \middle| \begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right) &= (1 - p_1)^2 p_2^2, \quad P\left(\begin{smallmatrix} 2 \\ 1 \end{smallmatrix} \middle| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right) = 2(1 - p_1)^2 p_2 (1 - p_2), \\
P\left(\begin{smallmatrix} 2 \\ 2 \end{smallmatrix} \middle| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right) &= (1 - p_1)^2 (1 - p_2)^2.
\end{aligned}$$

According to Mendel's law of segregation, only the first 5 of these 10 are possible haplotype combinations for the mother if the offspring genotypes are taken into account. The conditional probability that the mother has haplotype combination H_i , given it is one of the possible haplotype combinations, is

$$\begin{aligned}
P(H_1) &= \frac{p_1^2 p_2^2}{c}, & P(H_2) &= \frac{2p_1^2 p_2 (1 - p_2)}{c}, \\
P(H_3) &= \frac{2p_1 (1 - p_1) p_2^2}{c}, & P(H_4) &= P(H_5) = \frac{2p_1 (1 - p_1) p_2 (1 - p_2)}{c},
\end{aligned} \tag{7.6}$$

where $c = p_1 p_2 (p_1 - 2)(p_2 - 2)$ is the sum of the five numerators of Eq. (7.6). Note that we do not explicitly include the condition "compatibility with Mendelian inheritance in the mother" in Eq. (7.6).

Finally, we want to stress that we do not weight haplotype frequencies in Eq. (7.6) with respect to the observed number of alleles in the offspring. Alternatively, one could use the Bayes formula for updating maternal haplotype probabilities given the observed number of 1 alleles at the two loci in the offspring. In fact, this latter approach is followed in some software packages.

7.1.3.4 What is the kernel of the joint likelihood function? Finally, using Eqs. (7.5) and (7.6), the kernel of the joint likelihood function is obtained as

$$\begin{aligned}
L(\theta) &= \sum_{i=1}^5 P(H_i) \cdot L_i(\theta) \\
&= \frac{p_1^2 p_2^2}{c} \theta^2 (1 - \theta)^{10} + \dots \\
&\quad + \frac{2p_1 (1 - p_1) p_2 (1 - p_2)}{c} \theta^{10} (1 - \theta)^{10} \left(\theta^2 + (1 - \theta)^2 \right)^2.
\end{aligned}$$

Example 7.2. To illustrate the calculations, we assume $p_1 = 0.1$ and $p_2 = 0.5$. With the result from Eq. (7.6), the conditional probabilities for the haplotype combinations of the mother can be derived easily. They are given by

$$P(H_1) \approx 0.0175, P(H_2) \approx 0.0351, P(H_3) = P(H_4) = P(H_5) \approx 0.3158.$$

The kernel of the joint likelihood is approximately

$$\begin{aligned} L(\theta) &\approx 0.0175 L_1(\theta) + 0.0351 L_2(\theta) + 0.3158 L_3(\theta) \\ &\quad + 0.3158 L_4(\theta) + 0.3158 L_5(\theta) \\ &= 0.0175 \cdot \theta^2(1-\theta)^{10} \\ &\quad + 0.3509 \cdot \theta(1-\theta)^6 \\ &\quad + 0.3158 \cdot \theta^2(1-\theta)^{14} \left(\theta^2 + (1-\theta)^2 \right)^4 \\ &\quad + 0.3158 \cdot \theta^{10}(1-\theta)^{10} \left(\theta^2 + (1-\theta)^2 \right)^2. \end{aligned}$$

The kernel of this likelihood function should not be maximized analytically. Instead, numerical maximization is preferable. It is plotted in Figure 7.5, and one can observe that the maximum is approximately at $\theta = 0.14$.

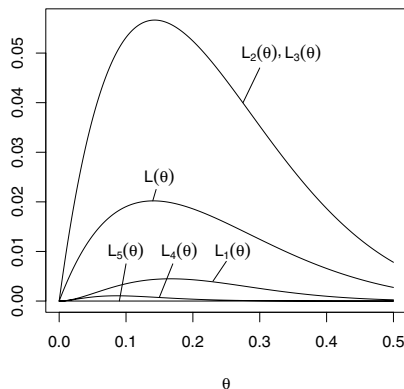


Fig. 7.5 Kernel of the likelihood function from Example 7.2. Five haplotype combinations are possible in the mother (see Section 7.1.3.1), leading to kernels L_1 to L_5 . Likelihood kernel L_1 has a weight of 0.0175. L_2 and L_3 are identical and, together, have a weight of 0.3509. L_4 and L_5 both have a weight of 0.3158. The joint kernel of the likelihood function is the weighted average of L_1 to L_5 .

In applications, even the numerical maximization of the kernel of the likelihood

Table 7.1 LOD score table for Example 7.2.

θ	0.0	0.01	0.05	0.1	0.2	0.3	0.4
LOD score	$-\infty$	0.08	0.68	0.84	0.83	0.66	0.38

function can be impractical, as it can get extremely complicated in multi-generation pedigrees. In these cases, *LOD score tables* are computed, i.e., the likelihood function is evaluated at specific values of θ . A standard choice is $\theta = 0, 0.01, 0.05, 0.1, 0.2, 0.3$, and $\theta = 0.4$ (see Table 7.1 for the data of Example 7.2).

An important issue is that the kernel of the likelihood function, and thus the LOD scores, depends on marker allele frequencies if founder genotypes are missing. This is illustrated in the next example. *Founders* are individuals without ancestors in a pedigree.

Example 7.3. *The purpose of this example is to illustrate the dependency of the LOD score function on marker allele frequencies. Starting points are the formulae that were used in Example 7.2. We therefore consider two constellations for allele frequencies: $p_1 = p_2 = 0.1$ and $p_1 = p_2 = 0.9$. For the first constellation, i.e. $p_1 = p_2 = 0.1$, haplotype frequencies in the mother are approximately 0.0028 for H_1 , 0.0499 for H_2 and H_3 , and 0.4488 for H_4 and H_5 . For the second constellation, haplotype frequencies in the mother are completely different. For $p_1 = p_2 = 0.9$, H_1 has a probability of 0.6694. H_2 and H_3 occur with a probability of 0.1488, and both H_4 and H_5 have a probability of 0.0165.*

The haplotype-specific kernels of the likelihood function and the joint kernels are displayed in Figure 7.6. It can be seen that the constellations result in entirely different kernels. While the kernel corresponding to H_5 is identical to the null, H_4 has a high weight on the joint kernel for the situation $p_1 = p_2 = 0.1$. Subsequently, the kernel of the likelihood is very similar to L_4 in this case. For $p_1 = p_2 = 0.9$, the kernel of the likelihood is dominated by L_1 to L_3 , which together have a probability of approximately 97%.

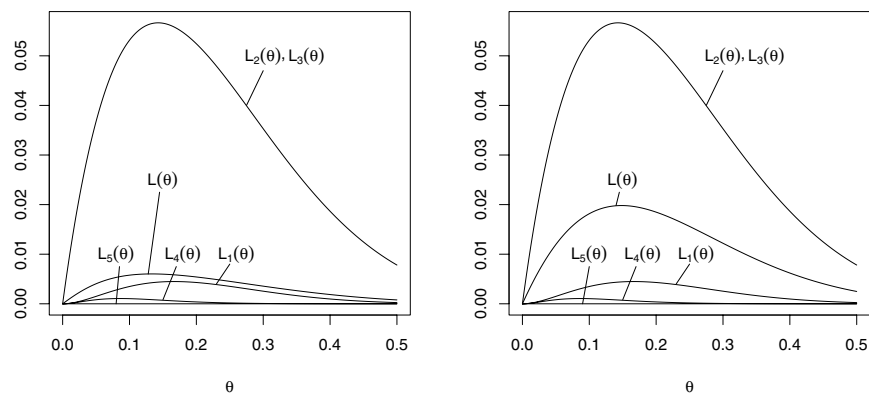


Fig. 7.6 Kernel of the likelihood function from Example 7.3. Five haplotype combinations are possible in the mother (see Section 7.1.3.1), leading to kernels L_1 to L_5 . On the left-hand side, $p_1 = p_2 = 0.1$ is assumed, and the joint likelihood is dominated by L_4 . On the right-hand side, haplotype-specific kernels and the joint kernel of the likelihood are shown for $p_1 = p_2 = 0.9$. It can be seen that the kernel is mostly influenced by likelihood kernels L_2 and L_3 , which are all identical.

For a general formulation of the likelihood in a nuclear family, recall that the kernel of the likelihood is given by

$$L(\theta) = P(g_{\text{par}}) \cdot P(g_{\text{off}}|g_{\text{par}}||\theta) = P(g_F)P(g_M) \prod_{\text{off}} P(g_{\text{off}}|g_F, g_M||\theta), \quad (7.7)$$

where, as before, g_F and g_M denote the paternal and maternal genotypes, respectively. With unknown parental phase, the likelihood in Eq. (7.7) is extended by incorporating the sum across all possible parental phases to

$$L(\theta) = \sum_{\text{phases}} P(g_F, \text{phase}_F) P(g_M, \text{phase}_M) \prod_{\text{off}} P(g_{\text{off}}|g_F, g_M, \text{phase}_F, \text{phase}_M||\theta).$$

In this section, we considered the phase to be an integral part of the missing parental genotype. In the case of missing parental genotypes, we can therefore replace the sum over all possible phases with the sum over all possible parental genotypes—with known phase. Then, the likelihood is given by

$$L(\theta) = \sum_{G_F} \sum_{G_M} P(G_F) P(G_M) \prod_{\text{off}} P(g_{\text{off}}|G_F, G_M||\theta),$$

where G_M and G_F denote the phase-known genotypes of the mother and the father, respectively. It is important to note that the phase is assumed to be unknown in the offspring. If information about the phase in the offspring is to be included in the likelihood, a simple modification of the likelihood is required:

$$L(\theta) = \sum_{G_F} \sum_{G_M} P(G_F) P(G_M) \prod_{\text{off}} \sum_{G_{\text{off}}} P(G_{\text{off}}|G_F, G_M||\theta)$$

This formulation of the likelihood allows the inclusion of offspring with unknown or ambiguous genotypes. The first term $P(G_F)P(G_M)$ also shows that only the parental genotypes—in a more general setting, the founder genotypes—depend on haplotype frequencies. With the usual assumptions as discussed in this section, the dependency on haplotype frequencies is reduced to a dependency on allele frequencies. Offspring genotypes, however, depend on parental genotypes and the recombination fraction only.

7.2 HOW CAN THE RECOMBINATION FRACTION BE ESTIMATED BETWEEN A GENETIC MARKER AND A DISEASE?

So far, we have considered only the linkage between two marker loci. In many applications, interest lies in the investigation of the linkage between a disease and a genetic marker. The two possible simple causal models underlying this idea, which ignore influences other than those of a single genetic locus, are depicted in Figure 7.7 (see also Chapter 11 for a similar figure depicting the concept of disease–marker association). The first part of Figure 7.7 displays a direct relationship between the

disease locus and the disease, and the marker locus is identical to the trait locus. In the second part of Figure 7.7, the marker locus is linked to an unknown disease-causing trait locus. In both cases, the disease is caused by the disease locus, and the relationship is established via the penetrance function (see Chapter 2).

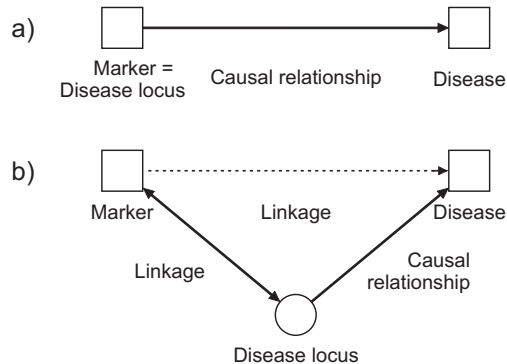


Fig. 7.7 Causal models for genetic linkage. a) The investigated marker is identical to the functional disease-causing locus. b) The marker locus is in linkage with the disease locus that, subsequently, causes the disease.

In Section 7.2.1 we begin by investigating linkage in phase-known pedigrees. The case of general pedigrees will be sketched in Section 7.2.2, and the increase in *LOD* scores by genotyping additional individuals in simple situations will be considered in Section 7.2.3.

7.2.1 How can the recombination fraction be estimated between a genetic marker and a disease in phase-known pedigrees?

If we want to analyze linkage between a disease and a genetic marker, the model becomes more complicated. In the classical approaches considered in this chapter, the disease needs to be fully described with regard to the underlying genetic model. Specifically, the following components have to be included:

- Chromosomal region: autosomal or sex-linked;
- Mode of inheritance: dominant, recessive, or co-dominant, or, more generally, penetrances including frequency of phenocopies;
- Disease allele frequency,
- Fraction of new mutations.

In the following sections, the disease allele frequency will be assumed to be low and we will consider only simple genetic models. These are autosomal and have complete penetrance, no phenocopies, and no mutations. Furthermore, the disease locus will be assumed to have only two alleles. These assumptions will allow a substantial simplification of the problem so that calculations can be carried out without using of a computer.

In this section, we will restrict the analysis to a single genetic marker and a disease locus, and we will estimate the recombination fraction between the two loci. As a result, this analysis is called *two-point analysis*. A term sometimes used synonymously in the literature is *single-point analysis* because only one genetic

marker is involved. However, it should more appropriately be called *single-marker analysis*.

7.2.1.1 How can the recombination fraction be estimated between a genetic marker and an autosomal dominant disease with complete penetrance without phenocopies?

The starting point is the three-generation pedigree displayed in Figure 7.8. The simplest case with an autosomal dominant inheritance is shown. The penetrances for a diallelic trait locus with alleles N and D therefore are $f_0 = 0$ and $f_1 = f_2 = 1$, with f_i as defined in Eq. (2.1). In analogy to Section 7.1.1, the mother is homozygous with the 1 allele at the marker locus, thus not informative for linkage. She is unaffected and therefore she is homozygous with the N allele.

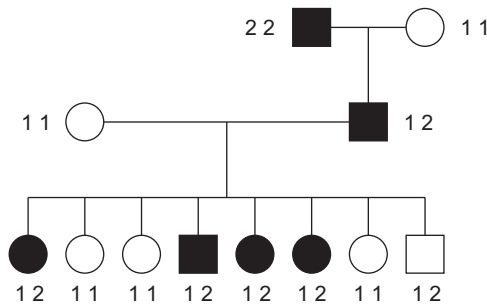


Fig. 7.8 Three-generation pedigree to illustrate linkage analysis between a genetic marker and an autosomal dominant disease with complete penetrance and without phenocopies in a phase-known pedigree. All individuals are genotyped at the marker locus.

The first step consists in determining parental phases. The grandmother is homozygous with the 1 allele at the marker locus. Furthermore, she is unaffected. Because complete penetrance is assumed, the grandmother is known to be homozygous with the N allele. The grandmother therefore transmits one $\frac{1}{N}$ haplotype to her son.

The grandfather is affected, and thus carries at least one D allele. The father has received one D allele from the grandfather, as he is also affected with the disease. The second paternal haplotype therefore is $\frac{2}{D}$.

Haplotypes can be determined easily in the last generation: all offspring carry the maternally transmitted $\frac{1}{N}$ haplotype, and the paternally transmitted allele at the marker locus is the remaining one. In addition, if an offspring is unaffected, it has received exactly one N allele from the father. Otherwise, it has obtained a D allele paternally (see Figure 7.9).

One recombination can be observed in eight informative meioses, yielding $\hat{\theta} = \frac{1}{8} = 0.125$. The kernel of the likelihood function subsequently is $L(\theta) = \theta (1 - \theta)^7$, and the LOD score is

$$LOD = \max_{\theta \in [0, 0.5]} \log \frac{\theta (1 - \theta)^7}{0.5^8} = \log \frac{0.125 (1 - 0.125)^7}{0.5^8} \approx 1.0992.$$

The more general cases of phase-unknown pedigrees and/or missing genotypes can be solved using the same techniques as described in Sections 7.1.2 and 7.1.3. We

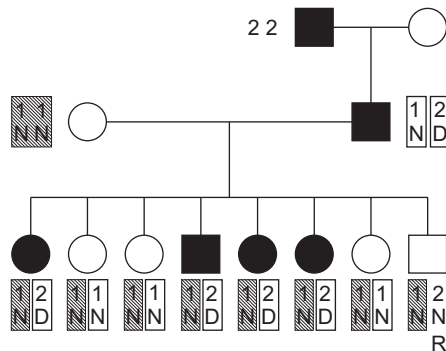


Fig. 7.9 Three-generation pedigree to illustrate linkage analysis between a genetic marker and an autosomal dominant disease with complete penetrance and without phenocopies in a phase-known pedigree. The mother is not informative for linkage. Thus, the maternal genotypes and her transmissions to the third generation are marked gray. The father's phase could be determined and his haplotypes $\begin{smallmatrix} 1 \\ N \end{smallmatrix} \mid \begin{smallmatrix} 2 \\ D \end{smallmatrix}$ are given in boxes. In the offspring, one recombination can be observed, marked by *R*.

want to stress that these techniques rely on the specification of the allele frequencies at both the trait and the marker locus.

7.2.1.2 How can the recombination fraction be estimated between a genetic marker and an autosomal recessive disease without phenocopies in phase-known pedigrees?

The starting point for this section is the three-generation pedigree displayed in Figure 7.10 following an autosomal recessive disease without phenocopies. The penetrances therefore are $f_0 = f_1 = 0$ and $f_2 = 1$. An even simpler example illustrating the fundamental principle of linkage analysis between a genetic marker and an autosomal recessive genetic disease is given as Problem 7.5. In order to determine phase in both unaffected parents, grandparents of both parents need to be genotyped. Furthermore, one affected and one unaffected grandparent is required for each parent.

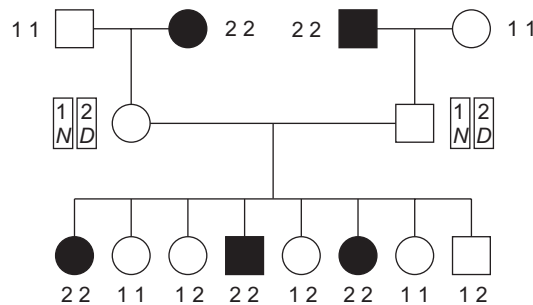


Fig. 7.10 Three-generation pedigree to illustrate linkage between a genetic marker and an autosomal recessive disease with complete penetrance and without phenocopies in a pedigree with both parents being unaffected. Boxes indicate determined parental haplotypes.

There are some issues that make the analysis of a recessive disease more complicated than a dominant one. If the disease is inherited in an autosomal recessive mode of inheritance and if some grandparents are unavailable for the study, the phase in an unaffected parent cannot be determined. Even if both grandparents are phenotyped and genotyped, unaffected parents are only phase-known if one grandparent is affected and the other is unaffected. If, however, both grandparents are unaffected, a parent can be either homozygous with the *N* allele or heterozygous *DN* at the trait

locus. Finally, if $f_0 = 0$, i.e., in the absence of phenocopies, an affected individual is homozygous D and thus non-informative for linkage.

In Figure 7.10, parental phase has already been determined. Both parents are heterozygous at both the marker and the trait locus. In both parents, the D allele is on a $\frac{2}{D}$ haplotype. The phase in the affected offspring can be easily determined because all affected offspring are homozygous D at the trait locus (see Figure 7.11). No recombination occurred in the three affected offspring, and the contribution of each affected offspring to the kernel of the likelihood function therefore is $(1 - \theta)^2$ because both parents are informative for linkage. The contribution of all three affected offspring thus is $((1 - \theta)^2)^3 = (1 - \theta)^6$.

Unaffected offspring can carry either two N alleles or one N and one D allele, the latter even with two different phases (see Figure 7.11). For instance, the second offspring is homozygous 1 1 at the marker locus. If she is also homozygous $N N$ at the trait locus, no recombinations can have occurred, corresponding to $(1 - \theta)^2$. Alternatively, one recombination can have occurred for the case that the second offspring is heterozygous at the trait locus. In this case, she contributes $\theta(1 - \theta)$ to the kernel of the likelihood. However, one needs to distinguish between a paternal and a maternal recombination. All three possible constellations are equally likely, and thus the contribution of the second offspring to the kernel of the likelihood is $\frac{1}{3}(1 - \theta)^2 + \frac{2}{3}\theta(1 - \theta)$. Though the third offspring is heterozygous at the marker locus, her contribution to the kernel of the likelihood is identical, as the same three combinations are possible (see Figure 7.11). The joint kernel of the likelihood is therefore given by

$$L(\theta) = (1 - \theta)^6 \left(\frac{1}{3}(1 - \theta)^2 + \frac{2}{3}\theta(1 - \theta) \right)^5.$$

The three affected offspring do not exhibit a recombination, thus hinting at $\hat{\theta} = 0$. The five unaffected offspring are less informative, and a recombination event cannot be determined unambiguously in them, either. Therefore, a plausible estimate for the recombination fraction is $\hat{\theta} = 0$, yielding

$$L(\hat{\theta}) = L(0) = \left(\frac{1}{3} \right)^5.$$

With $L(\theta)|_{\theta=0.5} = 0.5^{16}$, the maximum LOD score is $LOD = \log L(0)/L(0.5) \approx 2.4309$.

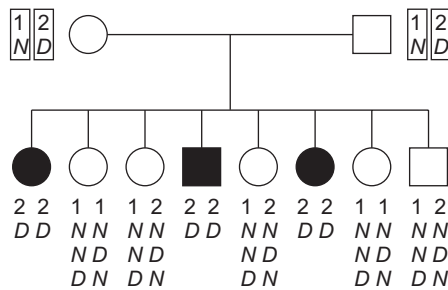


Fig. 7.11 Possible genotypes in the offspring for the three-generation pedigree from Figure 7.10.

7.2.1.3 How can the recombination fraction be estimated between a genetic marker and an autosomal disease with phenocopies and reduced penetrance in phase-known pedigrees?

So far, we have considered only simple Mendelian genetic models. A more general framework will be discussed in the next section. As discussed, affected and unaffected offspring can contribute differently to the kernel of the likelihood. They thus have different degrees of informativity. This will be considered more rigorously in Section 7.2.3. To give a flavor of what will follow in the next sections, we reconsider the pedigree from Figure 7.8. If the genetic model is modified in a way that penetrances are reduced and phenocopies are possible, analysis becomes more complicated.

In this situation, an unaffected individual may carry either one or two disease alleles, while an affected individual could well be homozygous NN . As a consequence, if $0 < f_0 \leq f_1 \leq f_2 < 1$, the mother from Figure 7.8 could have any of the possible three genotypes, leading to three different phases. The genotype is determined from the observed phenotype by using Bayes' formula. Specifically,

$$P(NN|U) = \frac{P(U|NN)P(NN)}{P(U|NN)P(NN) + P(U|DN)P(DN) + P(U|DD)P(DD)}, \quad (7.8)$$

where U denotes unaffected so that, e.g., $P(U|DN)$, denotes 1 minus a penetrance. Equation (7.8) shows that the linkage results can depend on both the penetrances and on the genotype frequencies at the trait locus. The father's phase also cannot be determined, though he is affected. However, with Bayes' formula and the genotype information in the grandparents, it is possible to deduce his genotype distribution at the trait locus. This, in turn, can be used for obtaining the paternal haplotype probabilities. In fact, the calculations are generally more cumbersome than those for phase-unknown pedigrees but simple genetic models, i.e., fully penetrant without phenocopies. Calculations are generally performed with the use of software packages such as ALLEGRO, FASTLINK, GENEHUNTER, LINKAGE, MERLIN, S.A.G.E., SUPERLINK, and VITESSE.

7.2.2 How can linkage analysis be conducted in the general case?

7.2.2.1 What are the main assumptions and how does misspecification affect the results?

In the previous section, we considered only simple genetic diseases following an autosomal mode of inheritance, complete penetrance, no phenocopies, and no mutations. In addition, implicit assumptions imposed have been rather strict and include the following:

- No pedigree errors,
- No misclassification of phenotypes,
- No genotyping errors,
- No errors in the mode of inheritance:
 - No errors in the penetrance function,
 - No misspecified allele frequencies at the trait locus,
 - No errors in mutation rates,

- No misspecified allele frequencies at the marker locus,
- No linkage disequilibrium between the trait and the marker locus (this point is seen by noting that the true genotypes at the trait locus are unknown; this leads to the setting discussed in Section 7.1.3.3).

To repeat, in model-based linkage analysis one assumes that the models describing the trait and marker loci are correctly specified. Violations of these assumptions may lead to increased type I or II error. Some of these assumptions can be generally examined in large pedigrees, such as the assumption of no pedigree errors and no genotyping errors. They usually lead to incompatibility with Mendelian law of inheritance. Other assumptions, such as no errors in the mode of inheritance or in the penetrance model, cannot be verified. One approach to overcome this limitation is to jointly estimate these model parameters and the recombination fraction in *combined linkage and segregation analyses* [74]. However, in most applications the kernel of the likelihood is fairly flat; therefore, this additional estimation of nuisance parameters can lead to very low power and imprecise estimation of parameters. If these nuisance parameters are misspecified, a bias can occur. Below, we describe the different misspecification scenarios and their effect on linkage analysis.

First, when the phenotype, i.e., the affection status is misclassified, a loss in power, i.e., increased type II error levels, is observed [449, 516, 721].

Second, if the penetrance model is misspecified, there is a decreased power to detect linkage [28, 721].

Third, misspecification of marker allele frequencies can lead to false-positive evidence for linkage, i.e., substantially inflated LOD scores, i.e., increased type I error levels, as well as false-negative linkage results, i.e., increased type II error levels [227, 363, 722]. The effects are most pronounced in small pedigrees with key individuals unavailable for genotyping. For example, if no parental marker data are available in nuclear families consisting of an affected sib-pair (ASP; for details, see Chapter 8) together with their parents, a correctly specified model for the trait locus still may result in increased type I error levels [28, 436]. If both the marker and the trait locus model are misspecified, tests are too liberal [721].

Fourth, incorrectly assuming linkage equilibrium when dense markers are available, false positive evidence of linkage can be observed, particularly when heterogeneity LOD score approaches are applied and in ASP analysis when parental data are missing [83, 323]. Bias in ASP analysis can be eliminated by inclusion of parental genotype data and the additional genotyping of unaffected siblings.

Finally, consanguineous pedigrees deserve special attention. Here, false positive evidence of linkage can be observed for markers in linkage disequilibrium when genotype data are missing for any pedigree member [407]. Furthermore, ignoring distant loops leads to false positives [423], and the use of classical test statistics that do not take into account the relatedness of family members almost always leads to false conclusions [406].

Model misspecification leads not only to increased or decreased type I and type II errors but also to biased estimates of the recombination fraction. For example, θ is biased upward if some phenotypes are misclassified, if the penetrance is too high, and

if genetic heterogeneity is ignored (see below). It is biased downward if the sporadic risk is assumed to be too high or if polygenic effects are ignored [28, 515, 721]. In multipoint linkage analysis, the maximum likelihood position of the disease locus may be even pushed off the map [564, 642]. This might be the reason why high LOD scores are observed at the telomeres in some applications. The reason for these biases is not entirely clear, but the following heuristic explanation given by Ott [515] seems reasonable. With unequal gene frequencies, with or without linkage, the disease gene tends to occur in the typed members of the pedigree most often together with the most frequent allele, just because that allele is so common. When it is taken to be less frequent than it really is, its apparently frequent occurrence with the disease allele appears as evidence for linkage. To avoid this telomere problem, an analysis should not be carried out beyond the length of the chromosome. For example, if the last marker on the chromosome is 5cM from the telomere, then a map distance of no greater than 5cM should be used after the last marker. In any case, one has to be specifically cautious when marker genotypes are missing and linkage findings are made at the telomere.

7.2.2.2 What is multipoint analysis and what problems arise with dense marker maps?

To avoid bias in estimates of the recombination fraction, one may use *multipoint analysis* instead of *two-point analysis*. In multipoint analysis, linkage analysis is carried out for a set of markers that are linked to each other against a disease. The marker order and the intermarker distances are assumed to be known. The disease is then placed in each marker interval and left and right of the leftmost and rightmost markers, respectively. Next, the likelihood is computed for all possible positions of the trait locus in the considered chromosomal region. The most likely position of the disease locus subsequently is the ML estimator. With this approach, one estimates a position along a chromosomal region rather than a recombination fraction per marker. This has the advantage that position estimates are more reliable if the disease locus falls in the interval between the leftmost and the rightmost markers. For example, if θ were estimated with an upward bias at a marker left of the disease locus because of model misspecification, a marker on the other side of the disease locus would push the estimate back toward the original position (see Ref. [259] for a detailed discussion). Though this approach has the advantage of giving unbiased parameter estimates, concerns about erroneous type I and type II error fractions still remain [43, 155, 280, 564].

Furthermore, the unbiasedness of the position estimate heavily depends on the correct marker order, the intermarker distance, and the assumption of no interference. If markers are spread out wide with intermarker distances of 5–10 cM, the marker order and the intermarker distance can generally be estimated well using linkage analysis between markers only (see Problem 7.1). Thus, for a standard genome-wide linkage scan with short tandem repeat (STR) markers, which usually involves not more than 1000 genetic markers, intermarker orders and distances can be estimated with sufficient precision.

However, the denser the marker map gets, the harder it is to correctly estimate the intermarker order and the intermarker recombination fraction. For example, if

the intermarker distance is as low as 1 cM, one has to estimate a probability of approximately 1%, and the required number of meioses for this task exceeds the sample size of conventional linkage studies. Furthermore, the denser the marker map gets, the more relevant undetected genotyping errors are. They can superimpose the estimation of intermarker distances and even the intermarker order. This problem increases with the more widespread use of single nucleotide polymorphism (SNP) chips. Here, markers are placed not every cM but several markers per cM. There is no way to reliably estimate the intermarker order and distance. Instead, intermarker order has to be determined using available sequence maps. The physical distance along the map is then used to interpolate, not to estimate, the intermarker distances from estimates of distances between STR markers flanking the SNP markers of interest. Several applications have shown that genome-wide linkage scans involving either STR or SNP markers coincide well in their results [720], this ad hoc approach seems to be appropriate.

For disorders following a clear-cut Mendelian mode of inheritance, the multipoint approach is generally well suited for mapping a chromosomal locus. Here, sufficient information about the disease locus model is generally available, and the misspecification of both the disease and the marker model may be negligible. For complex genetic diseases, however, little information is usually available about the disease locus, while sufficient information has generally been collected for the marker locus. In this case, one may prefer the use of model-free linkage analysis (see Chapter 8). This avoids the need to specify a genetic model for the trait locus. Nevertheless, a simple Mendelian model may well suffice to identify a region where a true disease gene is located, and a simple model-based linkage analysis can be a good compromise between a desirable and a practicable model. Therefore, we want to stress that the linkage step is not the final one in genetic epidemiology. Instead, it can be considered the first. The next steps usually include positional mapping of a region mapped by linkage analysis and identification of the disease gene, or, in complex genetic diseases, risk alleles or haplotypes contributing to the overall disease risk. Therefore, a misspecified genetic model used for linkage analysis is acceptable if it still correctly pinpoints a chromosomal region containing a susceptibility gene for the disease studied. However, especially for complex genetic diseases, there is no guarantee that the assumed simple model will give a correct answer.

7.2.2.3 What are common extensions of the standard linkage models considered so far?

The models considered so far can be extended to yield more realistic genetic models. For example, we have introduced only a single parameter for the recombination fraction, though more crossing over occurs in female than in male meioses. If one aims at modeling sex-specific recombination fractions, two parameters for the recombination fraction are used: one for females θ_f and one for males θ_m . In most examples of the preceding section, only one parent is informative for linkage, and therefore only one recombination fraction was estimable. Furthermore, there are cases where one cannot estimate θ_f without specifying θ_m . This non-independence of estimates may occur, for example, if one offspring exhibits a recombination, but it cannot be determined in which parent it occurred (for details,

see Ref. [516]; for a design that made θ_f indistinguishable from θ_m , see Ref. [299]). For applications, it is important to note that bivariate LOD scores from sex-specific recombination fractions usually are not published because differences in LOD scores are generally negligible (see Ref. [155]).

The penetrance functions chosen in the previous section were also very simple: age dependency of penetrances, the dependency on environmental factors, the interaction between two loci, and the dependency on the parental transmission were not considered. Especially maternal or paternal imprinting has received attention; see, e.g., Ref. [643]. Here, one replaces the three penetrances $P(\text{aff}|NN)$, $P(\text{aff}|DN)$, and $P(\text{aff}|DD)$ with four penetrances $f_0 = P(\text{aff}|NN)$, $f_{1P} = P(\text{aff}|DN)$, $f_{1M} = P(\text{aff}|ND)$, and $f_2 = P(\text{aff}|DD)$, where the first symbol indicates the paternal meiosis (Chapter 2). The linkage analyses are conducted analogously, as described in the preceding section, with an appropriately modified likelihood.

Age dependency of penetrances is important for many applications and is discussed in detail in Refs. [516, 660]. In practice, instead of modeling age dependency by a continuous function, step functions with constant penetrances per age group are generally used. Each of these strata is called a *liability class*. Because penetrances are usually pre-specified, the kernel of the likelihood needs to be modified only appropriately. Liability classes are also used in applications to account for environmental effects.

Another important assumption we introduced in the previous section is locus homogeneity. However, in most applications it may not be reasonable to expect all families to be linked to the same locus, and the assumption of heterogeneity might be appropriate. In this case, one assumes that only a proportion α of families is linked to the locus under investigation. Note that this α from the literature must not be confused with the significance level.

The null hypothesis considered so far is $\alpha = 1$ and $\theta = \frac{1}{2}$. This will also be true for $\alpha = 0$ and any value of θ or for $\theta = \frac{1}{2}$ and any value of α . The bivariate likelihood is therefore given by

$$L(\alpha, \theta) = \alpha L(\theta) + (1 - \alpha)L(\tfrac{1}{2}),$$

and the *heterogeneity LOD score* (*HLOD*) subsequently is

$$HLOD(\theta) = \log \frac{L(\hat{\alpha}, \theta)}{L(\alpha, \theta)|_{\alpha=0, \theta=1/2}}.$$

Under standard conditions, the MLRT statistic of the *HLOD* would be distributed as a χ^2 with 1 degree of freedom (d.f.). However, because both restrictions $\theta = \frac{1}{2}$ and $\alpha = 0$ lead to the same likelihood, standard asymptotic theory does not hold. Instead, the asymptotic distribution is a mixture of a χ_1^2 and a χ_2^2 [603], so that a χ^2 with 2 d.f. gives conservative *p*-values. It is important to note that the value of $\hat{\alpha}$ usually changes as θ changes. Finally, heterogeneity analysis can be considered a simple approach to dealing with locus heterogeneity. In the last few years, several methods have been proposed for two-locus analysis that directly model the joint effect of two different genetic loci. The two-locus models are discussed in detail in Ref. [516].

Four remarks are required to be made at the end of this section. First, X-linked linkage can be studied analogously to autosomal linkage. Only the likelihood changes if the investigated region is outside of the pseudo-autosomal region; the basic principle for linkage analysis remains the same as outlined in the previous Section 7.2.1 (for details, see Refs. [177, 401]).

Second, loops in pedigrees, i.e., marriage or consanguinity loops are usually resolved in computer packages by artificially doubling those individuals that break the loop.

Third, as illustrated in the previous section, not all individuals are informative for linkage at all genetic markers. For example, the affected daughter in the right side of the pedigree displayed in Figure 7.19 is homozygous at the first marker locus, and thus not informative for linkage. The joint consideration of several closely neighboring markers at a time may increase the informativity (see Problem 7.6). However, at the same time the computational burden increases substantially. As pointed out by Broman, “the likelihood is a mighty big sum.” Therefore, algorithms such as the Elston–Stewart algorithm and the Lander–Green algorithm have been developed to reduce the required computations to a minimum. These algorithms are crucial for many calculations in genetic epidemiology and will therefore be discussed both in brief in Section 8.6 and in detail in the Appendix.

Finally, model-based linkage analysis of quantitative traits can be carried out using the same principles as discussed in the previous section. However, rather than employing a penetrance function to establish a connection between the trait locus and the dichotomous outcome, the density function of a continuous distribution is utilized. In this book, we do not exploit this approach further because it is rarely applied in practice. Instead, model-free methods are preferred for quantitative traits in applications. These will be discussed in detail in Chapter 9. For model-based linkage analysis of quantitative traits, the interested reader may refer to Refs. [516, 738].

7.2.3 How much can be gained by genotyping an additional individual from a family? How can power calculations be performed?

The considerations in Section 7.2.1.2 showed that affected and unaffected offspring can contribute differently to the kernel of the likelihood, thus having different degrees of informativity. In applications, a pedigree including information on vital status and affection status is usually drawn during family recruitment. Therefore, the researcher has some idea about the phenotype of various family members. This information can be utilized to focus recruiting effort on the most informative individuals. In this section, we will therefore discuss the gain in informativity by typing an additional individual with known phenotype.

Figure 7.12 shows two pedigrees with one affected and one unaffected parent. The disease is assumed to be rare autosomal dominant with complete penetrance, no phenocopies, and no mutations. Because the disease is rare, it is very likely that the father is heterozygous DN rather than DD . If we further assume the ideal situation for linkage analysis, i.e., $\theta = 0$ between the marker and the trait locus, we

can determine parental phase from the first three individuals of the pedigree. The question now is how much information can be gained maximally by genotyping 1 an unaffected offspring—as in the pedigree on the left—or 2 an affected offspring—as in the pedigree on the right? This can best be measured on the LOD score scale.

The gray-marked unaffected offspring in the left pedigree is heterozygous $1\ 3$ at the marker locus. The haplotypes therefore are $\frac{1}{N}$ and $\frac{3}{N}$. Only the paternal meiosis is informative for linkage; no recombination occurs in this offspring, and the contribution to the kernel of the likelihood is $1 - \theta$. The contribution of this individual to the LOD score therefore is

$$LOD_{\text{all}}(\theta) = LOD_{\text{all but last}}(\theta) + \log \frac{1 - \theta}{\frac{1}{2}}.$$

In the optimal case of $\theta = 0$, the gain in the LOD score by genotyping an unaffected individual therefore is

$$\log \frac{(1 - \theta)}{\frac{1}{2}} \Big|_{\theta=0} = \log 2 \approx 0.3010.$$

Again, we want to stress that this result relies on the observation of no recombination. If there is at least one recombination, $\theta = 0$ is impossible and the maximum *LOD* score is $-\infty$.

In the pedigree on the right, the situation is similar: the gray-marked affected offspring exhibits no recombination and subsequently contributes $1 - \theta$ to the kernel of the likelihood. Consistently, the maximal increase in the *LOD* score by genotyping an affected individual also is ≈ 0.3012 .

In analogy to Figure 7.12, we now consider the case of a simple autosomal recessive trait. As before, we assume complete penetrance, no phenocopies, no mutation, and that the phase has been determined (see Figure 7.13). Again, we consider the best case scenario of $\theta = 0$. In contrast to Figure 7.12, where only one parent was informative for linkage, both parents are informative for linkage in Figure 7.13 if both parental phases are considered fixed. These can be derived from the available data if one is willing to assume first that the disease is rare autosomal recessive, so that both parents are heterozygous $D\ N$, and second that no recombination has occurred in the

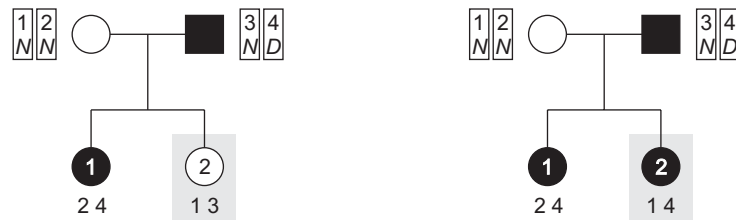


Fig. 7.12 Gain in information by additionally genotyping the individuals in the gray box. Assumed is a simple autosomal dominant genetic disease. In both pedigrees, the parental phase is assumed to be determined from the first offspring under the assumption of no recombination.

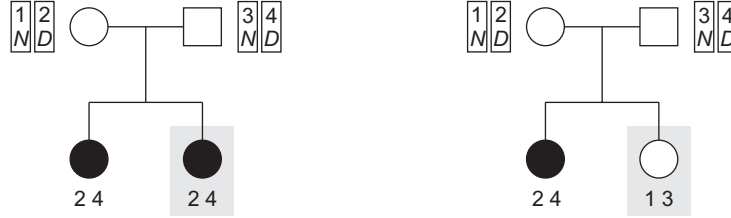


Fig. 7.13 Gain in information by additionally genotyping the individuals in the gray box. The mode of inheritance is assumed to be simple autosomal recessive. Furthermore, the parental phase is assumed to be known in both pedigrees.

meioses of the first affected offspring. Then, in the pedigree on the left, the offspring is affected and shows neither a maternal nor a paternal recombination. Consequently, the contribution of this individual to the kernel of the likelihood is $(1 - \theta)^2$, and the maximal gain in LOD score by genotyping an affected individual in this case is

$$\log \frac{(1 - \theta)^2}{\left(\frac{1}{2}\right)^2} \Big|_{\theta=0} = \log 4 \approx 0.6020.$$

As already seen in Section 7.2.1.2, unaffected individuals are less informative than affected individuals in the case of an autosomal recessive mode of inheritance. Their contribution to the kernel of the likelihood is only $\frac{1}{3}(1 - \theta)^2 + \frac{2}{3}\theta(1 - \theta)$. This is illustrated in the pedigree on the right in Figure 7.13. This unaffected individual can be either homozygous N at the trait locus or heterozygous DN . In the first case, no recombination has occurred. In the other situation, exactly one recombination has happened. This, however, could have taken place in either the maternal or paternal meiosis. All three events are equally likely and can not be distinguished.

In a simple autosomal recessive disease, the maximal contribution of an unaffected individual to the LOD score therefore is

$$\log \frac{\frac{1}{3}(1 - \theta)^2 + \frac{2}{3}\theta(1 - \theta)}{\left(\frac{1}{2}\right)^2} \Big|_{\theta=0} = \log \frac{4}{3} \approx 0.1249.$$

In Figures 7.12 and 7.13 we consider single families only and introduce several assumptions on the mode of inheritance and the available data. With these assumptions we are able to derive the maximal increase in the LOD score by genotyping additional individuals. Our approach is simple and applies only to offspring for whom parental phase has already been determined. It can be used for power calculations when simple Mendelian diseases are studied. In fact, one needs only to count how many informative meioses one needs to collect until the desired LOD score can be achieved with confidence. However, if the mode of inheritance is more complex and phenocopies have to be taken into account or reduced penetrance occurs, counting informative meiosis is insufficient. Several computer packages exist for estimating power in these situations, e.g., SIMLINK and SLINK [537, 700]. In the first step,

the user has to provide all the information about each pedigree. In the second step, the disease locus is simulated to segregate in each family conditional on all observed phenotypes. In the third step, marker genotypes are simulated given specific recombination fractions between marker and trait loci. These three steps are repeated over a sufficiently large number of replicates. In this way, the distribution of the LOD scores and the *expected LOD score (ELOD)* is obtained. In the ELOD approach, both the genetic model and the recombination fraction are considered fixed.

If the expected LOD score of a family is maximized per replicate over all values of the recombination fraction, one obtains the *maximized LOD score (MLOD)*. With a sufficient number of replicates, this yields the distribution of the MLOD score and an estimate of the *expected maximized LOD score (EMLOD)*. For a more thorough discussion, the interested reader is referred to the excellent textbook by Ott [516].

7.3 WHAT IS SIGNIFICANT EVIDENCE OF LINKAGE?

In the previous sections, we considered linkage between two pre-specified marker loci and between a single genetic marker and a disease, and we conducted a single test. Therefore, it might be reasonable in this case to work with the classical significance level $\alpha = 0.05$. In most applications, however, the situation will be different. A whole chromosomal region containing several candidate genes might be of interest, or several candidate genes from different chromosomes might be under consideration. Not knowing the position of a putative disease gene, one might prefer a systematic search in the whole genome. In any case, this leads to a multiple testing problem involving a few genetic markers and a small set of candidate genes from one or two different functional groups in the best case. In the worst case, hundreds of thousands of SNPs are typed on an SNP chip along the whole genome. How should we deal with this extreme case of multiple testing? Unfortunately, there is no answer to this question. Instead, the options for handling the multiple testing problem substantially vary.

Historically, it has become a rule of thumb to declare significant linkage if the *LOD* score exceeds 3. This critical value was derived by Morton [478] using the *sequential probability ratio test (SPRT)* [684]. He showed for a single-marker analysis and sequential sampling that an *LOD* score of 3 should be obtained on autosomes for a type I error of 5%. He also found that linkage is unlikely and can therefore be excluded for model-based linkage analysis if $LOD < -2$. Finally, for sex chromosomes, the *LOD* score should exceed 2 for significant linkage. Although the situation of sequential testing is no longer applied in practice, the critical value of $LOD = 3$ is still in use based on two lines of justification.

One approach considers the posterior probability for linkage and will be discussed below. The alternative approach relates *LOD* scores to *p*-values, and a pre-specified significance level α is used to check whether $p < \alpha$. However, the question is which significance level α should be used for claiming “significance.” Suppose that a complete genome-wide scan has been conducted, and we aim at controlling the *family-wise error fraction*. This means that the probability for rejecting any true

null hypothesis may not exceed the *global significance level* $\alpha = \alpha_G$. If tests are performed at each of the K genetic markers and if these tests are independent, then standard approaches can be used to determine *point-wise*, i.e., *local significance levels* α_P . The local significance level is the probability for rejecting the true null hypothesis at a specific chromosomal position. One classical method is the very conservative Bonferroni approach, which leads to $\alpha_P = \alpha_G/K$. For example, if the potential number of SNPs to be tested is 1,000,000, then $\alpha_P = 5 \cdot 10^{-8}$ for a global significance level of $\alpha_G = 0.05$. However, the statistical tests along a chromosome are not independent, and standard statistical approaches ignore the dependencies, i.e. the linkage between genetic markers.

To avoid the unduly extreme conservative Bonferroni adjustment for multiple testing, Lander and colleagues [391, 393, 394] (see also Refs. [206, 209]) determined the asymptotic relation between global and point-wise significance levels under the assumption of no interference and an incompletely dense marker map. In this case, many linkage test statistics are asymptotically normal and can therefore be approximated by an Ornstein–Uhlenbeck process [65]. A proof of this result has been given by Cienko [122].

The relation between global and point-wise significance levels is given by

$$\alpha_G \approx \alpha_P \left(C + 2 \varrho \cdot L \cdot (\Phi^{-1}(\alpha_P))^2 \right), \quad (7.9)$$

where

- C is the number of investigated chromosomes,
- L is the total length of the investigated chromosomal regions in M,
- ϱ is a constant measuring the crossing over intensity or, more precisely, is related to the first derivative of the covariance function between genotypes at two different loci, and
- Φ is the distribution function of the standard normal distribution.

For example, the total genome size L is approximately 33 M, and the total number of chromosomes C is 23 for the human. The crossing over intensity depends on the study design and the familial relationship; common values are given in Table 7.2.

Using Eq. (7.9), it can be seen that the point-wise significance level α_P has to be approximately $4.9 \cdot 10^{-5}$, in a classical genome-wide linkage study using the LOD score method as discussed in this chapter to guarantee $\alpha_G \leq 0.05$. This p -value corresponds to an LOD score of ≈ 3.3 by using the relationship $z = \sqrt{\chi^2} = 2 \ln(10) LOD$. Here, z is the value of a standard normally distributed variable. Similarly, $\alpha_P \approx 2.2 \cdot 10^{-5}$, corresponding to $LOD \approx 3.61$ for sib-pair studies with the same global significance level.

This approach has been strongly criticized for the unrealistic assumptions made to obtain the functional relationship between the global and the point-wise significance levels [180, 479, 725]. The concerns are related to both crucial assumptions: the absence of interference and the presence of an infinitely dense marker map. Because the criticisms related to the assumption of no interference cannot be refuted, some authors (see, e.g., Ref. [180]) have recommended reporting point-wise p -values from

Table 7.2 Linkage method, familial relationship and crossing over intensity.

Linkage method	ϱ	Linkage method	ϱ
LOD score analysis	1		
Allele sharing in human (Chapters 8 and 9)			
Sibs and half-sibs	2	Grandparent-grandchild	1
Avuncular	2.5	First cousin	8/3
First cousin, once removed	20/7	Second cousin	16/5

ϱ is the measure for crossing over intensity.

single-marker analyses instead of tackling the multiple testing problem. However, one should keep in mind that almost all linkage analyses are carried out assuming absence of interference.

The second concern is related to the assumption of an infinitely dense marker map. In fact, if marker informativity is incomplete, stringent application of the Lander and colleagues criterion leads to overly conservative tests. Therefore, Lander and Kruglyak [393] studied the effect of incomplete marker informativity. They found that the *LOD* score thresholds should be decreased by

- 20% for a 10 cM map,
- 15% for a 5 cM map,
- 10% for a 2 cM map, and
- 7% for a 1 cM map.

Thus, for sib-pair analyses (Chapters 8 and 9) point-wise *LOD* scores of approximately 3.07 and 2.89 are required for 5 cM and 10 cM maps, respectively, to ensure a global test level of 5%. This means that a genetic region shows linkage in genome-wide scans using sib-pairs if the *LOD* score is ≥ 3 . For standard *LOD* score analysis, the *LOD* score should exceed 2.80 for a 5 cM map and 2.97 for a 2 cM map. Genome-wide scans based on microsatellites usually have an average marker distance of 5–10 cM, thus an *LOD* of 3 and above can be considered significant evidence for linkage. If, however, a SNP scan is carried out, more stringent significance levels are required because the marker density and the information content (Chapter 3) are increased. It is important to note that a *LOD* score of 3 asymptotically corresponds to $p = 10^{-4}$; $p < 10^{-3}$ holds in any case (see Problem 7.8). Thus, for finite samples, an *LOD* score of 3 need not correspond to $p = 10^{-4}$.

Different point-wise *p*-values are required if the aim is the independent confirmation of a significant initial linkage signal in a second independent sample. In this second study, we assume that only a few markers from a single region of interest are investigated. The hypothesis to be tested is different from that of the initial whole genome scan. There is still a multiple testing problem, but it is less pronounced than for the genome-wide search. If a single chromosome with a genomic length of about 50 cM = 0.5 M is investigated at a global significance level of 5%, the required

LOD score at any point in the candidate region can be calculated using Eq. (7.9) to be approximately 1.63, corresponding to $\alpha_P = 0.003$, for a sib-pair study with an infinitely dense marker map. For a 1 cM grid, the required *LOD* score is about 1.63.

Lander and Kruglyak [393] have proposed standards “to promote rigor and communication.” Their work has indeed increased communication, as their proposal has been criticized by several authors with probably the most rigorous concerns raised by Elston [180]. Specifically, Lander and Kruglyak introduced a classification for linkage results based on the observed point-wise *p*-value of a genome-wide scan. They used the terms *suggestive linkage*, *significant linkage*, and *highly significant linkage*.

Though their concept for classifying linkage findings looks appealing at first glance, *p*-values are related only to significance. Specifically, small genetic effects become significant with a large sample size, whereas even large effects may remain non-significant in an underpowered small study. According to classical statistical test theory, a statistical test is a function with values in the interval 0 to 1, which reflect the probability that the null hypothesis of no linkage is rejected given a specific value of the test statistic. And the world of a statistician is simple because the null hypothesis is either true or false. The heart of the idea of classifying linkage results according to their degree of evidence is good, and it would therefore be helpful if the standard concept of testing for linkage is extended to the more appropriate profile—a test concept (Ref. [693]). Even simpler would be the routine use of appropriate effect size measures to judge the relevance of the linkage finding.

Furthermore, when working with the terms suggestive, significant, and highly significant linkage, the investigator automatically relates *p*-values from linkage analyses to the second concept, which is based on the posterior probability for linkage. Here, one aims at determining the probability that a significant linkage finding is indeed true or—when more information is available—estimating the probability for linkage to a chromosomal region given the data.

Finally, the terms seem to have been selected specifically for the situation of an a priori probability for linkage to a specific region of 5%. This assumption is most likely not fulfilled for most complex genetic diseases where more than one genetic locus is involved. In addition, a huge number of linkage and association studies have been carried out in the past decade; thus, the a priori probability for linkage to a chromosomal region deviates from 5%. Though the second concept relating a priori to a posteriori probabilities for linkage using the Bayes’ formula could still be employed in these cases, choosing a reasonable a priori probability is problematic. Indeed, no non-informative prior can be chosen, making this approach somewhat obsolete.

Nevertheless, one should be aware of the second, the Bayesian, argument for why an *LOD* score of 3 is sufficient for claiming evidence of linkage. For this, we define $PO = P(H_1)/P(H_0)$ as the *a priori odds* or *forward odds* for linkage. Because H_0

and H_1 are disjunct and a complete decomposition of the set of hypotheses, the *a posteriori probability for linkage* given the data is obtained via the Bayes' rule by

$$\begin{aligned}
 P(H_1|\text{data}) &= \frac{P(\text{data}|H_1)P(H_1)}{P(\text{data})} \\
 &= \frac{P(\text{data}|H_1)P(H_1)}{P(\text{data}|H_0)P(H_0) + P(\text{data}|H_1)P(H_1)} \\
 &= \frac{10^{LOD} \cdot PO}{1 + 10^{LOD} \cdot PO}.
 \end{aligned} \tag{7.10}$$

The *a priori* probability for linkage $P(H_1)$ is usually set to 0.02, yielding an *a priori* odds PO for linkage of approximately $0.0204 \approx 0.02$. The argument for the choice of the *a priori* probability for linkage is simple: there are approximately 50 independent chromosomal regions in the human genome. Assuming that there is a linked chromosomal region, we obtain $P(H_1) = 1/50 = 0.02$. Equation (7.10) can be used to calculate the *a posteriori* probability for linkage given the LOD score result. With *LOD* scores of 1, 2, and 3, the *a posteriori* probability for linkage is approximately 17, 67, and 95%, respectively. Thus, in 5% of all cases, there is no gene when a linkage study reveals an *LOD* score of 3.

In our opinion, one should report both the lowest *p*-value from multipoint analyses and results from single-marker evaluations. None of these should be corrected for multiple testing. More important than reporting of *p*-values is statement of linkage results in terms of effect estimates and confidence intervals, which is unfortunately seldom done in practice.

The whole discussion about applications of significance and non-significance can be reduced to a minimum if an initial finding can be validated in a second study with a different sample and, preferably, different investigators. More importantly, it can be completely avoided after identification and functional characterization of the disease gene and the relevant mutations.

7.4 PROBLEMS

Problem 7.1 Consider the pedigree displayed in Figure 7.14. Three markers M_1 , M_2 , and M_3 have been genotyped in this pedigree and are located on a single chromosome. Which marker order yields the lowest number of recombinations?

Problem 7.2 Consider the pedigree displayed in Figure 7.15.

Problem 7.2.1. How many informative meioses does the pedigree have?

Problem 7.2.2. What is the kernel of the likelihood of this pedigree?

Problem 7.2.3. Estimate the recombination fraction in this pedigree.

Problem 7.2.4. Determine the *LOD* score, the MLRT, and the asymptotic *p*-value of the MLRT associated with the estimated recombination fraction.

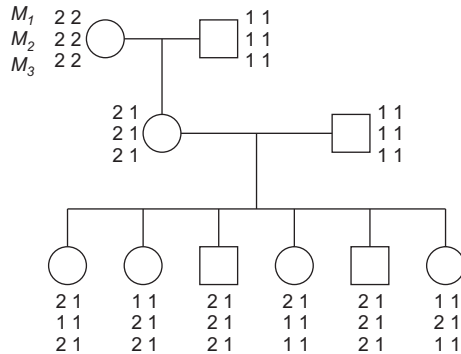


Fig. 7.14 Three-generation pedigree for Problem 7.1. The three genetic markers M_1 , M_2 , and M_3 have been genotyped in all individuals. Which order is the most plausible one?

Problem 7.3 Consider the pedigree displayed in Figure 7.16. It is similar to the pedigree from Section 7.1.1. However, the mother is now heterozygous at both marker loci.

Problem 7.3.1. What is the kernel of the likelihood of this pedigree?

Problem 7.3.2. What is the LOD score at $\theta = 0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4$?

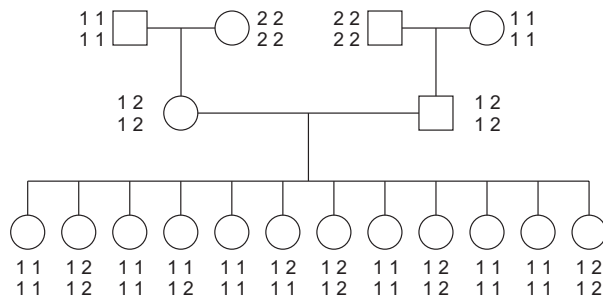


Fig. 7.15 Three-generation pedigree for Problem 7.2. Both parents are heterozygous at both marker loci. Can the phase be determined in both parents?

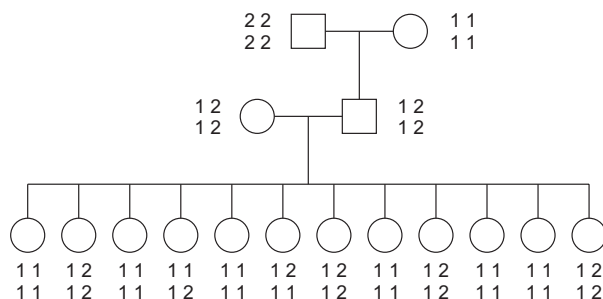


Fig. 7.16 Three-generation pedigree for Problem 7.3. The mother is heterozygous at both marker loci. Which haplotype combinations are possible in the mother?

Problem 7.4 Consider the pedigree displayed in Figure 7.17. It is similar to the pedigree from Section 7.1.1. However, we now assume that the grandmother has not been genotyped.

Problem 7.4.1. What are the possible haplotype combinations in the grandmother?

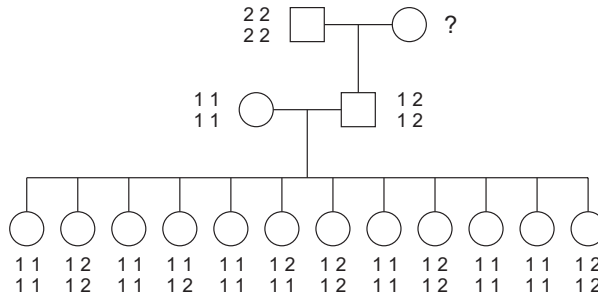


Fig. 7.17 Three-generation pedigree for Problem 7.4. The grandmother was unavailable for genotyping. What is the kernel of the likelihood function?

Problem 7.4.2. What is the kernel of the likelihood for each possible grandmaternal haplotype combination?

Problem 7.4.3. What is the probability for each possible grandmaternal haplotype combination?

Problem 7.4.4. What is the kernel of the joint likelihood function if the allele frequencies are 0.5 at both SNP markers?

Problem 7.5 There is a rare autosomal recessive disease segregating with complete penetrance and no phenocopies in the family displayed in Figure 7.18. All family members have been genotyped at all genetic markers.

Problem 7.5.1. Determine the parental haplotypes.

Problem 7.5.2. How many recombinations have occurred?

Problem 7.5.3. What is the kernel of the likelihood function? What is the LOD score function? What is the maximum *LOD* score? At which recombination fraction?

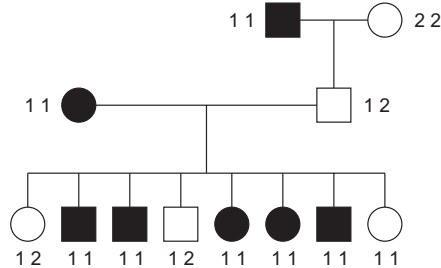


Fig. 7.18 Three-generation pedigree for Problem 7.5. Determine the parental haplotypes and the maximum *LOD* score.

Problem 7.6 There is a rare autosomal dominant disease segregating with complete penetrance and no phenocopies in the family displayed in Figure 7.19. The two genetic markers are closely linked at $\theta = 0.005$.

Problem 7.6.1. Which individuals of the second generation are informative for linkage analysis for the disease with marker 1?

Problem 7.6.2. Which individuals of the second generation are informative for linkage analysis for the disease with marker 2?

Problem 7.6.3. Construct haplotypes which, in turn, serve as a single new marker.

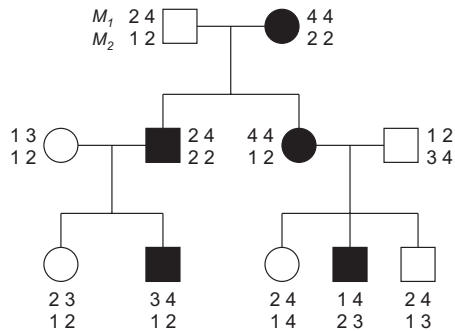


Fig. 7.19 Three-generation pedigree for Problem 7.6. Is the *LOD* score higher if both markers are considered jointly?

Problem 7.7 There is a rare autosomal dominant disease segregating with complete penetrance, no phenocopies, and no mutation in the families displayed in Figure 7.20.

Problem 7.7.1. Is there a difference in the *LOD* score function between pedigrees (a) and (b)?

Problem 7.7.2. Is there a gain in information between pedigrees (b) and (c)?

Problem 7.7.3. What becomes crucial in linkage analysis of pedigree (c) compared to pedigrees (a) and (b)?

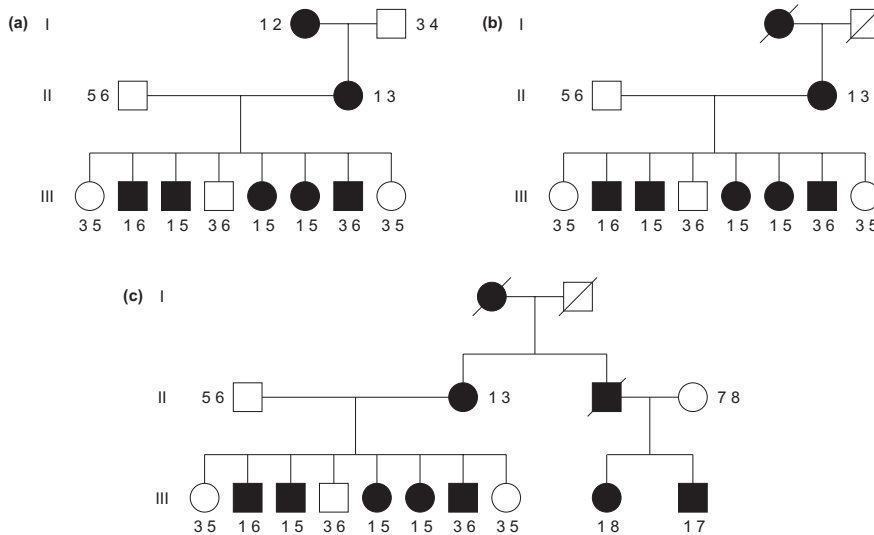


Fig. 7.20 Pedigrees for Problem 7.7. There is a rare autosomal dominant disease segregating in the families, and all or most family members have been genotyped at one genetic marker. What are the differences between the families and how do these differences affect the *LOD* score function?

Problem 7.8 Show that $LOD = 3$ always leads to $p < 10^{-3}$. Furthermore, show that an LOD score of 3 asymptotically corresponds to a p -value of 10^{-4} .

URLs

ALLEGRO

<http://linkage.rockefeller.edu/soft/list1.html>

EASYLINKAGE

A graphical user interface for automated linkage analyses

<http://compbio.charite.de/genetik/hoffmann/easyLINKAGE/>

FASTLINK

<http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html>

GENEHUNTER

<http://www.broadinstitute.org/ftp/distribution/software/genehunter/>

LINKAGE

<ftp://linkage.rockefeller.edu/software/linkage/>

MERLIN

<http://www.sph.umich.edu/csg/abecasis/Merlin/>

S.A.G.E.

<http://darwin.cwru.edu/sage/>

SIMLINK

<http://linkage.rockefeller.edu/soft/simlink.html>

SLINK

<ftp://linkage.rockefeller.edu/software/slink/>

SUPERLINK

<http://bioinfo.cs.technion.ac.il/superlink/>

VITESSE

<http://users.ox.ac.uk/ayoung/vitesse.html>

8

Model-Free Linkage Analysis for Dichotomous Traits

In the previous chapter, we discussed model-based methods for detecting linkage to a dichotomous trait and the underlying assumptions. These include the correct specification of models for both the trait and the marker loci. For complex diseases, however, it is impossible to correctly specify the entire model causing familial correlation. In contrast, model-free methods do not require the correct specification of the model for the trait locus and are therefore often preferred for complex traits in applications. In this chapter, we first discuss the fundamental idea underlying model-free linkage methods and introduce the most commonly used measures of concordance between pairs of individuals from the same family in Section 8.1. For simplicity, we here restrict our attention to autosomal single-locus models. In Section 8.2, we give a mathematical foundation for the affected sib-pair (ASP) analysis using simple genetic models. Several standard tests for ASPs are introduced in Section 8.3, and their properties are discussed in Section 8.4. The problem of sample size and power calculation using ASPs is considered in Section 8.5. The remaining sections are devoted to a number of extensions. In Section 8.6, linkage to multiple marker loci is described, and in Sections 8.7 and 8.8 we discuss the application of ASP methods to large sibships and extended pedigrees. Finally, some further extensions of the ASP approach are sketched in Section 8.9.

8.1 WHAT IS THE BASIC IDEA OF MODEL-FREE LINKAGE ANALYSIS?

The fundamental principle underlying all model-free approaches to linkage analysis is the *principle of similarity*, which can be traced to the work by Penrose in 1935 [525] (for an excellent review on the work of Penrose on sib-pairs, see Ref. [178]). According to this principle, sib-pairs that are phenotypically similar should also be genotypically similar, and sib-pairs that are phenotypically dissimilar should also be genotypically dissimilar. Obviously, this principle requires that there must be a gene that causes some variability on a trait. Furthermore, we assume the study of markers that are linked to this causal gene.

The first question is: How can phenotypic and genotypic similarity be measured? In this chapter, we will consider 0–1 traits. Therefore, pairs are either concordant, i.e., similar, or discordant, i.e., dissimilar. Consequently, because mostly sib-pairs with at least one affected are studied, there are either *concordant sib-pairs*, i.e., ASPs, consisting in a pair of affected siblings, or *discordant sib-pairs (DSPs)*, where one sibling is affected, while the other is unaffected, and no measure of phenotypic similarity is required. This simple situation is a bit more complicated for quantitative traits, where several measures of phenotypic similarity have been proposed. These will be discussed in Chapter 9.

For genotypic similarity, the simplest measure is the *identical by state (IBS)* score. Given two individuals who have been genotyped at a genetic marker, it is possible to determine the number of alleles that they have in common. For example, both offspring from Figure 8.1a have one 1 and one 3 allele. By state, they therefore have two alleles in common, thus $IBS = 2$. Similarly, the IBS value of the offspring in family (g) is $IBS = 2$, even though this family is not informative for linkage because both parents are homozygous at the marker locus. To incorporate this information for linkage analysis, it is therefore more important to see whether a sib-pair shares copies of the same parental alleles, i.e., the same grandparental gametes. Hence, an allele is said to be *identical by descent (IBD)* for a pair of relatives if it has been inherited from the same common ancestor.

This is illustrated in Figure 8.1. In the first row, families in which parents are heterozygous with different alleles are displayed. It is therefore possible to uniquely determine the number of alleles IBD in their offspring. In families (d) and (e) the mother is homozygous at the marker locus. Therefore, it is impossible to determine which 3 allele has been transmitted to the offspring. Because the sibs in pedigree (d) have received different alleles from their father, the paternally inherited alleles can be determined not to be IBD. As the IBD status of the maternally inherited alleles is indistinguishable, the IBD value of sib-pair (d) is either 0 or 1. Similarly, the IBD value is 1 or 2 for the sib-pair shown in Figure 8.1e, as the paternally transmitted allele is IBD.

The situation is slightly more complicated in pedigree (f). Both parents are heterozygous with identical alleles. If at least one offspring were homozygous, we could uniquely determine the IBD value of the sib-pair (see Problem 8.1). Because

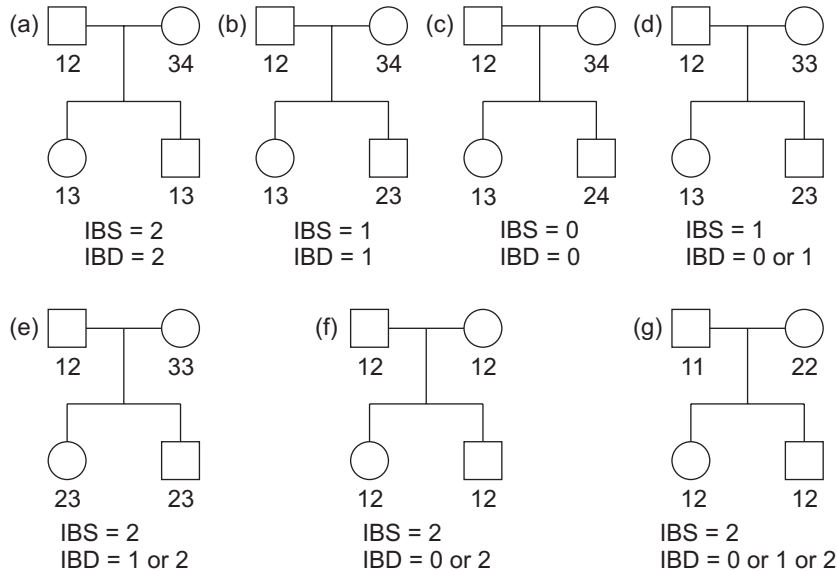


Fig. 8.1 Seven nuclear families with different constellations for identical by descent (IBD) and identical by state (IBS) scores.

both are heterozygous, the IBD value of the sib-pair is either 0 or 2. In family (g), both parents are homozygous at the marker locus, and no information about segregation is available. Accordingly, the IBD value of the sib-pair can be 0, 1, or 2. Finally, it is important to note that only those seven different constellations for IBD values shown in Figure 8.1 are possible for a sib-pair if a single genetic marker has been genotyped.

Using these measures of genotypic and phenotypic similarity, statistical tests for linkage can be constructed by deriving the expected degree of similarity under the assumption of no linkage and comparing the observed and expected numbers. These approaches do not require specification of a disease model and are therefore called *model-free*. Because they rely on the number of alleles shared IBD or IBS, they are sometimes also termed *allele-sharing* methods. A third name for these methods is *non-parametric linkage*. However, most of the approaches discussed in Section 8.3 involve parameters that vary with the degree of linkage. Therefore, we prefer the term “model-free” over “non-parametric.” Furthermore, some model-free linkage methods can be shown to be identical to specific model-based linkage approaches (see Section 8.4) and thus are weakly model-based. Furthermore, they may have maximal power under certain genetic models and are therefore weakly model-based in this sense.

At the end of this section we ask what kind of IBD values would be expected for sib-pairs if we assume the null hypothesis of no linkage, meaning that phenotypes and genotypes are independent. Considering, therefore, the IBD distribution in

the absence of any phenotype and assuming that the parents are heterozygous with different alleles, the IBD value in the offspring can be determined unambiguously.

For simplicity, we assume that the parents have genotypes 1 2 and 3 4 respectively. The possible IBD values in sib-pairs are listed in Table 8.1. There are in all 16 different genotype combinations. Four combinations yield an IBD value of 2 and 0, respectively. The remaining eight combinations result in an IBD value of 1. With $z_j = P(\text{IBD} = j)$, the null hypothesis of no linkage is therefore given by

Table 8.1 Possible identical by descent (IBD) values of a sib-pair where parental genotypes are given by 1 2 and 3 4.

Genotype of 2nd child	Genotype of 1st child			
	1 3	1 4	2 3	2 4
1 3	2	1	1	0
1 4	1	2	0	1
2 3	1	0	2	1
2 4	0	1	1	2

$$(z_0, z_1, z_2) = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right). \quad (8.1)$$

8.2 WHY IS AFFECTED SIB-PAIR ANALYSIS REASONABLE?

Under the alternative hypothesis of linkage, allele sharing in ASPs should be different from the expected distribution of Eq. (8.1). Indeed, the functional relationship between the IBD probabilities and the recombination fraction has been derived for general single-locus models by Suarez and colleagues [649].

For example, for a simple recessive genetic model, i.e., an autosomal, fully penetrant recessive model without phenocopies and new mutations, the IBD probabilities as functions of the recombination fraction are given by (see Problem 8.2)

$$\begin{aligned} P(\text{IBD}_m = 2|\text{FD}) &= \theta^4 + 2\theta^2(1 - \theta)^2 + (1 - \theta)^4, \\ P(\text{IBD}_m = 1|\text{FD}) &= 4\theta(1 - \theta)\Psi, \\ P(\text{IBD}_m = 0|\text{FD}) &= 4\theta^2(1 - \theta)^2 \end{aligned}$$

for ASP families consisting of two affected offspring and two unaffected parents.

Similarly, for a simple dominant model and an ASP family with two affected offspring and one affected and one unaffected parent, the relationship is (see Problem 8.3)

$$\begin{aligned} P(\text{IBD}_m = 2|\text{FD}) &= \frac{1}{2}\Psi, \\ P(\text{IBD}_m = 1|\text{FD}) &= \frac{1}{2}, \\ P(\text{IBD}_m = 0|\text{FD}) &= \theta(1 - \theta). \end{aligned}$$

Figure 8.2 illustrates the functional relationship between the IBD probabilities and the recombination fraction. It can be seen that all six displayed IBD probabilities are

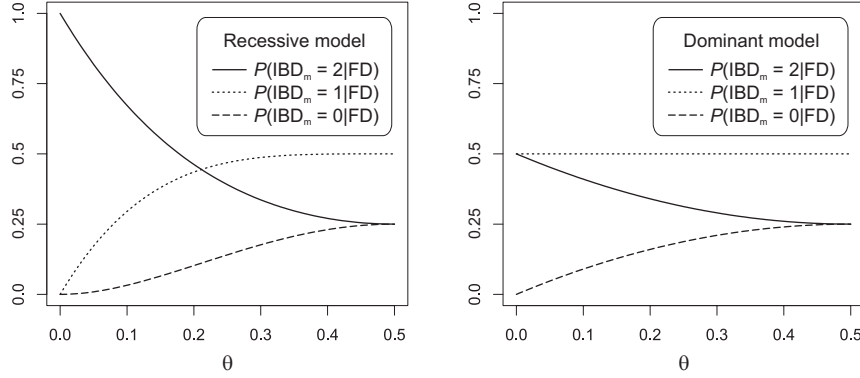


Fig. 8.2 The identity by descent (IBD) probability z_i at a marker locus for two specific family constellations and simple genetic models as functions of the recombination fraction θ . The left and right parts of the figure show the relationships for the recessive and the dominant model, respectively.

identical to their values expected under the null hypothesis of no linkage if $\theta = \frac{1}{2}$. Furthermore, the probability that an ASP shares two alleles IBD z_2 is always greater than $\frac{1}{4}$ unless $\theta = \frac{1}{2}$. For some genetic models, it therefore seems reasonable to test for linkage by investigating whether z_2 is indeed larger than its expected value. However, the probability decreases quadratically with increasing θ . Thus, the power to detect linkage will steeply decrease with increasing distance between trait and marker locus.

It can also be seen that $z_0 \leq \frac{1}{4}$. Finally, the simple dominant model shows that z_1 need not be of value for linkage analysis for some genetic models. Here, z_1 is identical to its expected value under the null hypothesis of no linkage. This result already indicates that statistical tests for linkage that focus on the trinomial distribution of IBD values, thus having 2 degrees of freedom (d.f.), may not be as powerful as tests that are based on a one-parameter statistic that is derived from the trinomial distribution. However, statistical tests using IBD sharing will in general be reasonable for testing linkage.

8.3 WHAT ARE COMMON TESTS FOR AFFECTED SIB-PAIR ANALYSIS?

Increased IBD and IBS sharing can be tested by various methods. Brief overviews are given in Refs. [303, 359]. Many different test statistics and their implementations in software packages have been compared under several two-locus disease models by Davis and Weeks [154]. In the following sections we consider a sample of n independent ASPs. As already shown in Section 8.1 for a single genetic marker, the IBD distribution cannot be determined unambiguously in all cases but has to

be estimated. In the single-marker case, only seven different IBD distributions are possible, and they can easily be estimated. For the multipoint situation, we assume that the IBD distributions have been estimated for all n ASPs at the chromosomal position of interest using a suitable algorithm as described in Section 8.6 and the Appendix. The true and estimated probabilities that sib-pair i shares j alleles IBD are denoted by z_{ij} and $\hat{z}_{ij}$, respectively.

In Section 8.3.1, we consider popular likelihood ratio tests, while Section 8.3.2 deals with common score- and Wald-type tests, which could well be derived as likelihood ratio tests [540]. The use of the different tests is illustrated at the end of Section 8.3.2 by reanalyzing the data of chromosome 16 from a genome-wide search for type I diabetes. Finally, we discuss the use of IBS statistics for ASPs in Section 8.3.3.

8.3.1 What are the maximum LOD score and the triangle test?

8.3.1.1 Which of the null hypotheses are tested in the maximum LOD score test and the triangle test? If the IBD distribution can be determined unambiguously for all n ASPs, the χ^2 goodness-of-fit test [102, 303] is the simplest and most intuitive approach to test for linkage. It compares the observed number of ASPs sharing 0, 1, and 2 alleles IBD with the expected and results in a 2 d.f. test. The hypotheses for the χ^2 goodness-of-fit test are

$$H_0: (z_0, z_1, z_2) = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) \quad \text{versus} \quad H_1: (z_0, z_1, z_2) \neq \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right).$$

Alternative tests for this null hypothesis are the *maximum LOD score (MLS)* (see Section 8.3.1.2) and *Holmans' possible triangle test*, also called *possible triangle test* or simply *triangle test* (see Section 8.3.1.3), both are based on the trinomial distribution [302, 560, 561].

8.3.1.2 What is the maximum LOD score test? If the genetic marker is fully informative, the kernel of the likelihood of all observed data is from a trinomial distribution

$$z_0^{n_0} z_1^{n_1} z_2^{n_2}, \quad (8.2)$$

where n_j is the number of ASPs sharing j alleles IBD. Note that the parameter of interest is not the recombination fraction θ but the allele sharing. With $\hat{z}_j = n_j/n$ denoting the ML estimates, i.e., the estimated proportion of sib-pairs sharing j alleles IBD, the LOD score for the likelihood ratio test based on Eq. (8.2) is given by

$$\text{MLS} = \lg \frac{\hat{z}_0^{n_0} \hat{z}_1^{n_1} \hat{z}_2^{n_2}}{\left(\frac{1}{4}\right)^{n_0} \left(\frac{1}{2}\right)^{n_1} \left(\frac{1}{4}\right)^{n_2}}. \quad (8.3)$$

MLS is an LOD score using the ML estimate from the trinomial and asymptotically χ^2 distributed with 2 d.f. It can be computed as illustrated in the following example.

Example 8.1. *The data for this and the following examples in this section are given in Table 8.2. First, we assume that all ASPs are completely informative, and we ignore*

the right part of the table. The sample of completely informative ASPs consists of $n = 38$ ASPs. Twenty-one and 13 of them have an IBD value of 2 and 1, respectively, giving $\hat{z}_2 = 21/38 \approx 0.5526$, $\hat{z}_1 = 13/38 \approx 0.3421$, and $\hat{z}_0 = 4/38 \approx 0.1053$. The MLS therefore is

$$MLS = \lg \frac{0.1053^4 0.3421^{13} 0.5526^{21}}{\left(\frac{1}{4}\right)^4 \left(\frac{1}{2}\right)^{13} \left(\frac{1}{4}\right)^{21}} \approx 3.8592.$$

The χ^2 statistic subsequently is $T = 2 \ln(10) MLS \approx 16.5290$. This corresponds to $p = 2.58 \cdot 10^{-4}$ using the χ^2 distribution with 2 d.f.

Table 8.2 Sample data for illustrating the use of the affected sib-pair (ASP) statistics.

# ASPs	$\hat{z}_{i0}$	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{w}_{i0}$	$\hat{w}_{i1}$	$\hat{w}_{i2}$
21	0	0	1	0	0	1
13	0	1	0	0	1	0
4	1	0	0	1	0	0
25	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{2}{3}$	$\frac{1}{3}$
5	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
7	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{3}$	$\frac{2}{3}$	0
6	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$\hat{z}_{ij}$ denotes the estimated probability that ASP i shares j alleles identical by descent (IBD), and $\hat{w}_{ij}$ denotes the weight of allele j from ASP i to the maximum LOD score test statistic. We assume that both parents have been genotyped. The hash symbol # denotes number.

In almost all applications, the IBD value cannot be determined unambiguously for all sib-pairs. Therefore, the MLS statistic needs an extension that allows for incompletely informative ASPs. In this case, the likelihood is the probability for the genotypes of a sib-pair given that the sib-pair is an ASP. Using the law of total probability, the likelihood of sib-pair i is $L_i = \sum_{j=0}^2 w_{ij} z_j$, where z_j is the unknown probability of an ASP sharing j alleles IBD, and w_{ij} is the probability for the marker genotypes, say $g_{\text{off},i}$ of sib-pair i given that they share j alleles IBD. The joint likelihood is the product of the likelihood of all sib-pairs,

$$L = \prod_{i=1}^n (w_{i0} \cdot z_0 + w_{i1} \cdot z_1 + w_{i2} \cdot z_2).$$

The MLS subsequently is

$$MLS = \lg \frac{\prod_{i=1}^n (\hat{w}_{i0} \cdot \hat{z}_0 + \hat{w}_{i1} \cdot \hat{z}_1 + \hat{w}_{i2} \cdot \hat{z}_2)}{\prod_{i=1}^n (\hat{w}_{i0} \cdot \frac{1}{4} + \hat{w}_{i1} \cdot \frac{1}{2} + \hat{w}_{i2} \cdot \frac{1}{4})}, \quad (8.4)$$

which is asymptotically χ^2_2 distributed.

The weights w_{ij} are constant if parental genotypes are available. For example, consider a nuclear family i where all subjects, i.e., both parents and offspring, are heterozygous 1 2. Here, $\hat{z}_{i2} = P(\text{IBD}_i = 2|g_{\text{off},i}) = \hat{z}_{i0} = P(\text{IBD}_i = 0|g_{\text{off},i}) = \frac{1}{2}$. Application of the Bayes formula yields

$$w_{ij} = P(g_{\text{off},i}|\text{IBD}_i = j) = \frac{P(\text{IBD}_i = j|g_{\text{off},i})P(g_{\text{off},i})}{P(\text{IBD}_i = j)}. \quad (8.5)$$

In our example family, the first term of the numerator on the right-hand side of Eq. (8.5) is estimated as $\frac{1}{2}$, thus identical for $\hat{w}_{i0}$ and $\hat{w}_{i2}$. The second term is independent of the IBD status, thus also equal. Finally, the denominator is $P(\text{IBD}_i = 0) = P(\text{IBD}_i = 2) = \frac{1}{4}$ for both $\hat{w}_{i0}$ and $\hat{w}_{i2}$. The estimated weights $\hat{w}_{i0}$ and $\hat{w}_{i2}$ are therefore identical for both IBD values and equal to $\frac{1}{2}$. Finally, $\hat{w}_{i1} = 0$ because $\hat{z}_{i1} = 0$. In this example, the estimated weights w_{ij} indeed represent a probability distribution. We stress, however, that only the relative values of the w_{ij} for $j = 0, 1, 2$ are important, not their absolute values.

Consider a different example and assume that one parent is homozygous 1 1, that the other parent is heterozygous 2 3, and that both offspring are heterozygous 1 3. In this case, $\hat{z}_{i1} = \hat{z}_{i2} = \frac{1}{2}$ but $\hat{w}_{i1} = 2\hat{w}_{i2}$ because the denominator of Eq. (8.5) is $\frac{1}{2}$ for $\text{IBD}_i = 2$ and $\frac{1}{4}$ for $\text{IBD}_i = 1$.

Finally, we note that for non-informative ASPs, where both parents are homozygous at the marker locus, the estimated weights are $\hat{w}_{ij} = \frac{1}{3}$ for all $j = 0, 1, 2$. Note that these non-informative ASPs contribute both to the MLS statistic of Eq. (8.4) and to the estimated IBD probabilities in the entire sample.

These sample estimates $\hat{z}_0, \hat{z}_1$, and $\hat{z}_2$ can be determined with an expectation maximization (EM) algorithm, and the update formula from step k to $k + 1$ is

$$\hat{z}_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{w}_{ij} \hat{z}_j^{(k)}}{\hat{w}_{i0} \hat{z}_0^{(k)} + \hat{w}_{i1} \hat{z}_1^{(k)} + \hat{w}_{i2} \hat{z}_2^{(k)}}. \quad (8.6)$$

Its application is illustrated in Example 8.2. Furthermore, we illustrate the computation of the MLS using all example data from Table 8.2 in Example 8.3.

Table 8.3 Estimated probabilities $\hat{w}_{ij}$ for marker genotypes given identical by descent status for 10 affected sib-pairs.

i	$\hat{w}_{i0}$	$\hat{w}_{i1}$	$\hat{w}_{i2}$	i	$\hat{w}_{i0}$	$\hat{w}_{i1}$	$\hat{w}_{i2}$
1	0	0	1	6	1	0	0
2	0	0	1	7	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
3	0	1	0	8	0	$\frac{1}{3}$	$\frac{1}{3}$
4	0	1	0	9	$\frac{1}{3}$	$\frac{1}{3}$	0
5	0	1	0	10	$\frac{1}{2}$	0	$\frac{1}{2}$

$\hat{w}_{ij}$ are used for estimating identical by descent probabilities in the sample.

Example 8.2. We consider $n = 10$ ASPs with estimated weights $\hat{w}_{ij}$ as given in Table 8.3. As starting values we choose $\hat{z}_0^{(0)} = 0.25$, $\hat{z}_1^{(0)} = 0.5$, and $\hat{z}_2^{(0)} = 0.25$. The first update gives $\hat{z}_0^{(1)} = 0.1950$, $\hat{z}_1^{(1)} = 0.5100$, and $\hat{z}_2^{(1)} = 0.2950$. Using the update formula, we proceed in estimating $\hat{z}_j^{(2)}$, which turns out to be $\hat{z}_0^{(2)} \approx 0.1763$, $\hat{z}_1^{(2)} \approx 0.5125$, and

$$\begin{aligned}\hat{z}_2^{(2)} &= \frac{1}{10} \sum_{i=1}^{10} \frac{\hat{w}_{i2} \cdot \hat{z}_2^{(1)}}{\hat{w}_{i0} \hat{z}_0^{(1)} + \hat{w}_{i1} \hat{z}_1^{(1)} + \hat{w}_{i2} \hat{z}_2^{(1)}} \\ &= \frac{1}{10} \sum_{i=1}^{10} \frac{\hat{w}_{i2} \cdot 0.1950}{\hat{w}_{i0} 0.1950 + \hat{w}_{i1} 0.5100 + \hat{w}_{i2} 0.2950} \\ &= \frac{1}{10} \cdot \left(\frac{1 \cdot 0.2950}{0 \cdot 0.1950 + 0 \cdot 0.5100 + 1 \cdot 0.2950} + \frac{1 \cdot 0.2950}{0 \cdot 0.1950 + 0 \cdot 0.5100 + 1 \cdot 0.2950} \right. \\ &\quad + \frac{0 \cdot 0.2950}{0 \cdot 0.1950 + 1 \cdot 0.5100 + 0 \cdot 0.2950} + \frac{0 \cdot 0.2950}{0 \cdot 0.1950 + 1 \cdot 0.5100 + 0 \cdot 0.2950} \\ &\quad + \frac{0 \cdot 0.2950}{0 \cdot 0.1950 + 1 \cdot 0.5100 + 0 \cdot 0.2950} + \frac{0 \cdot 0.2950}{1 \cdot 0.1950 + 0 \cdot 0.5100 + 0 \cdot 0.2950} \\ &\quad + \frac{\frac{1}{3} \cdot 0.2950}{\frac{1}{3} \cdot 0.1950 + \frac{1}{3} \cdot 0.5100 + \frac{1}{3} \cdot 0.2950} + \frac{\frac{1}{3} \cdot 0.2950}{0 \cdot 0.1950 + \frac{2}{3} \cdot 0.5100 + \frac{1}{3} \cdot 0.2950} \\ &\quad \left. + \frac{0 \cdot 0.2950}{\frac{1}{3} \cdot 0.1950 + \frac{2}{3} \cdot 0.5100 + 0 \cdot 0.2950} + \frac{\frac{1}{2} \cdot 0.2950}{\frac{1}{2} \cdot 0.1950 + 0 \cdot 0.5100 + \frac{1}{2} \cdot 0.2950} \right) \\ &\approx 0.3121.\end{aligned}$$

Six updates are required to estimate the IBD probabilities with a precision of 10^{-4} . The update steps are given in Table 8.4. The final ML IBD estimates are $\hat{z}_0 \approx 0.1640$, $\hat{z}_1 \approx 0.5137$, and $\hat{z}_2 \approx 0.3222$.

Table 8.4 Updates from the expectation maximization algorithm for estimating the identical by descent (IBD) probabilities in the sample of 10 affected sib-pairs given in Table 8.3.

Update step	IBD probability estimates		
	$\hat{z}_0$	$\hat{z}_1$	$\hat{z}_2$
0 (Initialization)	0.2500	0.5000	0.2500
1	0.1950	0.5100	0.2950
2	0.1753	0.5125	0.3121
3	0.1681	0.5133	0.3185
4	0.1654	0.5135	0.3210
5	0.1644	0.5136	0.3219
6	0.1640	0.5137	0.3222

Example 8.3. In this example, we illustrate the calculation of the MLS using all data from Table 8.2. In the first step, the EM update formula (8.6) is used to estimate $\hat{z}_j$.

The parameter estimates for the total sample of 81 ASPs are $\hat{z}_0 = 0.0684$, $\hat{z}_1 = 0.4907$, and $\hat{z}_2 = 0.4409$ (compare with Problem 8.4). The MLS for these data is

$$\begin{aligned}
 \text{MLS} = & \lg \left(\underbrace{\frac{(0 \cdot 0.0684 + 0 \cdot 0.4907 + 1 \cdot 0.4409)^{21}}{(0 \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4})^{21}}}_{21 \text{ families with IBD} = 2} \cdot \underbrace{\frac{(0 \cdot 0.0684 + 1 \cdot 0.4907 + 0 \cdot 0.4409)^{13}}{(0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4})^{13}}}_{13 \text{ families with IBD} = 1} \right. \\
 & \cdot \underbrace{\frac{(1 \cdot 0.0684 + 0 \cdot 0.4907 + 0 \cdot 0.4409)^4}{(1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4})^4}}_{4 \text{ families with IBD} = 0} \cdot \underbrace{\frac{(0 \cdot 0.0684 + \frac{2}{3} \cdot 0.4907 + \frac{1}{3} \cdot 0.4409)^{25}}{(0 \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{4})^{25}}}_{25 \text{ families with IBD} = 1 \text{ or } 2} \\
 & \cdot \underbrace{\frac{(\frac{1}{3} \cdot 0.0684 + \frac{2}{3} \cdot 0.4907 + 0 \cdot 0.4409)^7}{(\frac{1}{3} \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{2} + 0 \cdot \frac{1}{4})^7}}_{7 \text{ families with IBD} = 0 \text{ or } 1} \cdot \underbrace{\frac{(\frac{1}{2} \cdot 0.0684 + 0 \cdot 0.4907 + \frac{1}{2} \cdot 0.4409)^5}{(\frac{1}{2} \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{4})^5}}_{5 \text{ families with IBD} = 0 \text{ or } 2} \\
 & \cdot \underbrace{\frac{(\frac{1}{3} \cdot 0.0684 + \frac{1}{3} \cdot 0.4907 + \frac{1}{3} \cdot 0.4409)^6}{(\frac{1}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{4})^6}}_{6 \text{ non informative families}} \Big) \\
 \approx & 3.7283.
 \end{aligned}$$

Thus, the LOD score decreased from 3.86 to 3.73 in this example by adding the partly informative families compared to the use of the completely informative families from Example 8.1. The LOD score of 3.73 corresponds to $T \approx 17.1694$, giving $p = 1.87 \cdot 10^{-4}$ by using the χ^2_2 distribution. The question is why the LOD score dropped although families were added to the entire sample. This can be easily explained by looking at the IBD estimates. While the estimated probability for an ASP to share 2 alleles IBD was approximately 55% in the completely informative families, this fraction decreased to 44% in the entire sample. We finally note that the completely non-informative families are omitted from calculations in some of the freely available software packages.

8.3.1.3 What is the triangle test? The power of the MLS test can be increased by restricting the maximization of the IBD probabilities in the sample $\hat{z}_j$ to a set of inequalities. Specifically, it has shown that the set of IBD probabilities corresponding to a genetic model satisfies the inequalities (see Problem 8.5; see Refs. [201, 302])

$$z_1 \leq \frac{1}{2} \quad 2 \cdot z_0 \leq z_1 \quad z_0 \geq 0. \quad (8.7)$$

These restrictions form a triangle as displayed in Figure 8.3. Holmans [302] has proposed an LR test with the IBD estimates restricted to this possible triangle. Unfortunately, the null hypothesis is represented by a point on the boundary of the possible triangle, i.e., the restricted parameter space, and thus standard asymptotic theory cannot be applied. Instead, the ML theory for non-standard conditions has

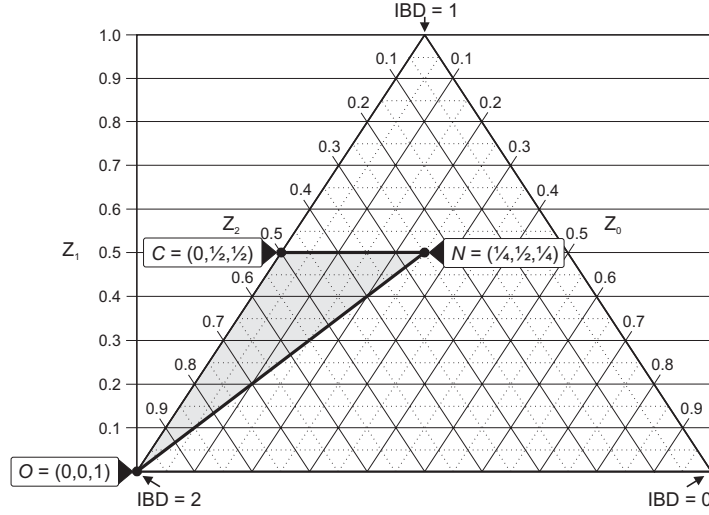


Fig. 8.3 The space of identity by descent (IBD) distributions displayed as de Finetti-like diagram and the possible triangle for affected sib-pairs. Each point in the triangle corresponds to a specific IBD distribution. The possible triangle is the gray area. The points N , C , and O will be used in Section 8.4.

to be used [603]. It appears that the *triangle test statistic* (TTS) is asymptotically distributed as a mixture of χ^2 distributions with $(\frac{1}{2} - c)\chi_0^2 + \frac{1}{2}\chi_1^2 + c\chi_2^2$, where χ_0^2 denotes a point mass at 0 [321]. The weight c of this mixture depends on the degree of informativity, i.e., the number of marker alleles, and on whether parents have been genotyped [359]. Therefore, c varies with the data set, and it is difficult to determine its value in general. To derive a general recommendation that can be used in practice, Holmans [302] conducted an extensive Monte-Carlo simulation study. He found that for $\alpha = 0.001$ when parents are not genotyped, the criterion ranges from 2.23 to 2.32 for 2 to 1000 marker alleles; when parents are genotyped, the criterion ranges from 2.29 to 2.32. Hence, using a general LOD score cutoff of 2.3 is reasonable for $\alpha = 0.001$. For $\alpha = 0.0001$, an LOD score of 3.25 has been shown to be a sensible compromise.

As a conservative approximation, i.e., assuming complete informativity, one may use the following approach for calculating p -values given the TTS LOD score [359]:

$$\frac{1}{2}P(\chi_1^2 > 2 \ln(10) TTS) + \frac{\arccos \sqrt{2/3}}{2\pi} P(\chi_2^2 > 2 \ln(10) TTS)$$

It is important to note that the restrictions have been obtained by considering genetic effects only. However, in practice, environmental effects might move the alternative IBD probabilities outside Holmans' triangle [261]. Therefore, it has been argued that an ASP test statistic should be able to detect any deviation from the null hypothesis $H_0: (z_0, z_1, z_2) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, and the alternative hypothesis should not be restricted to $H_1: (z_0, z_1, z_2) \neq (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, subject to the restrictions of Eq. (8.7).

We illustrate the use of the TTS in a simple example.

Example 8.4. Suppose that 10 ASPs share zero, 55 ASPs share one, and 35 ASPs share two marker alleles IBD in a sample of $n = 100$ ASPs. The unrestricted ML estimates for the IBD probabilities are $\hat{z}_0 = 10/100$, $\hat{z}_1 = 55/100$, and $\hat{z}_2 = 35/100$, which do not satisfy the possible triangle constraints as $\hat{z}_1 > 0.5$. The restricted ML IBD estimates turn out to be $\tilde{z}_0 \approx 0.1111$, $\tilde{z}_1 = 0.5$, and $\tilde{z}_2 \approx 0.3889$. Subsequently, the TTS is given by

$$TTS \approx \lg \frac{0.1111^{10} 0.5^{55} 0.3889^{35}}{\left(\frac{1}{4}\right)^{10} \left(\frac{1}{2}\right)^{55} \left(\frac{1}{4}\right)^{35}} \approx 3.1942.$$

The asymptotic p -value corresponding to this LOD score is

$$\frac{1}{2}P(\chi_1^2 > 2 \ln(10) \cdot 3.1942) + \frac{\arccos \sqrt{2/3}}{2\pi} P(\chi_2^2 > 2 \ln(10) \cdot 3.1942) \approx 1.25 \cdot 10^{-4}.$$

8.3.1.4 What is the maximum LOD score test with no dominance variance?

Both the MLS and the TTS are two-parameter tests. It has been shown, however, that one-parameter allele-sharing tests are more powerful for most genetic models [719]. In Section 8.2, it was shown that $z_1 = \frac{1}{2}$ for some genetic models. This IBD probability corresponds to the dominance variance (see Section 8.5) of a genetic model. If this dominance variance is neglected, the likelihood can be maximized subject to the restrictions $z_0 \leq \frac{1}{4}$ and $z_1 = \frac{1}{2}$. In fact, only one parameter is being maximized and the resulting test is a one-parameter test. The “free” parameter is maximized along a half-line, and the resulting *MLS test with no dominance variance* is asymptotically distributed as a 50:50 mixture of χ^2 distributions with 0 and 1 d.f. It is preferable to the standard MLS without restrictions (for details, see Refs. [302, 320, 433]).

We illustrate the use of the MLS with two simple examples. Another extensive illustration is provided at the end of Section 8.3.2 in Example 8.7.

Example 8.5. In the first example, we use the data from Example 8.4. The restricted IBD estimates with the assumption of no dominance variance are identical to those of Example 8.4 because $\hat{z}_1$ assuming the dominance variance is already 0.5.

For the second example, suppose that 20 ASPs share zero, 30 ASPs share one, and 50 ASPs share two marker alleles IBD in a total sample of $n = 100$ ASPs. The unrestricted ML estimates for the IBD probabilities are $\hat{z}_0 = 20/100$, $\hat{z}_1 = 30/100$, and $\hat{z}_2 = 50/100$, which do not satisfy the possible triangle constraints as $2 \cdot z_0 > z_1$. The restricted ML IBD estimates assuming the presence of dominance variance turn out to be $\tilde{z}_0 \approx 0.1667$, $\tilde{z}_1 \approx 0.3333$, and $\tilde{z}_2 = 0.5$. Assuming no dominance variance, the estimates differ, however. They are given by $\tilde{z}_0 \approx 0.1429$, $\tilde{z}_1 = 0.5$, and $\tilde{z}_2 \approx 0.3571$, giving $MLS \approx 2.9907$ and $p \approx P(\chi_1^2 > 2 \ln(10) \cdot 2.9907) \approx 1.23 \cdot 10^{-4}$.

8.3.2 What are common score- or Wald-type 1 degree of freedom tests?

Linkage tests for ASPs involving 1 d.f. are generally more powerful than those with 2 d.f., and various tests have been proposed. All 1 d.f. tests of score- or Wald-type may be represented as weighted sums of the estimated IBD, i.e.,

$$w_0 \hat{z}_0 + w_1 \hat{z}_1 + w_2 \hat{z}_2. \quad (8.8)$$

As before, $\hat{z}_j = \frac{1}{n} \sum_{i=1}^n \hat{z}_{ij}$ is the estimated proportion of ASPs sharing j alleles IBD, but w_j is now independent of i and represents weights for the total sample. The weighted sum in Eq. (8.8) converges to a univariate normal distribution. Because the square of a standard normal is a χ_1^2 distribution, all tests based on Eq. (8.8) are termed 1 d.f. tests (see Problem 8.6).

The test statistic for a 1 d.f. test is invariant under linear transformations of the weights w_j [719] and can therefore be rewritten in a more convenient form that will be helpful in the following example. The weights can be standardized arbitrarily so that $w_0 = 0$ and $w_2 = 1$ can be chosen. In fact, $\hat{z}_0 = 1 - \hat{z}_1 - \hat{z}_2$, and thus $w_0 = 0$ is a reasonable choice. Furthermore, with $w_2 = 1$, the test is completely determined by the weight w_1 assigned to $\hat{z}_1$, and the non-standardized test statistic is

$$T = \hat{z}_2 + w_1 \hat{z}_1 = \frac{1}{n} \sum_{i=1}^n (\hat{z}_{i2} + w_1 \hat{z}_{i1}) = \frac{1}{n} \sum_{i=1}^n T_i. \quad (8.9)$$

Mean and variance of T are given by

$$\mathbb{E}(T) = z_2 + w_1 z_1 \quad \text{and} \quad \text{Var}(T) = \frac{1}{n} \left(z_2(1 - z_2) - 2w_1 z_1 z_2 + w_1^2 z_1(1 - z_1) \right).$$

Under the null hypothesis of no linkage $(z_0, z_1, z_2) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ so that $\mathbb{E}_{H_0}(T) = \frac{1}{4} + \frac{1}{2}w_1$. Similarly, $\text{Var}_{H_0}(T) = \frac{1}{4n}(\frac{3}{4} - w_1 + w_1^2)$. For ASPs, one expects increased allele sharing under the alternative hypothesis, and the hypotheses for this one-sided test problem are

$$H_0: \mathbb{E}(T) = \frac{1}{4} + \frac{1}{2}w_1 \quad \text{versus} \quad H_1: \mathbb{E}(T) > \frac{1}{4} + \frac{1}{2}w_1.$$

This score and Wald test versions of the 1 d.f. tests are therefore given by

$$T_S = \frac{T - \mathbb{E}_{H_0}(T)}{\sqrt{\text{Var}_{H_0}(T)}} \quad \text{and} \quad T_W = \frac{T - \mathbb{E}_{H_0}(T)}{\sqrt{\widehat{\text{Var}}(T)}}. \quad (8.10)$$

Here, $\widehat{\text{Var}}(T) = \frac{1}{n(n-1)} \sum_{i=1}^n (T_i - T)^2$ is the natural variance estimator for $\text{Var}(T)$, with T_i as defined in Eq. (8.9).

Both T_S and T_W are asymptotically standard normally distributed, and the t distribution with $n - 1$ d.f. is a good choice for applications to get a better approximation to an asymptotic distribution. In standard statistics, $n = 30$ or $n = 40$ are usual rules

of thumb for sufficient correspondence of the asymptotic and the exact distribution of the underlying test statistic. The comparisons between these two distributions are, however, often made for significance levels of 5%. In linkage studies, p -values of interest have to be lower than 0.001 (see Section 7.3). An alternative for obtaining p -values are therefore Monte–Carlo simulation methods, which will be discussed in Chapter 9.

The test statistics of Eq. (8.10) are still formulated in a general way as they contain the weight w_1 for z_1 . Three choices for the weight w_1 are common in applications and will be considered in detail. These are

1. $w_1 = \frac{1}{2} = 0.5 \quad \Rightarrow \quad \text{mean test}$
2. $w_1 = 0 \quad \Rightarrow \quad \text{proportion test}$
3. $w_1 = \frac{11}{40} = 0.275 \quad \Rightarrow \quad \text{minimax test}$

Ad 1. The most common choice for w_1 in Eq. (8.9) is $\frac{1}{2}$, yielding the *mean test*. In general, $\bar{\tau}$ and $\hat{\tau}_i$ are used to denote T and T_i , respectively, for the mean test. Since $\mathbb{E}(2\hat{z}_{i2} + 1\hat{z}_{i1}) = 2z_2 + z_1$ is the expected number of alleles shared IBD by ASP i ,

$$\tau_i = z_{i2} + \frac{1}{2}z_{i1} = \mathbb{E}(\hat{z}_{i2} + \frac{1}{2}\hat{z}_{i1}) = \frac{1}{2}\mathbb{E}(2\hat{z}_{i2} + 1\hat{z}_{i1})$$

is the proportion of alleles shared IBD by ASP i . Therefore, $\bar{\tau}$ is the estimated mean proportion of alleles shared IBD. It can easily be shown that the probability $\tau = \mathbb{E}(\bar{\tau})$ may also be interpreted as the probability that an allele is IBD.

For the mean test, one specifically obtains $\mathbb{E}_{H_0}(\bar{\tau}) = \frac{1}{2}$ and $\mathbb{V}_{H_0}(\bar{\tau}) = \frac{1}{8n}$ as mean and variance under the null hypothesis of no linkage. Furthermore, the variance estimator of $\bar{\tau}$ is $\widehat{\mathbb{V}\text{ar}}(\bar{\tau}) = \frac{1}{n(n-1)} \sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})^2$. Therefore, the two versions of the mean test are given by

$$T_{\text{S-M}} = \frac{\bar{\tau} - \frac{1}{2}}{\sqrt{\frac{1}{8n}}} \quad \text{and} \quad T_{\text{W-M}} = \frac{\bar{\tau} - \frac{1}{2}}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})^2}}. \quad (8.11)$$

An important question is: Which of the two mean test statistics should be preferred in applications? The solution is obtained easily by a comparison of the variances. τ_i is a probability so that $\mathbb{V}\text{ar}(\hat{\tau}_i) = \tau_i(1 - \tau_i)$. $\mathbb{V}\text{ar}(\hat{\tau}_i)$ takes its largest value at $\tau_i = \frac{1}{2}$, thus $T_{\text{W-M}} \geq T_{\text{S-M}}$. The calculation of the mean test is illustrated in Example 8.6 at the end of this section.

Ad 2. A second common choice for w_1 is $w_1 = 0$, leading to the *proportion test*. It tests whether the proportion of families sharing two alleles IBD equals $\frac{1}{4}$ or whether this sharing is increased. Mean and variance under H_0 are given

by $\mathbb{E}_{H_0}(\hat{z}_2) = \frac{1}{4}$ and $\mathbb{V}ar_{H_0}(\hat{z}_2) = \frac{3}{16n}$, respectively. With $\widehat{\mathbb{V}ar}(\hat{z}_2) = \frac{1}{n(n-1)} \sum_{i=1}^n (\hat{z}_{i2} - \hat{z}_2)^2$, the two versions of the proportion test are given by

$$T_{S-P} = \frac{\hat{z}_2 - \frac{1}{4}}{\sqrt{\frac{3}{16n}}} \quad \text{and} \quad T_{W-P} = \frac{\hat{z}_2 - \frac{1}{4}}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (\hat{z}_{i2} - \hat{z}_2)^2}}.$$

The computation of the proportion test statistics is illustrated in Example 8.6.

Ad 3. One choice for w_1 in Eq. (8.9) is $w_1 = 0.275 = \frac{11}{40}$. It differs slightly from the midpoint 0.25 of the weights $w_1 = 0.5$ used by the mean test and $w_1 = 0$ used by the proportion test. The weight $w_1 = 0.275$ has been shown to yield a minimax test [719] and performs well regardless of the underlying true IBD probabilities (see Section 8.4). In the following example, we use the symbols $\hat{\tau}^*$ and $\hat{\tau}_i^*$ instead of T and T_i , respectively, for the minimax test.

Under the null hypothesis of no linkage, mean and variance are given by $\mathbb{E}_{H_0}(\hat{\tau}^*) = 0.25 + 0.275 \cdot 0.5 = 0.3875 = \frac{31}{80}$ and $\mathbb{V}ar_{H_0}(\hat{\tau}^*) = \frac{881}{6400n}$. With $\widehat{\mathbb{V}ar}(\hat{\tau}^*) = \frac{1}{n(n-1)} \sum_{i=1}^n (\hat{\tau}_i^* - \hat{\tau}^*)^2$, the two versions of the minimax test are given by

$$T_{S-mm} = \frac{\bar{\hat{\tau}^*} - \frac{31}{80}}{\sqrt{\frac{881}{6400n}}} \quad \text{and} \quad T_{W-mm} = \frac{\bar{\hat{\tau}^*} - \frac{31}{80}}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (\hat{\tau}_i^* - \hat{\tau}^*)^2}}.$$

The computation of the minimax test statistics is illustrated in Example 8.6.

Although the score- and Wald-type test statistics are widely used, they have been criticized for their reduced power when marker informativity is incomplete. However, with the advent of the single nucleotide polymorphism (SNP) chip technology (Chapter 3), this problem can be easily overcome in applications.

Example 8.6. Table 8.5 gives the necessary extension of the data from Table 8.2 for calculating the mean test, the proportion test, and the minimax test statistics. For the mean test, one obtains

$$\begin{aligned} \bar{\hat{\tau}} &= \frac{1}{n} \sum_{i=1}^n \hat{\tau}_i = \frac{1}{81} (21 \cdot 1 + 13 \cdot \frac{1}{2} + \dots + 6 \cdot \frac{1}{2}) = \frac{53.5}{81} \approx 0.6605, \\ \widehat{\mathbb{V}ar}(\hat{\tau}) &= \frac{1}{n(n-1)} \sum_{i=1}^n (\hat{\tau}_i - \bar{\hat{\tau}})^2 \approx \frac{1}{81 \cdot 80} (21 \cdot 0.1153 + \dots + 6 \cdot 0.0258) \approx 9.51 \cdot 10^{-4}. \end{aligned}$$

It is easily seen that $\widehat{\mathbb{V}ar}(\hat{\tau}) \approx 9.51 \cdot 10^{-4} \ll \mathbb{V}ar_{H_0}(\hat{\tau}) = \frac{1}{8 \cdot 81} \approx 1.54 \cdot 10^{-3}$. Because the numerators of T_{W-M} and T_{S-M} are identical, T_{W-M} is approximately 1.27 larger than T_{S-M} in this example. The values of the test statistics and the asymptotic p -values using the t_{n-1} distribution are

$$\begin{aligned} T_{W-M} &\approx (0.6605 - 0.5) / \sqrt{9.5 \cdot 10^{-4}} \approx 5.2039 \quad \text{with} \quad p_{W-M} \approx 7.37 \cdot 10^{-7}, \\ T_{S-M} &\approx (0.6605 - 0.5) / \sqrt{1.5 \cdot 10^{-3}} \approx 4.0855 \quad \text{with} \quad p_{S-M} \approx 5.19 \cdot 10^{-5}. \end{aligned}$$

Table 8.5 Data and observed proportion of alleles shared identical by descent (IBD) for the affected sib-pair (ASP) statistics.

# ASPs	$\hat{z}_{i0}$	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{\tau}_i$	$\hat{\tau}_i^*$	# ASPs	$\hat{z}_{i0}$	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{\tau}_i$	$\hat{\tau}_i^*$
21	0	0	1	1	1	25	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{51}{80}$
13	0	1	0	$\frac{1}{2}$	$\frac{11}{40}$	5	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
4	1	0	0	0	0	7	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{4}$	$\frac{11}{80}$
						6	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{31}{80}$

ASPs denotes the number of ASPs with the specific IBD configuration. $\hat{z}_{ij}$ is the estimated probability that ASP i shares j alleles IBD, $\hat{\tau}_i$ is the estimated proportion of alleles shared IBD, and $\hat{\tau}_i^*$ is the corresponding contribution of an ASP to the non-standardized minimax test statistic.

The corresponding LOD scores are $LOD_{W-M} = T_{W-M}^2 / (2 \ln(10)) \approx 5.8805$ and $LOD_{S-M} \approx 3.6245$.

Similarly, one calculates for the proportion test $\hat{z}_2 \approx 0.4815$, $\widehat{\text{Var}}(\hat{z}_2) \approx 0.0017$, and $\text{Var}_{H_0}(z_2) \approx 0.0023$, yielding

$$T_{W-P} \approx (0.4815 - 0.25) / \sqrt{0.0017} \approx 5.5624 \quad \text{with} \quad p_{W-P} \approx 1.71 \cdot 10^{-7},$$

$$T_{S-P} \approx (0.4815 - 0.25) / \sqrt{0.0023} \approx 4.8113 \quad \text{with} \quad p_{S-P} \approx 3.47 \cdot 10^{-6}.$$

These correspond to $LOD_{W-P} \approx 6.7187$ and $LOD_{S-P} \approx 5.0266$.

Finally, one computes for the minimax test $\hat{\tau}^* \approx 0.5716$, $\widehat{\text{Var}}(\hat{\tau}^*) \approx 0.0012$, and $\text{Var}_{H_0}(\hat{\tau}^*) \approx 0.0017$, yielding

$$T_{W-mm} \approx (0.5716 - 0.3875) / \sqrt{0.0012} \approx 5.2525 \quad \text{with} \quad p_{W-mm} \approx 6.06 \cdot 10^{-7},$$

$$T_{S-mm} \approx (0.5716 - 0.3875) / \sqrt{0.0017} \approx 4.4659 \quad \text{with} \quad p_{S-mm} \approx 1.29 \cdot 10^{-5}.$$

These correspond to $LOD_{W-mm} \approx 5.9907$ and $LOD_{S-mm} \approx 4.3309$.

Example 8.7. We illustrate the use of the MLS statistics and the score and Wald statistics of Sections 8.3.1 and 8.3.2 by reanalyzing the data of Mein and colleagues [457]. We gratefully acknowledge the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory for providing the genotype data (from Ref. [457], data version 1.0) for independent analysis by us. On the basis of a genome-wide scan that included a total of 356 independent ASPs from the United Kingdom, Mein and colleagues reported a susceptibility locus on chromosome 16q22–16q24 (D16S515–D16S520) for type I diabetes (see Figure 1b in Ref. [457]). Four markers were additionally typed in this region showing strong evidence of linkage to increase information content which varied between 65 and 90% for chromosome 16.

Figure 8.4 displays the MLS scores for different curves for 5 different test statistics. It is available in GENEHUNTER. The MLS is given with both the triangle restriction and the restriction of no dominance variance. The graphs for these two statistics

are identical for this chromosome because $\hat{z}_1$ has already been restricted to 0.5. In fact, the unrestricted ML estimator is $(\hat{z}_0, \hat{z}_1, \hat{z}_2) \approx (0.13, 0.58, 0.29)$, which does not fulfill the triangle restriction. This estimator is above the line C–N of Figure 8.3. For both the TTS and the MLS without dominance variance, the estimator is therefore restricted to this line.

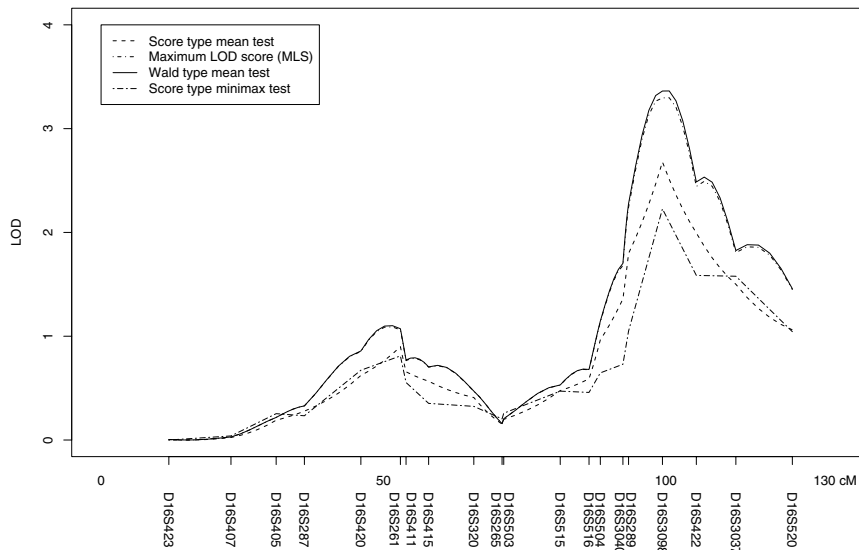


Fig. 8.4 Multipoint linkage analysis of chromosome 16 for the type I diabetes data in Ref. [457]. Displayed are multipoint analyses of the maximum *LOD* score (MLS) statistic with the triangle restriction, score- and Wald-type mean test statistic, and the score-type minimax statistic.

The figure also shows that the score type of the mean test yields a larger *LOD* score than the minimax test. This observation can be expected by taking into account the results on optimality of ASP test statistics, which will be discussed in Section 8.4. The score-type version of the minimax test is implemented in the module LODPAL from the S.A.G.E. package.

Furthermore, we can see that the Wald-type of the mean test leads to substantially larger *LOD* scores than the score-type version of the mean test. The Wald-type of the mean test is available, e.g., in S.A.G.E..

All statistics clearly show evidence for linkage of type I diabetes to chromosome 16. The maximum *LOD* score is at D16S3098 for all methods. Although the general shapes of the *LOD* score curves are similar for all displayed test statistics, there are pronounced differences in the magnitude of the *LOD* scores. In addition, because of the different limiting distributions, the asymptotic *p*-values also vary considerably.

8.3.3 What is the current status for affected sib-pair tests using alleles shared identical by state?

Instead of using the IBD information of ASPs, one could base tests on IBS values [397]. In general, these methods are less powerful than IBD methods. Furthermore, they depend on the marker informativity even if parental data are available and IBD values can be determined uniquely [61]. Therefore, they are not applied anymore.

8.4 IS THERE AN OPTIMAL AFFECTED SIB-PAIR TEST? ARE AFFECTED SIB-PAIR TESTS RELATED TO MODEL-BASED LINKAGE TESTS?

The first large study for power comparisons of ASP statistics was conducted by Blackwelder and Elston [64]. Assuming a single disease locus model, they compared three different score test statistics: the proportion test, the mean test, and the χ^2 goodness-of-fit test. Their conclusion was that although the most powerful test depends on the true state of nature—which includes the mode of inheritance, the disease allele frequency, and the recombination fraction between the genetic marker under study and the disease locus—the mean test is generally more powerful than the proportion and the goodness-of-fit tests. In fact, Knapp et al. [365] have proven for a diallelic single trait locus model that the mean test is the uniformly most powerful test in the recombination fraction θ if

$$(f_1^2 - f_0 f_2) \cdot (2K(f_0 - 2f_1 + f_2) + (f_1^2 - f_0 f_2)) = 0,$$

where $K = p^2 f_2 + 2pqf_1 + q^2 f_0$ is the population prevalence of the disease for a diallelic locus with disease allele frequency p and $q = 1 - p$, and f_j are the penetrances (see Chapter 2). This condition is satisfied for any multiplicative genetic model and a simple recessive genetic model. Furthermore, the mean test is locally most powerful in θ irrespective of the mode of inheritance [365]. Anticipating a later result, the score-type version of the mean test is equivalent to the non-parametric linkage (NPL; see Section 8.8) score in the case of independent ASPs. Furthermore, the maximum likelihood binomial method [3] reduces to the mean test when there is one ASP per family.

Instead of basing optimality on genetic models, optimality can be defined in terms of allele sharing. Whittemore and Tu [719] have considered the class of score tests T_S (Eq. (8.10)). They have shown that $0 \leq w_1 \leq \frac{1}{2}$ in order to be the efficient score statistic of the trinomial likelihood from Eq. (8.2). As a consequence, a 1 d.f. test based on T_S with weight w_1 is optimal if the true allele sharing lies on a straight line with points $N = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ and $(0, w_1, 1 - w_1)$ with $0 \leq w_1 \leq \frac{1}{2}$ in the de Finetti triangle of IBD distributions (see Figure 8.3). Specifically, the proportion test is optimal when the true allele sharing is on the line $N-O$, and the mean test is optimal when the true allele sharing is on the line $N-C$. A test should perform well when the true allele sharing is near the straight line. Because the mean test and

proportion test are extreme choices in the class of optimal models, their performance can be unfavorable if the true allele sharing is close to the other family of models. Therefore, it is plausible to construct a more robust 1 d.f. test by choosing a point that is almost midway between the extreme lines for the mean test and the proportion test. This means that w_1 should be close to 0.25. The optimal test is T_S having the lowest maximal loss in asymptotic power among the possible allele sharings subject to $0 \leq w_1 \leq \frac{1}{2}$. This optimal test is termed the minimax test, and Whittemore and Tu [719] have shown that the optimal choice is $w_1 = 0.275$.

In this section, it has been shown that some model-free methods are weakly model-based in the sense that they have maximal power under specific genetic models. As already mentioned in Section 8.1, some model-free methods are indeed equivalent to certain model-based approaches. Knapp et al. [366] have shown for ASPs that the mean test is equivalent to an LOD score test with a simple recessive genetic model if the phenotype of the parents is unknown or if both parents are unaffected. Furthermore, the proportion test is equivalent to an LOD score test with a simple dominant genetic model if both parents are unaffected and if both the disease allele and the significance level are small. Equivalences have also been shown between particular forms of LOD scores and some forms of the MLS [320, 563].

8.5 HOW CAN SAMPLE SIZE OR POWER BE CALCULATED FOR AN AFFECTED SIB-PAIR STUDY?

Once the IBD allele-sharing probabilities are known, sample size and power calculations can be done with standard methods, as will be illustrated below for the score test versions of the 1 d.f. tests. The question is, however, how the parameters of a genetic model are related to the IBD probabilities. In the literature, two different approaches have been proposed for connecting genetic models and IBD probabilities. The first approach traces back to Suarez [649], who expressed the IBD probabilities in terms of the population prevalence of the disease and the additive variance and dominance variance (these terms will be defined below) of a single autosomal major gene locus. The Suarez method is not considered here because it is rarely used in applications. Instead, we discuss in more detail an approach based on recurrence risk ratios which have been introduced in Chapter 6.

8.5.1 How are identical by descent probabilities and recurrence risk ratios functionally related?

Before we can establish the functional relationship between locus-specific recurrence risk ratios and IBD probabilities, some notation is required. For a diallelic single-locus model, the variability between biological relatives that is caused by the trait locus is given by

$$\sigma_a^2 = 2pq(p f_2 + (1 - 2p)f_1 + q f_0)^2, \quad \sigma_d^2 = p^2 q^2 (f_2 - 2f_1 + f_0)^2. \quad (8.12)$$

More specifically, σ_a^2 is the *additive variance* of the trait locus that is due to average differences between carriers of alleles at the trait locus. The *dominance variance* σ_d^2 refers to the deviation in heterozygotes that are not always exact intermediates between homozygotes. If $\sigma_d^2 = 0$, the resulting model is called an *additive single-locus model*, *additive model* for short, or a model with *no dominance variance* (see Section 6.3.1).

On the basis of a short but seminal paper by James [338], Risch [561, 562] showed that the locus-specific recurrence risk ratios can be written as functions of the kinship coefficient c_R , Jacquard's $\Delta_{7,R}$ coefficient, the population prevalence K , and the additive and dominance variances σ_a^2 and σ_d^2 :

$$\lambda_R - 1 = \frac{1}{K^2} (2 \cdot c_R \sigma_a^2 + \Delta_{7,R} \sigma_d^2), \quad (8.13)$$

if there is no linkage between different trait loci. c and Δ_7 are discussed in detail in Section 6.3.3.

With these notations, the locus-specific recurrence risk ratios at the trait locus are

$$\lambda_M - 1 = \frac{1}{K^2} (\sigma_a^2 + \sigma_d^2), \quad \lambda_S - 1 = \frac{1}{K^2} (\frac{1}{2} \sigma_a^2 + \frac{1}{4} \sigma_d^2), \quad \lambda_O - 1 = \frac{1}{K^2} \frac{1}{2} \sigma_a^2. \quad (8.14)$$

Using elementary algebra for Eq. (8.14), one can easily see that $\lambda_M = 4\lambda_S - 2\lambda_O - 1$. Furthermore, $\lambda_S = \lambda_O$ if $\sigma_d^2 = 0$.

Equation (8.14) shows that the λ 's are functions of K , σ_a^2 , and σ_d^2 . These in turn are functions of the parameters p , f_2 , f_1 , and f_0 of the underlying genetic model. The genetic model also determines the IBD probabilities. It is therefore possible to rewrite λ_M , λ_S , and λ_O as functions of the IBD probabilities $z_{t,j}$ at the trait locus t

$$z_{t,2} = \frac{1}{4} + \frac{\lambda_M - \lambda_S}{4\lambda_S} = \frac{1}{4} + \frac{3\lambda_S - 2\lambda_O - 1}{4\lambda_S}, \quad z_{t,1} = \frac{1}{2} - \frac{\lambda_S - \lambda_O}{2\lambda_S}, \quad (8.15)$$

and $z_{t,0} = 1 - z_{t,1} - z_{t,2}$. The IBD probabilities $z_{m,j}$ at a marker locus m can be obtained from the IBD probabilities at the trait locus $z_{t,l}$ via

$$z_{m,j} = P(\text{IBD}_m = j) = \sum_{l=0}^2 P(\text{IBD}_m = j | \text{IBD}_t = l) P(\text{IBD}_t = l). \quad (8.16)$$

$P(\text{IBD}_t = l)$ have been stated in Eq. (8.15), and $P(\text{IBD}_m = j | \text{IBD}_t = l)$ are given in Table 8.6 for a recombination fraction θ between t and m . The elements in the table can be derived as follows. If the ASP shares zero alleles IBD at the trait locus, the ASP shares zero alleles IBD at the marker locus only if the ASP has either no or two recombinations for each parent, yielding $\Psi = \theta^2 + (1 - \theta)^2$ for a single parent. The parental meioses are assumed to be independent, giving Ψ^2 as the upper left probability in the table. The same holds for the lower right probability in the table.

If the ASP has zero alleles at the trait locus but two alleles at the marker locus, exactly one recombination per ASP and parent is required. This occurs with probability $1 - \Psi$. As before, meioses in both parents are assumed to be independent. Therefore, the entry in the lower left corner of the table is $(1 - \Psi)^2$. The upper right entry in the

Table 8.6 Identical by descent (IBD) probabilities $z_{m,j}$ at a marker locus given the IBD status at the trait locus. The recombination fraction between the trait locus t and the marker locus m is θ , and $\Psi = \theta^2 + (1 - \theta)^2$.

IBD status at marker locus m	IBD status at trait locus t		
	0	1	2
0	Ψ^2	$\Psi(1 - \Psi)$	$(1 - \Psi)^2$
1	$2\Psi(1 - \Psi)$	$\Psi^2 + (1 - \Psi)^2$	$2\Psi(1 - \Psi)$
2	$(1 - \Psi)^2$	$\Psi(1 - \Psi)$	Ψ^2
Total	1	1	1

table can be obtained analogously. The probabilities are conditional; thus, columns in the table sum to 1, and the first and the last column are already complete.

Because the probabilities in the first and last rows of the middle column are identical due to symmetry, we finally have to determine the central entry of the table. If the IBD value is 1 at the trait locus, it is also 1 at the marker locus if the ASP has either no or two recombinations per parent, yielding Ψ^2 for both parents. Alternatively, exactly one recombination may occur in the two meioses per parent. This event has a probability of $1 - \Psi$ per parent, giving $(1 - \Psi)^2$ in total for both parents.

With a little algebra, using Eq. (8.16) and Table 8.6 we obtain

$$\begin{aligned}
 z_{m,2} &= \frac{1}{4} + \frac{1}{4\lambda_S} (2\Psi - 1) ((\lambda_S - 1) + 2\Psi(\lambda_S - \lambda_O)), \\
 z_{m,1} &= \frac{1}{2} - \frac{1}{2\lambda_S} (2\Psi - 1)^2 (\lambda_S - \lambda_O), \\
 z_{m,0} &= \frac{1}{4} - \frac{1}{4\lambda_S} (2\Psi - 1) ((\lambda_S - 1) + 2(1 - \Psi)(\lambda_S - \lambda_O)).
 \end{aligned} \tag{8.17}$$

A simplification of Eq. (8.17) is obtained for a model without dominance variance. In this case, $\lambda_S = \lambda_O$ and subsequently

$$z_{m,2} = \frac{1}{4} + (2\Psi - 1) \frac{\lambda_S - 1}{4\lambda_S} \quad \text{and} \quad z_{m,1} = \frac{1}{2}. \tag{8.18}$$

8.5.2 How are sample size and power calculated for the mean test using recurrence risk ratios?

The values $z_{m,j}$ of Eqs. (8.17) and (8.18) form the basis of sample size and power calculations using the MLS and the score test versions of the 1 d.f. tests. This will be illustrated in this section with the mean test of Eq. (8.11) as an example. Calculations are straightforward under the assumption of complete marker informativity as

measured by the polymorphism information content (PIC) value (see Chapter 3). In practice, however, informativity is almost complete in genome-wide searches only if SNP chip scans are conducted, while it is generally incomplete for microsatellite scans. Therefore, we will discuss an intuitive solution for sample size and power calculations in the presence of incomplete informativity in the latter part of this section.

The null hypothesis H_0 of no linkage is rejected at a significance level α if $P_{H_0}(T_{S-M} = \sqrt{8n}(\hat{\tau} - \frac{1}{2}) > z_{1-\alpha})$, where P_{H_0} denotes the probability distribution under H_0 . The *power* of the test is the probability that the test rejects H_0 in favor of H_1 if H_1 indeed holds true. The expected value of the test statistic $\hat{\tau}$ under H_1 is given by τ_m . Its variance can be calculated in two different ways. Firstly, one can use the results from Section 8.3.2, yielding

$$\mathbb{V}ar_{H_1}(\hat{\tau}) = \frac{1}{n} \left(z_{m,2}(1 - z_{m,2}) - z_{m,1}z_{m,2} + z_{m,1}(1 - z_{m,1})/4 \right).$$

Alternatively, and more easily, τ_m is interpretable as the probability that an allele is IBD (see Section 8.3.2). Its variance is $\tau_m(1 - \tau_m)$. In a sample of n ASP, there are $2n$ meioses; therefore, the variance of $\hat{\tau}$ under H_1 can be more easily approximated by $\mathbb{V}ar_{H_1}(\hat{\tau}) = \tau_m(1 - \tau_m)/2n$. Note that $\tau_m = z_{m,2} + \frac{1}{2}z_{m,1}$ can be expressed in terms of the λ -values using Eq. (8.17) for a general model by

$$\tau_m = \frac{1}{2} + \frac{1}{4\lambda_S} (2\Psi - 1)(2\lambda_S - \lambda_O - 1) = \frac{1}{2} + \delta_m. \quad (8.19)$$

Table 8.7 Number of affected sib-pairs n for the one-sided mean test required to detect linkage at a significance level of $\alpha = 0.0001$ and a power of $1 - \beta = 0.8$. Required sample sizes are given for a fully informative marker and, in parenthesis, for a chromosomal position with polymorphism information content (PIC) = 0.8. The distance between the trait and the marker locus is given by the recombination fraction θ .

λ_O	$\theta = 0$				$\theta = 0.05$			
	λ_S				λ_S			
	1.2	1.5	2	3	1.2	1.5	2	3
1.2	1488 (1832)	136 (168)	41 (51)	16 (20)	2273 (2798)	213 (263)	68 (84)	30 (37)
1.5		365 (450)	10 (13)	21 (26)		561 (691)	20 (25)	37 (64)
2			157 (194)	32 (40)			244 (301)	54 (67)
3				84 (104)				133 (164)

Under the null hypothesis of no linkage, $\tau_m = \frac{1}{2}$, and thus $\delta_m = 0$. Therefore, δ_m is a non-standardized measure of effect size. It turns out (see Problem 8.8) that the mean test has a power of

$$1 - \Phi\left(\frac{z_{1-\alpha} - 2\delta_m\sqrt{2n}}{2\sqrt{\frac{1}{4} - \delta_m^2}}\right)$$

for a sample of n independent ASPs and a genetic effect δ_m at a significance level α . Furthermore, in order to have a power of $1 - \beta$ of detecting linkage at a significance level α , the required number of independent ASPs n is at least (see Problem 8.3)

$$n \geq (z_{1-\beta} + z_{1-\alpha})^2 \cdot \frac{(\frac{1}{4} - \delta_m^2)}{2\delta_m^2} = (z_{1-\beta} + z_{1-\alpha})^2 \cdot \frac{\frac{1}{2}\tau_m(1 - \tau_m)}{(\tau_m - \frac{1}{2})^2}. \quad (8.20)$$

This formula is identical to standard formulae for sample size calculations, with $(z_{1-\beta} + z_{1-\alpha})^2$ being the *design effect* and $\frac{\frac{1}{2}\tau_m(1 - \tau_m)}{(\tau_m - \frac{1}{2})^2}$ being the *reciprocal of the squared standardized effect size* $1/\Delta^2$, where Δ is the *standardized effect size*. Recall that Δ is given by $\Delta = \frac{\mu_1 - \mu_0}{\sigma}$ for the one-sample Gauss-test, where μ_0 and μ_1 denote the means under the null and the alternative hypothesis, respectively, and σ is the standard deviation. Furthermore, note that $\tau_m(1 - \tau_m)/2$ is the variance of a single ASP, as an ASP has two informative meioses.

If marker informativity is incomplete, more ASPs are required. Guo and Elston [270] argued that typing of n ASPs with marker informativeness linkage information content (*LIC*) is equivalent to typing $n \cdot \text{LIC}$ ASPs at a fully informative marker locus. Thus, the required number of ASPs of Eq. (8.20) needs to be increased by a factor of $1/\text{LIC}$ if marker informativity is incomplete. As shown in Chapter 3, the *LIC* value can be expressed as a function of the heterozygosity (*HET*) and the *PIC* value assuming equally frequent marker alleles so that an adaptation of the required sample size n given any common informativity measure is possible.

Table 8.7 displays required sample sizes given locus-specific recurrence risk ratios λ_S and λ_O assuming a fully informative marker position and a position with *PIC* = 0.8. The latter corresponds to the common informativity at a chromosomal position in a genome-wide scan with microsatellite markers. Here, inter-marker distance varies between 5 and 15 cM, and thus $\theta = 0.05$ is a reasonable compromise for the largest possible distance between a marker and the trait locus. The required number of ASPs is also shown for $\theta = 0$. A point-wise significance level of $\alpha = 10^{-4}$ and a power of $1 - \beta = 0.8$ have been chosen. Furthermore, $\lambda_S \geq \lambda_O$ according to Eq. (8.14).

The table clearly indicates that samples sizes increase with the distance between the trait locus and the investigated chromosomal position. Furthermore, the lower the locus-specific λ , the larger the computed sample size. It may even deviate from 42 [4].

8.6 HOW ARE MODEL-FREE METHODS EXTENDED TO MULTIPLE MARKER LOCI?

In the last section, we saw that the required sample size to detect linkage to a specific chromosomal position depends on marker informativity. In applications, no single marker yields complete informativity, and microsatellite markers used in genome-wide scans usually have a heterozygosity between 70 and 90%. Considering a number of markers simultaneously instead of a single one shows some advantages. First, in the previous chapter on model-based linkage analysis (see Problem 7.6), we illustrated that using several markers can increase linkage information. This idea might also be useful for model-free linkage analysis. Second, extra information on IBD sharing may be obtained by condensing IBD information from different chromosomal positions. Finally, as already discussed in Section 7.2.2.2, when multiple markers are considered jointly, location estimates may be provided, though they generally will not be very precise [59, 246, 415].

It was mentioned in the previous chapter (see Section 7.2.2.3) that the likelihood may become quite complicated and that efficient algorithms are required for fast and easy calculations. The reason for this is fairly simple: in standard statistics, all subjects can be assumed to be independent, and the joint log likelihood is the sum of the individual likelihoods. Family members are, however, not independent, and thus the log likelihood is not a simple sum over all study subjects. Genetic markers that are linked to each other add to the dependency and aggravate the dependency problem. Fortunately, conditional independence can be assumed.

Two approaches that use conditional independence are possible. The first concept uses *conditional independence in pedigrees*, leading to the so-called *Elston–Stewart* algorithm [187, 481]. Here, one uses Markov properties by assuming that offspring genotypes are independent of grandparental genotypes given parental genotypes. The Elston–Stewart algorithm is extremely important for applications, which is best illustrated by the fact that the journal *Human Heredity* dedicated a special issue to celebrate the twentieth anniversary of its publication.

The second approach is as prominent as the Elston–Stewart algorithm and is called the *Lander–Green* algorithm [392]. For the Lander–Green algorithm, *conditional independence along a chromosome* is obtained if absence of interference is assumed so that the Markov property is utilized in a different way. For example, if we consider three genetic markers with the order M_1 – M_2 – M_3 , the markers M_1 and M_3 are independent given complete genetic information at marker M_2 .

In both algorithms, one also assumes that the phenotype of an individual depends only on its genotype and, possibly, individual-specific environmental factors. Thus, conditional independence is kept for phenotypes. Because of their great importance to applications, both algorithms have been improved and extended in several ways and can be used for calculating IBD distributions [25, 326, 378, 381, 444, 510]. If only the proportion of alleles shared IBD at an arbitrary chromosomal position is of interest, rather than the joint segregation pattern, a third alternative is available: the *Cardon–Fulker* or *Fulker–Cardon* algorithm [229, 231], which has been proposed for mapping

quantitative traits using random samples (see Chapter 9). This assumption of random sampling leads to an IBD distribution that is identical to that for ASPs under the null hypothesis of no linkage. The Cardon–Fulker approach is available in software packages and has been extended in several ways [18, 513]. All three algorithms are described in the Appendix. Some terms from the Lander–Green algorithm are required for understanding the statistical methods that will be discussed in this and the next chapter. These terms are explained in the following paragraphs.

The inheritance pattern of a pedigree can be completely described by an *inheritance vector*. The idea underlying the inheritance vector is that one determines for every individual in a pedigree whether the paternally or maternally inherited allele has been transmitted to his/her offspring. As a consequence, a meiosis can be described by a bit where 1 denotes the maternally inherited allele and 0 denotes the paternally inherited allele. All inheritance bits are collected to a vector, and the result is the inheritance vector, usually denoted by v . The dimension of the inheritance vector is identical to the number of meioses in the pedigree. For example, there are $b = 2n$ meioses in a pedigree with n non-founders (see Section 7.1.3.4 for the definition of the term founder), and the inheritance vector therefore is of dimension 2^b . It is important to note that the dimension of the inheritance vector increases exponentially with the number of non-founders, and the computational limit is around 20 non-founders. If the pedigree becomes too large, either individuals are removed or the pedigree is split in most applications, which, in turn, can influence the results.

In practice, the inheritance vector cannot be determined uniquely, but its distribution has to be estimated. This is done by using the Lander–Green algorithm. The distribution of the inheritance vectors forms the basis for calculating the IBD distribution of pairs of relatives. For a sib-pair, one has to check for both offspring and both maternally and paternally inherited alleles whether they are identical or not. This can be done by a comparison of the respective inheritance bits: if both respective inheritance bits are 0 or if both are 1, the allele is IBD, while it is not IBD if they are different.

8.7 WHAT ARE STANDARD APPROACHES FOR THE ANALYSIS OF LARGE SIBSHIPS?

In the previous sections, we have considered the simple situation that the sample consists of n independent ASPs. In many applications, however, the collected sample will contain several sibships with more than two affected siblings. In practice, one forms all possible pairs from a sibship in the analysis. Thus, for a sibship of size s , there are $\frac{s(s-1)}{2}$ possible pairs, though only $s - 1$ of these pairs are independent. Several authors have therefore suggested weighting schemes to downweight the number of possible pairs to the number of independent pairs [298, 648].

For the score- and Wald-type tests for ASPs from Section 8.3, there is, however, no need to downweight sib-pairs of large sibships. For these, Hodge [298] and Blackwelder and Elston [64] have shown that the number of alleles shared IBD are

pairwise—not mutually—independent. This pairwise independence implies that the covariance between the IBD value for any one sib-pair and the IBD value for any other sib-pair is zero. Therefore, the asymptotic distribution of the various test statistics is not affected by the dependence if all pairs are formed, and downweighting seems unnecessarily conservative.

Nevertheless, there is evidence that the MLS yields inflated type I errors if sib-pairs from large sibships are treated independently. Specifically, Abel and Müller-Myhsok [3] observed liberality in their simulation study for small type I error levels regardless of whether parents were typed (see also Ref. [303]). For this test statistic as well as in the case of small sample sizes, estimation of significance levels by Monte-Carlo simulation is advisable ([222, 303]; see also Chapter 9).

An alternative is to explicitly allow large sibships in the test statistic. A comparison of a broad variety of large sibship test statistics has been given by Sengul and colleagues [605]. One of these methods is the *maximum likelihood binomial (MLB)* method [3]. Other test statistics that explicitly allow large sibships will be discussed in the next section because they can also be used for extended pedigrees.

8.8 HOW CAN THE AFFECTED SIB-PAIR METHOD BE EXTENDED TO ARBITRARY UNILINEAL RELATIONSHIPS?

The classical approach for model-free analysis of extended pedigrees is the *affected pedigree member (APM)* method, originally proposed by Weeks and Lange [698]. In its original version, one compares the sum of the observed number of alleles shared IBS by each affected pair of relatives, weighted according to the frequency of the shared allele(s), with its expected value in the absence of linkage. Although it has been extended in several ways so that unaffected relatives could be included and analyses on the X chromosome were possible [450, 699, 701], the APM method had two major drawbacks. First, the results heavily depended on the weight function, i.e., the allele frequencies at the marker loci. Second, the APM method used only the IBS information and was hence less powerful than methods based on the IBD distribution [154, 252, 697]. One alternative is the *non-parametric linkage (NPL)* approach, originally proposed by Kruglyak and colleagues [378]. It can be considered an extension of the work by Whittemore and Halpern [716, 717] and will be described in the following paragraphs.

We start by reconsidering the score test version of the mean test from Eq. (8.11) for a sample of n independent ASPs

$$T_{S-M} = (\bar{\tau} - \frac{1}{2}) / \sqrt{\frac{1}{8n}}, \quad (8.21)$$

where $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \hat{\tau}_i$. It can also be written as

$$T_{\text{S-M}} = \sum_{i=1}^n \frac{1}{\sqrt{n}} \frac{\hat{\tau}_i - \frac{1}{2}}{\sqrt{\frac{1}{8}}} = \sum_{i=1}^n \underbrace{\frac{1}{\sqrt{n}}}_{=\gamma_i} \underbrace{\frac{\hat{\tau}_i - \mathbb{E}_{H_0}(\hat{\tau}_i)}{\sqrt{\text{Var}_{H_0}(\hat{\tau}_i)}}}_{=Z_i \sim (0,1)} \quad (8.22)$$

$$= \sum_{i=1}^n \gamma_i \frac{T_i - \mathbb{E}_{H_0}(T_i)}{\sqrt{\text{Var}_{H_0}(T_i)}} = \sum_{i=1}^n \gamma_i Z_i = Z. \quad (8.23)$$

While the contributions from families are averaged in Eq. (8.22) in the first step and standardized in the second, they are standardized first and averaged second in Eq. (8.21). Thus, Z_i has mean 0 and variance 1. Equation (8.22) is the key of the generalization to multiplex families. One needs only standardized scores for all families and a weighting factor fulfilling $\sum_{i=1}^n \gamma_i^2 = 1$ so that the weighted sum Z has mean 0 and variance 1 under the null hypothesis of no linkage. If the number of families is sufficiently large, Z is approximately normal according to the central limit theorem. Z as defined in Eq. (8.22) is the *NPL pairs* score [378]. We have already shown that the NPL score statistic from Eq. (8.22) is identical to the score-type version of the mean test, thus inheriting its optimality properties. We want to stress that the NPL test statistic is one-sided because we expect an increased allele sharing under the alternative hypothesis of linkage.

With the availability of the distribution of inheritance vectors, it is easy to generalize the NPL approach to the multipoint situation. One needs only to evaluate the conditional expectation of the score by summing over the scores given the inheritance vectors v at position x , multiplied by the respective probabilities:

$$Z_{i,x} = \sum_{v \in \mathcal{V}} P_{\text{complete},x}(v) Z_{i,x}(v), \quad (8.24)$$

where $Z_{i,x}(v)$ denotes the score of family i at position x given inheritance vector v .

In the past few years, many different scores T_i (see Eq. (8.23)) and weights γ_i have been discussed. A thorough technical discussion can be found in Ref. [455]. The most important scoring function for extended pedigrees was first proposed by Whittemore and Halpern [716]. Let a denote the number of affected individuals in a pedigree, let h be a collection of alleles obtained by choosing one allele from each of these individuals, and let $b_j(h)$ denote the number of times that the j th founder allele appears in h . The scoring function is defined as

$$T_{i,\text{all},x}(v) = 2^a \sum_h \left(\prod_j b_j(h)! \right),$$

where the sum is over the 2^a possible ways to choose h , and it is usually termed S_{all} in the literature.

Although the NPL approach is appealing, it was recognized as conservative [369]. As already discussed in Section 8.3.2, the score test from Eq. (8.22) evaluates the variance under the null hypothesis of no linkage. It is larger than the estimated

variance, and the denominator of the test statistic is larger than required. Furthermore, the test statistic from Eq. (8.24) can be overly conservative in the presence of missing data. Kong and Cox [369] modified the NPL approach by introducing a one-parameter likelihood associated with the NPL framework (for a nice description, see also Ref. [144]). Several implementations of the NPL statistics are freely available, including the packages GENEHUNTER, GENEHUNTER-PLUS, MERLIN, or ALLEGRO.

8.9 WHAT ARE POSSIBLE EXTENSIONS OF MODEL-FREE METHODS FOR DICHOTOMOUS TRAITS?

In the previous two sections, we discussed extensions of the ASP approach to large sibships and to extended pedigrees. There are other important modifications that are briefly discussed in this final section.

8.9.1 How can covariates in sib-pair analyses be handled?

The disease risk is often influenced by environmental factors, and the power for finding loci linked to the studied disease using ASPs can often be increased by covariate adjustments. One approach is to adjust for covariates prior to linkage analysis in a regression framework. Appropriate residuals can then be used for further analyses [140]. These approaches are problematic if covariates have to be defined on *pairs* instead of *individuals*; see Ref. [692].

An alternative approach that allows direct modeling of covariates is based on re-parameterizing the MLS (Eq. (8.3)) in terms of recurrence risk ratios λ (see Refs. [251, 514]; also see Refs. [260, 261] for a different parameterization):

$$\text{MLS} = \frac{\sum_{i=1}^n (\hat{z}_{i0}\lambda_0 + \hat{z}_{i1}\lambda_1 + \hat{z}_{i2}\lambda_2)}{\sum_{i=1}^n (z_{i0}\lambda_0 + z_{i1}\lambda_1 + z_{i2}\lambda_2)} \quad (8.25)$$

Here, $\hat{z}_{ij}$ and z_{ij} are the estimated and theoretical under H_0 probabilities that ASP i shares j alleles IBD, and the λ_j are the risk ratios for an ASP sharing j alleles IBD, for $j = 0, 1, 2$. Because $\lambda_0 = 1$, the model in Eq. (8.25) has two free parameters. If the λ values are written in their loglinear form $\ln(\lambda_j) = \alpha_j + \beta_j^T \mathbf{x}$, covariates $\mathbf{x}$ can be included in the model. By assuming a specific genetic model such as the minimax model or an additive genetic model, Eq. (8.25) can be reduced to a one-free-parameter model [251]. An implementation of this approach is available in S.A.G.E.

8.9.2 How can multiple disease loci be handled in sib-pair analyses?

In complex diseases, even if there is a locus with a major effect, not all families need to be linked to it. Therefore, heterogeneity LOD scores (HLOD) are of interest for some of the test statistics (see Chapter 7). These are implemented in MLBGH based on the MLB statistic. A different approach is taken by jointly analyzing two chromosomal regions. A theoretically straightforward extension of the single-locus

model to a general two-locus model can be obtained for the MLS. Instead of letting z_{ij} denote the probability that ASP i shares j alleles at a marker, one can use $z_{i,jk}$, which is the probability that ASP i shares j alleles at marker 1 and k alleles at marker 2. If the two loci interact in a multiplicative or an additive fashion (see Section 6.2.2), only four instead of five parameters $z_{i,jk}$ are required to complete the model [148, 204]. A nice summary on this topic has been given by Holmans [303].

8.9.3 Is it possible to estimate the recombination fraction in sib-pair analyses?

An important disadvantage of the model-free linkage analysis compared to the model-based linkage analysis is that no position estimates for the trait locus are provided in terms of the recombination fraction θ . However, Liang and colleagues [415] proposed an IBD-based procedure for estimating the location of a susceptibility gene within a chromosomal region framed by multiple markers. Their procedure is useful for estimating the location of a susceptibility gene when there is preliminary evidence that the chromosomal region includes a disease gene. It is model-free in the sense that no specification of penetrances or the mode of inheritance is required.

Liang and colleagues [415] derived a simple expression for the IBD sharing among ASPs at any position x in a chromosomal region as a function of the IBD sharing at the trait locus t in that region and the recombination fraction θ between x and t under some regularity conditions. For ASPs, the obtained functional relationship is given by $\mathbb{E}(2\hat{\tau}_x) = 1 + (1 - 2\theta_{xt})^2 (\mathbb{E}(2\hat{\tau}_t) - 1)$, where the indices denote the specific chromosomal positions. With this equality and by combining the data from all linked markers, Liang and colleagues [415] estimated the expected IBD sharing curve, thereby providing a location estimate and an IBD sharing estimate in ASPs at the trait locus. An advantage of this approach is that not only point but also interval estimates for the trait locus position can be derived, which in turn can be used for defining regions for fine-mapping the disease locus. The method has already been implemented in the freely available package `GENEFINDER` and generalized in several ways [59, 246, 416]. Nevertheless, it is not meant as a substitute to standard model-free linkage analyses [59, 415].

8.9.4 Should unaffected individuals be genotyped in sib-pair analyses?

Two strands of discussion can be found in the literature on the topic of whether unaffected individuals should be genotyped or not. While one strand is related to power, the other is related to the validity of the results from ASP or affected relatives analyses.

The power of sib-pair analyses, for either qualitative or quantitative traits, heavily depends on the certainty of the IBD status (see Section 8.3.2). Therefore, genotyping parents generally increases power, with the greatest gain occurring for markers with low informativity [304]. If parents are unavailable, genotypes from unaffected

siblings may be helpful in reconstructing parental genotypes, thus increasing power [305].

The validity of test statistics can also be improved by typing unaffected siblings. For example, in the Lander–Green algorithm, linkage equilibrium (see Chapters 4 and 11) is assumed among genetic markers being studied. This assumption is appropriate for studying sparse marker maps with high intermarker distances. However, with the use of SNP chips, this assumption is generally violated. Huang and colleagues [323] have demonstrated that this assumption can lead to inflated type I error levels if parental genotypes are missing. This bias can be eliminated when additional unaffected siblings are included in the analysis.

The most crucial assumption for ASP analysis is, however, that the allele sharing is $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ under the null hypothesis of no linkage. Zöllner and colleagues [770] have shown a modest genome-wide shift toward excess genetic sharing among sib-pairs sampled from the Hutterite population in South Dakota, which had not been selected with respect to a phenotype. Similar results have been reported before both for other species [519] and for the human genome.

These results are extremely important for interpreting linkage results from ASP or APM studies because these approaches correspond to studies where there are no controls available. They are therefore very sensitive to departure from the assumption of random allele sharing. This assumption can also be violated for several reasons including inbreeding in the population [242], assortative mating [628], selection or meiotic drive [186, 770], or, in more general terms, deviation from random sampling of sib-pairs or Mendelian segregation of the marker. As pointed out by Elston and colleagues [179, 186], any ASP or APM design that does not include some DSPs may be unable to validly detect these effects (also see Refs. [64, 304, 354]). Therefore, DSPs consisting of an affected and an unaffected subject should be collected additionally and analyzed separately from ASPs to quantify and possibly control for these effects. In contrast, any statistical method combining ASPs and DSPs in a joint analysis to gain power should be used with caution.

8.10 PROBLEMS

Problem 8.1 Determine the IBD and the IBS values of the sib-pairs in the two pedigrees depicted in Figure 8.5.

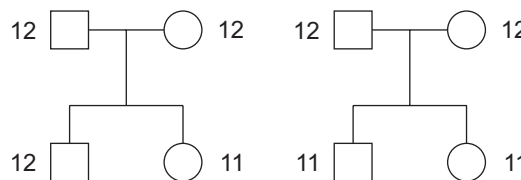


Fig. 8.5 Two nuclear families have been genotyped at a single genetic marker. What are the IBD and IBS values in the sib-pairs?

Problem 8.2 Consider a nuclear family with both parents being unaffected and both offspring being affected (Figure 8.6). The disease is simple autosomal recessive with complete penetrance, no phenocopies, and no mutations. Determine the IBD distribution in the offspring.

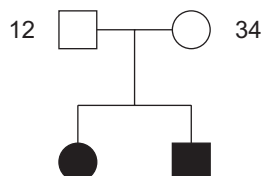


Fig. 8.6 Affected sib-pair family. The disease is assumed to be autosomal recessive with complete penetrance, no phenocopies, and no mutations. What is the IBD distribution in the offspring?

Problem 8.3 Consider a nuclear family with the mother and both offspring being affected (Figure 8.7). The father is supposed to be unaffected, thus not informative for linkage. The disease is assumed to be rare autosomal dominant with complete penetrance, no phenocopies, and no mutations. Determine the IBD distribution in the offspring.

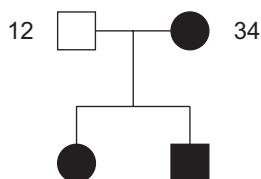


Fig. 8.7 Affected sib-pair family. The disease is assumed to be rare autosomal dominant with complete penetrance, no phenocopies, and no mutations. The parents are heterozygous at the marker locus with different alleles. Their genotypes at the trait locus are also given. What is the IBD distribution in the offspring?

Problem 8.4 Consider the data from Table 8.2. Note that we assumed that all parents have been genotyped.

Problem 8.4.1. Derive the update formula for this example.

Problem 8.4.2. What are the updates if starting values from the fully informative families are used, i.e., $z_0 = 0.1053$, $z_1 = 0.3421$, and $z_2 = 0.5526$?

Problem 8.5 Verify the validity of triangle conditions from Eq. (8.7), i.e., $z_1 \leq \frac{1}{2}$, $2 \cdot z_0 \leq z_1$ and $z_0 \geq 0$ using Eqs. (8.14) and (8.17).

Problem 8.6 Two unbiased estimators are sitting in a bar, having a few beers. The first one asks the other: “How do you like being married?”

Problem 8.7 Consider a sample of 10 ASP families with estimated two-point IBD distributions as given in Table 8.8.

Problem 8.7.1. Does the estimated IBD distribution fall in the possible triangle? Assume that both parents have been genotyped.

Problem 8.7.2. Calculate the score- and the Wald-type versions of the mean test.

Table 8.8 Notional identical by descent (IBD) distributions at a marker locus for 10 affected sib-pair (ASP) families.

Family	$\hat{z}_{i0}$	$\hat{z}_{i1}$	$\hat{z}_{i2}$	Family	$\hat{z}_{i0}$	$\hat{z}_{i1}$	$\hat{z}_{i2}$
1	0	0	1	6	0	0.5	0.5
2	0	0	1	7	0.5	0.5	0
3	0	0	1	8	0.5	0.5	0
4	0	0	1	9	0.5	0	0.5
5	0	0.5	0.5	10	0.5	0	0.5

Problem 8.8 Determine the formulae for both power and sample size calculation for the mean test.

URLs

ALLEGRO

<http://linkage.rockefeller.edu/soft/list1.html>

EASYLINKAGE

<http://compbio.charite.de/genetik/hoffmann/easyLINKAGE/>

GENEFINDER

<http://www.biostat.jhsph.edu/~wmchen/gf.html>

GENEHUNTER

<http://www.broadinstitute.org/ftp/distribution/software/genehunter/>

GENEHUNTER-IMPRINTING

<http://www.staff.uni-marburg.de/~strauchk/software.html>

GENEHUNTER-MODSCORE

<http://www.staff.uni-marburg.de/~strauchk/software.html>

GENEHUNTER-TWOLOCUS

<http://www.staff.uni-marburg.de/~strauchk/software.html>

GENEHUNTER-PLUS

<http://galton.uchicago.edu/genehunterplus/>

MERLIN

<http://www.sph.umich.edu/csg/abecasis/Merlin/download/>

MLBGH

abel@necker.fr

PARALLEL GENEHUNTER

<http://www.cs.sandia.gov/~sjplimp/genehunter.html>

S.A.G.E.

<http://darwin.cwru.edu/sage/>

9

Model-Free Linkage Analysis for Quantitative Traits

In the previous chapter, we focused on model-free linkage analysis of a disease. We pointed out in Section 2.3.5 that for complex diseases, these analyses might be strongly influenced by further genetic or environmental effects and that it may therefore be more promising to analyze an intermediate underlying phenotype, which usually is quantitative, rather than to search for causes of the disease status per se. As a result, this chapter is devoted to linkage analysis of quantitative traits, and Section 9.1 discusses the principal advantages and disadvantages of quantitative traits over dichotomous traits.

The simplest approach for model-free linkage analysis of quantitative traits is the Haseman–Elston (HE) method, which will be considered in Section 9.2. Many extensions of this regression-based approach have been proposed in the past 30 years to make it more powerful, and some of them will be described in Section 9.3.

An alternative method is the variance components (VC) model in which one assumes that the phenotypes of family members follow a multivariate normal distribution and whose analysis is based on maximum likelihood (ML) estimation. It has advantages in adjustments for covariates, is widely applied in applications, and therefore is discussed in some detail in Section 9.4.

Instead of using an ML approach for estimation, other groups have proposed score tests with similar properties. These will not be discussed in this book, and the interested reader may refer to the literature [140, 207, 253, 550, 654, 670, 690]. Another track would be non-parametric approaches for quantitative traits, including the *maximum likelihood binomial approach for quantitative traits (MLBQT)* [9], the *weighted pairwise rank correlation (WPC) method* [140], and the non-parametric linkage (NPAR) statistic based on the Wilcoxon signed rank test [380].

For the power of a specific study, the ascertainment scheme of the individuals is an extremely important factor, as this determines both phenotyping and genotyping costs as well as type I and II error levels. It will thus be detailed in Section 9.5.

Almost all statistical methods for linkage analysis that are discussed in this book rely on asymptotic properties of the test statistic. The goodness of the approximations to the limiting distribution depends on both the sample size and the test level. In many applications, the approximation of the exact distribution to the asymptotic distribution is not sufficient. Therefore, either exact tests or Monte–Carlo simulation or permutation approaches are preferable for determining p -values. For this reason, we discuss in Section 9.6 several methods for obtaining empirical p -values.

9.1 WHAT ARE ADVANTAGES AND DISADVANTAGES OF QUANTITATIVE TRAITS?

In the previous chapters, we regarded individuals in a study as either affected or unaffected. However, underlying this crude practical distinction is often a quantitative trait; if the trait value of a patient is above a specified level, he or she is viewed as affected. Hence, the clinically relevant disease is defined by a threshold relation. Other diseases are defined by a combination of criteria. But again, there may be *intermediate quantitative traits* leading to these criteria that could be available for linkage analysis in addition to or instead of the dichotomous trait. Examples of quantitative traits underlying disease definitions are given in Table 9.1. This leads to the question: Which phenotype should be preferred, the qualitative disease definition or the quantitative trait? Our own opinion is that if an analysis can be based on either a qualitative or a quantitative trait for a multifactorial disease, generally the quantitative measurement should be preferred.

- In different studies, different thresholds may be used for disease definition. For example, in some studies obesity is defined by a body mass index (BMI; kg/m^2) of at least 30, while an age- and gender-specific BMI percentile of 90 or 95 is used in other studies, and the comparability of the studies is questionable.

Table 9.1 Quantitative and related clinically relevant phenotypes.

Quantitative trait	Disease
IgE levels	Asthma
Blood pressure	Hypertension
Body weight	Obesity
Bone mineral density	Osteoporosis
Microfilaria density	Onchocercosis
Number of trematodes in excrement or urine	Bilharzia
Reading ability, writing ability	Dyslexia
Sleeping efficiency	Somnipathy

- In some diseases, reference ranges are defined only for specific age groups. For example, norms of current standardized tests for writing abilities are available in Germany only up to the VIth grade. Therefore, a general disease definition for higher age groups is difficult. Furthermore, reference values are mostly unsuitable for genetic epidemiological studies because the dependency on environmental and other factors is usually ignored. For example, obesity in adults is generally defined by a BMI of at least $30\text{kg}/\text{m}^2$. While less than 2% of 18-year-old German men exceed this threshold, 16% of 60-year-old German men have a BMI $\geq 30\text{kg}/\text{m}^2$. The use of the standard threshold would lead to disease severity varying with age. Thus, the quantitative trait generally is a more precise definition of the phenotype of interest than the dichotomous subdivision.
- It is likely that a complex genetic disease is not the result of a single gene. As pointed out by Allison and Faith [13] in the context of eating disorders, “it seems more plausible that a particular gene codes for a particular protein that produces some variation in a physiological or behavioral variable that, in turn, might place one at greater risk for the development of” the disease. Because of the underlying path model, the study of a so-called *intermediate quantitative trait* can result in a greater power than the study of the resulting dichotomous disease. Well-known examples of successful mapping of intermediate phenotypes are plasma levels of the angiotensin converting enzyme (ACE) and myocardial infarction, which are both related to the ACE gene [622]. Another example is the Lyon hypertensive rat, in which the phenotypes diastolic blood pressure and pulse pressure were mapped to rat chromosomes 13 and 2, respectively [172]. If the disease phenotype hypertension had been investigated instead of the two quantitative traits, these two loci might not have been identified.
- Statistical procedures for quantitative variables are usually more powerful than those in which the quantitative characteristic is reduced to a qualitative variable.

There are, however, two disadvantages of using quantitative traits compared to standard disease definitions. First, formal genetics, which includes determining the mode of inheritance, is difficult for quantitative traits because investigators do not see a disease running through the pedigree: quantitative traits are given instead of closed and/or open symbols. Second, a disease definition is directly used for treatment decisions. This is generally not the case for quantitative traits.

9.2 WHAT IS THE HASEMAN–ELSTON METHOD?

Model-free linkage analysis between a genetic marker and a quantitative trait using sib-pairs was first considered by Penrose [526], who proposed comparison of the covariance of the sib-pair phenotype and genetic marker differences with those expected when the trait and the genetic marker are not linked. The *Haseman–Elston (HE) method* [287] modified the Penrose approach and proposed using the principle of similarity in a different way. The second important aspect of the HE paper is

that they showed how partially informative families can be included in the analyses. The ideas of this paper can be considered the root of modern model-free linkage analysis of quantitative traits. It is therefore no wonder that the journal *Human Heredity* dedicated a special issue in 2003 to this method. Although he unjustly criticized the method at another point, the fundamental idea of the HE method has been well described by Robertson [568]: “The more similar the genotypes of sibs at the marker locus, the less will be the difference between them in the metric trait.” Put differently, the idea can be formulated the same way as for sib-pairs that have been randomly sampled from the population and a gene with a strong genetic effect on the quantitative trait:

- If a sib-pair is genetically similar at the linked genetic locus, the sib-pair should also be phenotypically similar.
- If a sib-pair is genetically dissimilar at the linked genetic locus, the sib-pair should also be phenotypically dissimilar.

Two major questions arise from this formulation. First, what are appropriate measures for genotypic and phenotypic (dis)similarity? Second, is there a statistical foundation for this approach? As already noted above, the IBD value of a sib-pair is an appropriate measure for genotypic similarity. An intuitive measure for phenotypic dissimilarity is the squared phenotypic difference of a sib-pair. This squared trait difference is also the Euclidian distance between the phenotypes of a sib-pair. Using these distance measures, the genetic marker and the phenotype of interest should be negatively correlated if they are linked [139]. Several alternative distance measures are possible, and some of them will be discussed in Section 9.3.

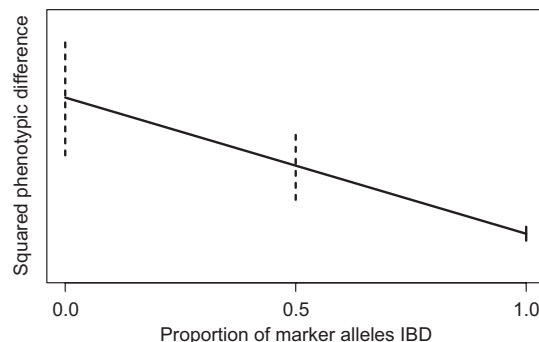


Fig. 9.1 Illustration of the Haseman–Elston method. The squared trait difference of a sib-pair is regressed on its proportion of alleles shared identical by descent (IBD). The greater the proportion of alleles shared IBD, i.e., the greater the genetic similarity, the greater the phenotypic similarity, i.e., the lower the squared trait difference.

The important contribution of Haseman and Elston was that they were able to give a simple answer to the second question. By using the simple Falconer model (Section 6.3), they derived a linear regression model for the squared trait difference on the proportion of alleles shared IBD at the trait locus (Section 9.2.1). The regression model can be easily extended to the case of a marker instead of a trait locus (Section 9.2.2). If marker and trait locus are linked, the slope of the regression of the squared trait difference on the proportion of alleles shared identical by descent

(IBD) should be negative (Figure 9.1). If the genetic marker and the phenotype are unlinked, the regression line should be parallel to the x -axis.

9.2.1 What is the expected squared phenotypic difference at the trait locus?

To formally develop the HE model, consider a sample of n independent sib-pairs. The phenotypes of the sib-pair i are denoted by x_{it} for $i = 1, \dots, n$ and $t = 1, 2$. The starting point is the simple Falconer model of Section 6.3.1, and we use the same model assumptions. Specifically, the phenotype x_{it} is additively decomposed into a general mean μ , a genetic effect g_{it} that is due to the trait locus, and an error term e_{it} , $i = 1, \dots, n$, $t = 1, 2$:

$$x_{i1} = \mu + g_{i1} + e_{i1} \quad x_{i2} = \mu + g_{i2} + e_{i2}$$

Neither a polygenic component nor shared environmental effects is included in the model. They are absorbed in the error term e_{it} so that there is a residual correlation $\text{Corr}(x_{i1}, x_{i2}) = \text{Corr}(e_{i1}, e_{i2}) = \rho_{SS}$. We assume $\mathbb{E}(e_{it}) = 0$ and $\text{Var}(e_{it}) = \sigma_e^2$. Furthermore, the diallelic trait locus has alleles A_1 and A_2 with allele frequencies p and q for alleles A_1 and A_2 and genetic effects as described in Chapter 6. Random mating, and Hardy–Weinberg equilibrium (HWE) are assumed to hold, and there is no major gene by residual interaction. The additive genetic variance is $\sigma_a^2 = 2pq(a - d(p - q))^2$, and the dominance genetic variance is given by $\sigma_d^2 = (2pqd)^2$.

In this section, we show that the squared trait difference $y_i = (x_{i1} - x_{i2})^2$ of a sib-pair, a measure of phenotypic dissimilarity, can be expressed as a simple linear function of the IBD information at the trait locus.

The calculation of the conditional expectation of y_i given z_{ij} is lengthy but straightforward and will be explained in detail. It is essentially based on Table 9.2, which is derived as follows. Column 1 shows the 9 cases of possible genotype combinations (g_{i1}, g_{i2}) of a sib-pair. These form the basis for the squared trait difference y_i given the genotype constellations. For example, consider case 4 where the first sib has genotype A_1A_1 and the second sib has genotype A_1A_2 . Then $x_{i1} = \mu + a + e_{i1}$ and $x_{i2} = \mu + d + e_{i2}$. Subsequently, $y_i = (x_{i1} - x_{i2})^2 = (a - d + \varepsilon_i)^2$ with $\varepsilon_i = e_{i1} - e_{i2}$. Columns 3–5 give the conditional probability of the genotype combination (g_{i1}, g_{i2}) given the IBD value z_{ij} , $j = 0, 1, 2$.

These can be derived as follows. For an IBD value of 2 (last column in Table 9.2), there are only three possible sib-pair genotype constellations. Put differently, if the sib-pair has two alleles IBD, the second sib has to have the same alleles as the first sib. And there are only three different genotype constellations for the first sib. If HWE holds, these occur with probabilities p^2 , $2pq$, and q^2 for genotypes A_1A_1 , A_1A_2 , and A_2A_2 , respectively.

The other two columns can be derived similarly. If the sib-pair shares null alleles IBD, four different alleles from the population have to be sampled. For example, the probability that one sib is heterozygous A_1A_2 at the trait locus is $2pq$. The probability

Table 9.2 Offspring genotype combinations at the trait locus, squared trait difference y_i given genotype combination at the trait locus, and conditional probability of the genotype combination given the identical by descent (IBD) value z_{ij} at the trait locus.

$g_{i1} - g_{i2}$	y_i	Conditional probability		
		z_{i0}	z_{i1}	z_{i2}
$A_1A_1 - A_1A_1$	ε_i^2	p^4	p^3	p^2
$A_2A_2 - A_2A_2$	ε_i^2	q^4	q^3	q^2
$A_1A_2 - A_1A_2$	ε_i^2	$4p^2q^2$	pq	$2pq$
$A_1A_1 - A_1A_2$	$(a - d + \varepsilon_i)^2$	$2p^3q$	p^2q	0
$A_1A_2 - A_1A_1$	$(-a + d + \varepsilon_i)^2$	$2p^3q$	p^2q	0
$A_1A_2 - A_2A_2$	$(a + d + \varepsilon_i)^2$	$2pq^3$	pq^2	0
$A_2A_2 - A_1A_2$	$(-a - d + \varepsilon_i)^2$	$2pq^3$	pq^2	0
$A_1A_1 - A_2A_2$	$(2a + \varepsilon_i)^2$	p^2q^2	0	0
$A_2A_2 - A_1A_1$	$(-2a + \varepsilon_i)^2$	p^2q^2	0	0

Conditional probability is $P(g_{i1}, g_{i2} | z_{ij}) = P(g_{i1}, g_{i2} | \text{IBD}_i = j)$ of genotype combination (g_{i1}, g_{i2}) given IBD value j . $\varepsilon_i = e_{i1} - e_{i2}$.

that both sibs are heterozygous (case 3) is the product of the individual probabilities and subsequently is given by $4p^2q^2$.

Finally, column 4 of Table 9.2 can be derived by noting that exactly three alleles need to be drawn from the population because one allele is shared IBD. Therefore, the probability for the first genotype combination is p^3 , which is the product of the probability for drawing three A_1 alleles from the population. The derivation of the probabilities displayed for cases 2 and 4–7 is done analogously because these sib-pairs have only one allele in common that has to be IBD. For example, case 4 displays one homozygous A_1 individual and one heterozygous individual. Therefore, the A_1 allele of the second sib has to be the one that is IBD, and two A_1 alleles for the first sib and one A_2 allele have to be drawn from the population. This combination has probability p^2q . Genotype combinations 8 and 9 have zero probability, as these sib-pairs cannot share one allele IBD. Finally, in case 3 the sib-pair shares either the A_1 allele or the A_2 allele IBD. The two possible genotype combinations occur with probabilities pq^2 and p^2q , respectively, and the sum of these two probabilities is pq . This completes the table.

The expected value of the squared trait difference given the IBD value can now be obtained as follows. First, we consider the case of z_{i2} :

$$\begin{aligned}
 \mathbb{E}(y_i | z_{i2}) &= \sum_{g_{i1}, g_{i2}} P(g_{i1}, g_{i2} | z_{i2}) \cdot \mathbb{E}(y_i | g_{i1}, g_{i2}) \\
 &= p^2 \mathbb{E}(\varepsilon_i^2) + 2pq \mathbb{E}(\varepsilon_i^2) + q^2 \mathbb{E}(\varepsilon_i^2) = \mathbb{E}(\varepsilon_i^2) = \sigma_\varepsilon^2,
 \end{aligned}$$

where $\mathbb{E}(\varepsilon_i^2) = \sigma_\varepsilon^2 = 2\sigma_e^2(1 - \varrho_{SS})$. Analogously, we obtain (see Problem 9.1)

$$\mathbb{E}(y_i|z_{i1}) = \sigma_\varepsilon^2 + \sigma_a^2 + 2\sigma_d^2, \quad \mathbb{E}(y_i|z_{i0}) = \sigma_\varepsilon^2 + 2\sigma_a^2 + 2\sigma_d^2,$$

which can be summarized to

$$\mathbb{E}(y_i|\text{IBD}_t) = \sigma_\varepsilon^2 + 2\sigma_g^2 - 2\sigma_g^2\tau_{t,i} + \sigma_d^2z_{t,i1} = \alpha + \beta\tau_{t,i} + \gamma z_{t,i1}, \quad (9.1)$$

where IBD_t denotes the IBD information at the trait locus t , $\tau_{t,i}$ is the proportion of alleles shared IBD at the trait locus t by family i , and $z_{t,i1}$ denotes the probability that sib-pair i shares one allele IBD at the trait locus t . Furthermore, $\alpha = \sigma_\varepsilon^2 + 2\sigma_g^2$, $\beta = -2\sigma_g^2$, and $\gamma = \sigma_d^2$.

Equation (9.1) shows that the squared trait difference at the trait locus is a linear function of τ_t and $z_{t,1}$. If there is no linkage between the genetic locus and the phenotype, $\sigma_g^2 = \sigma_d^2 = 0$. In this case, the regression of y_i on the τ is parallel to the x -axis with intercept σ_ε^2 . However, if the marker and the trait locus are linked, the linear regression of y_i on $\tau_{t,i}$ has negative slope $\beta = -2\sigma_g^2$ (see Figure 9.1).

The conditional expectation of the squared trait difference y_i given the IBD information has been modeled as a linear function. This conditional expectation is identical to the square of the random variable $x_{i1} - x_{i2}$, which in turn is similar to the conditional variance of $x_{i1} - x_{i2}$ given the IBD information. Subsequently, there is linkage between the genetic marker and the trait locus if at least two of the three conditional variances $\sigma_j^2 = \text{Var}(x_{i1} - x_{i2}|\text{IBD}_t = z_{ij})$, $j = 0, 1, 2$ are different. This idea leads to the VC approach, which will be discussed in detail in Section 9.4.

The squared trait difference y_i has been parameterized in terms of τ_t and $z_{t,1}$ in Eq. (9.1). The functional relationship between the phenotypes and the genotypes can also be formulated in τ_t and $z_{t,2}$ because

$$2\sigma_g^2\tau_t - \sigma_d^2z_{t,1} = 2\sigma_a^2\tau_t + 2\sigma_d^2\tau_t - \sigma_d^2z_{t,1} = 2\sigma_a^2\tau_t + 2\sigma_d^2z_{t,2}, \quad (9.2)$$

resulting in

$$\mathbb{E}(y_i|\text{IBD}_t) = \sigma_\varepsilon^2 + 2\sigma_g^2 - 2\sigma_a^2\tau_{t,i} - 2\sigma_d^2z_{t,i2} = \alpha + \beta^*\tau_{t,i} + \gamma^*z_{t,i2}. \quad (9.3)$$

With this parameterization, the regression can be on τ_t and $z_{t,2}$. The parameters β^* and γ^* can be interpreted as two times the negative additive and dominance variance, respectively, while β from Eq. (9.1) is two times the total negative genetic variance of the trait locus. Although both parameterizations give similar results, one should note that the original HE method has been derived in terms of Eq. (9.1), while the VC approach (see Section 9.4) is generally based on the parameterization from Eq. (9.3).

9.2.2 What is the expected squared phenotypic difference at the marker locus?

The trait locus is unknown in applications, and the aim is to map the phenotype to the trait locus given one or several marker loci. We therefore need to derive a

relationship between the squared trait difference and the IBD information IBD_m at a marker locus. This can be easily achieved by noting that

$$\mathbb{E}(y_i|IBD_m) = \sum_{j=0}^2 \underbrace{\mathbb{E}(y_i|IBD_t, IBD_m)}_{=\mathbb{E}(y_i|IBD_t)} P(IBD_t|IBD_m). \quad (9.4)$$

$\mathbb{E}(y_i|IBD_t)$ was derived in the previous section, and $P(IBD_t|IBD_m)$ is given in Table 8.6 if t and m are interchanged.

After appropriate replacement, one obtains [287]

$$\mathbb{E}(y_i|IBD_m) = \alpha_m + \beta_m \tau_{m,i} + \gamma_m z_{m,i1}, \quad (9.5)$$

where the regression coefficients are given by

$$\begin{aligned} \alpha_m &= \sigma_e^2 + 2\Psi\sigma_g^2 + 2(1-\Psi)\sigma_d^2, \\ \beta_m &= -2(1-2\theta)^2\sigma_g^2, \\ \gamma_m &= -(1-2\theta)^2\sigma_d^2, \end{aligned} \quad (9.6)$$

where θ denotes the recombination fraction between the trait and the marker locus and $\Psi = \theta^2 + (1-\theta)^2$.

Equations (9.5) and (9.6) indicate that linkage between a genetic marker locus and a quantitative trait can be detected if the regression coefficient β_m of a linear regression of the squared trait difference y_i on the estimated proportion of alleles shared IBD at the marker locus $\hat{\tau}_{m,i}$ and the estimated probability that the sib-pair shares one allele IBD $\hat{z}_{m,i1}$ are negative. We stress that this simple regression approach can also be applied to phenotypes that are not continuous. Indeed, the HE method has been applied to the dichotomous trait schizophrenia in a selected sample [184].

Equation (9.5) is a function of the proportion of alleles shared IBD and the probability of sharing one allele IBD; basically, there are three different possible hypotheses to test for linkage in this situation. First, one could conduct a test for γ alone. This is, however, not reasonable with the parameterization of Eq. (9.6) because a strict additive genetic model already implies $\gamma = 0$ although there might be a linkage. Therefore, only two hypotheses remain, and one could ask whether the test of the joint null hypothesis $H_0: \beta = \gamma = 0$ has greater power than the test of the single parameter β . This question has been answered by Blackwelder [62]. He showed by Monte-Carlo simulations that the joint linkage test has lower power than a linear regression ignoring γ in the estimation process and testing β only.

If γ is, however, set to zero in the regression estimation, a bias in the parameter estimate $\hat{\beta}$ might arise. In which situations is this bias relevant? There are two factors to the answer. First, we can borrow from results for the classical linear regression model. For this, $\hat{\beta}$ remains unbiased if either γ is indeed zero or the columns $\hat{\tau}_i$ and $\hat{z}_{i1}$ are orthogonal (see Ref. [169], p. 236). While the orthogonality condition generally will not be fulfilled, $\gamma = 0$ if there is no dominance variance or if the genetic marker and the phenotype are unlinked, i.e., $\theta = \frac{1}{2}$.

For traits with dominance variance, Amos and colleagues [23] have shown that the bias from ignoring $\hat{z}_{i1}$ when estimating β is negligible if parental genotypes are available in the analysis. If parental genotypes are missing, the bias increases with the dominance effect. It is largest for rare recessive traits and may become appreciably large when the allele frequency of the high allele A_1 , leading to high phenotypic values, exceeds 0.8 in the sample analyzed, which is rarely the case as a result of common sampling strategies. They therefore concluded that the bias is generally small [23, 27] and that the simple regression

$$y_i = \alpha + \beta \hat{\tau}_i + \varepsilon_i \quad (9.7)$$

can be used for linkage analysis and to test the hypothesis $H_0 : \beta = 0$ versus $H_1 : \beta < 0$. The test statistic is a standard t -test statistic and is given by

$$T = \hat{\beta} / \sqrt{\widehat{\text{Var}}(\hat{\beta})}, \quad (9.8)$$

$$\text{where } \hat{\beta} = \frac{s_{\tau y}}{s_{\tau\tau}} = \frac{\sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})(y_i - \bar{y})}{\sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{\tau}, \quad \hat{y}_i = \hat{\alpha} + \hat{\beta} \hat{\tau}_i,$$

$$\text{and } \hat{\sigma}_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \widehat{\text{Var}}(\hat{\beta}) = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})^2}.$$

Regardless of the true underlying distribution of the squared trait difference y_i , T is asymptotically normally distributed if the conditions are met allowing the use of the central limit theorem [582]. We again stress that the normality of the squared trait difference is not required, although this has been erroneously stated in the literature.

Nevertheless, for a better approximation with an asymptotic distribution, the central t -distribution with $n-2$ d.f. is preferable over the standard normal distribution. Furthermore, power and type I error levels may be better kept if either the individual phenotypes x_{it} are Winsorized [613] or the phenotype is Box–Cox transformed [192]. In *Winsorizing*, extreme values exceeding certain predefined upper and lower thresholds are replaced by the threshold values so that the extreme values are moved toward the center of the distribution. Although Winsorization is a useful technique, one should keep in mind that important information might be lost by truncating specific extreme values. Another approach to overcome liberality caused by extreme deviations from non-normality is to replace the standard ordinary linear regression with a generalized linear regression model [44].

The HE method is illustrated in a worked out example in Problem 9.2.

9.3 WHAT ARE COMMON EXTENSIONS OF THE HASEMAN–ELSTON METHOD?

The HE method is the roots of modern quantitative trait linkage analysis, and many extensions have been proposed. Some of them will be discussed in this section.

Many more can be found in the special issue of *Human Heredity*, 2003, volume 55, issue 2–3.

We begin with the extension of the HE method to the double squared trait difference (Section 9.3.1). The extension to large sibships, and how to deal with extended pedigrees, is sketched in Section 9.3.2. The phenotypic covariance of a sib-pair is a different measure for phenotypic similarity than the squared trait difference. This measure of phenotypic similarity leads to the *Haseman–Elston revisited* method, which will be discussed together with the *new Haseman–Elston* method in Section 9.3.3. Finally, we propose a simple approach for sample size and power calculation in Section 9.3.4.

9.3.1 How can the Haseman–Elston method be extended to the double squared trait difference?

In the previous section, we established the linear relationship between the squared trait difference and the IBD information for sib-pairs. Amos and Elston [26] derived similar regression relationships for the double squared phenotypic difference $y_i^2 = (x_{i1} - x_{i2})^4$ of sib-pairs and other relative pairs. More specifically, they showed that $\mathbb{E}(y_i^2 | \text{IBD}_i) = \alpha_2 + \beta_2 \hat{\tau}_i + \gamma_2 \hat{z}_{i1}$ also holds for sib-pairs and that $\beta_2 = 0$ under H_0 of no linkage. Furthermore, $\beta_2 < 0$ in the presence of linkage.

As a consequence, $\text{Var}(y_i | \text{IBD}_i) = \mathbb{E}(y_i^2 | \text{IBD}_i) - (\mathbb{E}(y_i | \text{IBD}_i))^2$ depends on the IBD distribution under H_1 . More specifically, the variance of the squared trait difference is different for sib-pairs with $\tau = 0, \frac{1}{2}$, and 1 in the case of linkage. The lowest variance can be observed for sib-pairs sharing two alleles IBD, and sib-pairs sharing zero alleles IBD have the largest variance in the squared trait difference (see Figure 9.1).

Subsequently, the linear regression model has heteroscedastic and not homoscedastic variances. Thus, $\text{Var}(y_i | \text{IBD}_i) \neq \sigma^2$ for all i . Therefore, Amos and colleagues [23] proposed to use not the common *ordinary least squares* as in the standard HE method but the so-called *weighted least squares*, where the weights are inversely proportional to the estimated variance of y given the IBD information. Unfortunately, the variances can become negative, if not adequately estimated, and the estimates are unstable for a small number of families. However, if variances can be estimated stably and if the weights are positive, this approach is optimal: according to the *Gauss–Markov theorem* it has the lowest variance among all unbiased estimation methods.

9.3.2 How can the Haseman–Elston method be extended to large sibships?

So far, we have considered only nuclear families consisting of two parents and exactly two offspring. If nuclear families with more than two offspring are recruited, dependencies arise as described in Section 8.7 for affected sib-pairs (ASPs). Unfortunately, the results of Blackwelder and Elston [64] for ASPs cannot be applied

to quantitative traits, i.e., the sib-pairs are not asymptotically pairwise independent. Instead, the correlation of two squared trait differences between two sib-pairs from the same family is typically of the magnitude of $\frac{1}{3}$ and $\frac{1}{4}$ [62, 723]. Therefore, the question arises whether the standard HE method is valid in the case of large sibships. A favorable result is that the standard HE method ignoring the within-family correlation is asymptotically valid [23, 139]. Power can, however, be increased if the within-family correlation structure is adequately estimated [63]. For example, in the S.A.G.E. package, two correlations are estimated: one for the correlation of the squared trait difference between two sib-pairs that have one sib in common ρ_1 and one for the correlation of the squared trait difference between two sib-pairs that have no sib in common ρ_0 . Because this correlation need not be correctly specified—e.g., environmental effects could well have an influence on the correlation between pairs—an adjustment is performed in the package for the possible misspecification of the correlation. The generalized estimating equations (GEE) approach is used to implicitly correct for this possible misspecification; see, e.g., Refs. [217, 765].

An alternative to the use of a different variance estimator is the adjustment for the degrees of freedom (d.f.) in the t -distribution [723]. Although an adjustment by the d.f. seems appealing, it is overly conservative in many situations [139]. Therefore p -values should be determined by Monte–Carlo simulations instead (see Section 9.6).

Questions on the validity of the HE approach in the case of large sibships are important for applications. Another question, however, is: Do large sibships increase the power to detect linkage? The answer is: Yes. Large sibships always have greater power than the same number of independent sib-pairs [10, 664]. As a consequence, fewer individuals need to be genotyped in total in the presence of large sibships. For example, 107 quintets, i.e., 535 siblings, have the same power in a single locus model as 952 independent sib-pairs, i.e., 1904 siblings. These results are also valid for families that are ascertained by a *single proband sib-pair (SPSP)* design (see Section 9.5).

In summary, large sibships should be ascertained if possible, and adjustments for the dependency should be performed. These adjustments can be carried out by using appropriate corrections of the variance estimator or by Monte–Carlo permutation tests. Alternatives to these approaches are the MLBQT [10] and the VC methods; the latter will be discussed below. While MLBQT has been specifically designed for sib-pairs, all other approaches can be used for extended pedigrees.

9.3.3 What are the Haseman–Elston revisited and the new Haseman–Elston method?

The squared trait difference is the measure of phenotypic similarity in the standard HE method. An alternative measure of phenotypic similarity that is used in the *Haseman–Elston revisited (HER)* method is the trait covariance of the sib-pairs, i.e., $(x_{i1} - \mu)(x_{i2} - \mu)$, where μ is the general mean in the population. Several questions for this measure of similarity might be posed:

1. Is there a simple linear relationship between genotypic similarity and phenotypic similarity as there was for the standard HE method? The answer is “yes.”
2. Which method has greater power, the standard HE method, which utilizes the squared trait difference, or the HEr method, which is based on the trait covariance? The answer is “it depends.”
3. Can the HEr method be made even more powerful? The answer is “yes.” This will lead to the *new Haseman–Elston (nHE)* method.
4. The trait covariance of the sib-pair requires an estimate of the general mean μ . How should μ be estimated for applications? The answer to this question is tricky, although some authors have discussed this issue [687, 692].

In this section, we will discuss these four questions step by step. We will not fully exploit the technical details of the HEr and the nHE method, but instead will sketch its derivation. For technical details, the interested reader may refer to the original work [170, 181]; also see the review by Feingold [208]. We have not prepared any problems for this section because computations can be done analogously to the standard HE method.

The method sparked off through two papers by Fulker and Cherny [230] and Wright [731], who pointed out that the standard HE method does not use all the information that is available in sib-pairs. More generally, if the random variables x_1 and x_2 are bivariate normally distributed with means 0 and a correlation different from $|1|$, the difference $x_1 - x_2$ does not capture the whole information. Only a bivariate distribution will contain the same amount of information as the original data. For example, the pair of random variables $(x_1 - x_2, x_1 + x_2)$ contains all available information. Standard statistics show that these new random variables are uncorrelated and thus are independent of bivariate normally distributed random variables. This property that the sum and the difference of two random variables are uncorrelated is an interesting technical issue for the derivation of the nHE method.

In the Falconer model (see Chapter 6), both x_1 and x_2 have mean μ so that the *mean-corrected trait difference* $(x_{i1} - \mu) - (x_{i2} - \mu)$ is considered instead of $x_1 - x_2$, and the *mean-corrected trait sum* $(x_{i1} - \mu) + (x_{i2} - \mu)$ is used instead of $x_1 + x_2$. The nHE method can now be derived as follows. In the first step, one shows that the squared mean-corrected trait sum $s_i = ((x_{i1} - \mu) + (x_{i2} - \mu))^2$ is a linear function of the proportion of alleles shared IBD by a sib-pair and the probability that a sib-pair shares one allele IBD. More specifically [170, 181],

$$\mathbb{E}(s_i | \text{IBD}_m) = \alpha_m^{(s)} - \beta_m \tau_{m,i} - \gamma_m z_{m,i1}, \quad (9.9)$$

with regression coefficients β_m and γ_m identical to those given in Eq. (9.6). Only the signs within the regression and the intercept $\alpha_m^{(s)}$ change.

In the second step, the squared trait difference is subtracted from the squared mean-corrected phenotypic sum, yielding $s_i - y_i = 4(x_{i1} - \mu)(x_{i2} - \mu)$. As a consequence, the mean-corrected cross-product of the traits has a mean of

$$\mathbb{E}((x_{i1} - \mu)(x_{i2} - \mu) | \text{IBD}_m) = \alpha_m^{(c)} - \frac{1}{2}\beta_m \tau_{m,i} - \frac{1}{2}\gamma_m z_{m,i1}. \quad (9.10)$$

Equation (9.10) implies that the trait covariance of the sib-pairs can be directly modeled as a linear function of the IBD information, and the sign of the regression coefficient is positive because $\beta_m = -2(1 - 2\theta)\sigma_g^2$ is negative. The statistical test for the nHE method can be carried out analogously to the standard HE test.

We can now proceed to the second question of power comparison between the standard HE and the HEr method. For randomly sampled sib-pairs, the HEr method generally has greater power than the standard HE method [181, 518], but it is suboptimal and can be inferior to the standard HE method in some cases [218].

This directly leads to the third question. The suboptimality results from the fact that the squared mean-corrected trait sum s_i and the squared trait difference y_i are not differentially weighted. In fact, if $\hat{\beta}_y$ and $\hat{\beta}_s$ denote the estimated regression coefficients of the squared difference and the squared sum regressions having variances σ_y^2 and σ_s^2 , respectively, then the optimal combination is

$$\hat{\beta} = \frac{\sigma_s^2}{\sigma_y^2 + \sigma_s^2} \hat{\beta}_y + \frac{\sigma_y^2}{\sigma_y^2 + \sigma_s^2} \hat{\beta}_s. \quad (9.11)$$

As a consequence, the single regression coefficients should be combined inversely proportional to their variances, and the resulting HE method is termed the nHE method. Several authors proposed using empirical variance estimates for combining the regression coefficients $\hat{\beta}_y$ and $\hat{\beta}_s$ (see, e.g., Ref. [218]). Sham and Purcell [608] have shown, however, that the variances σ_y^2 and σ_s^2 can be written as functions of the sibling correlation. They subsequently proposed a version of $\hat{\beta}$ from Eq. (9.11) with weights calculated from the correlation rather than estimated empirically. More specifically, for standardized phenotypes x_1 and x_2 with correlation ρ_{SS} , they proposed to regress $\frac{s_i}{(1+\rho_{SS})^2} - \frac{y_i}{(1-\rho_{SS})^2}$ on the proportion of alleles shared IBD τ_i . This directly leads to the fourth question from the beginning of this section: How should phenotypes be centered or adjusted? Before we answer this question, we will summarize the interesting properties of the HEr and the nHE methods.

First, the nHE method is approximately equivalent to the VC method [118, 608] (see Section 9.4), which is the optimal method in many situations if the assumptions made are met [208]. For example, if phenotypes deviate from normality, the VC approach can be liberal, or a loss of power can result. In contrast, both the HEr method and the nHE method are not liberal even if the phenotypes of the siblings markedly deviate from normality [15], although the nHE approach is asymptotically equivalent to the VC approach. However, the HEr generally suffers from low power in these situations; in contrast, the nHE method performs quite well. Furthermore, both the HEr and the nHE methods have the advantage of being computationally simple; we therefore consider the nHE method to be one of the best methods for applications. Even though the Sham and Purcell [608] approach requires some population parameters for calculating the optimal combination coefficient, Monte-Carlo simulation studies show that the results are fairly robust with respect to the choice of the parameters.

Now we can discuss the fourth question, which we have split into two questions:

1) How should the mean be chosen for the HEr? 2) How should phenotypes be

standardized for the nHE method by Sham and Purcell? One of the first papers on this topic was published by Wang and colleagues [687], who suggested using the family-wise mean in large sibships rather than the general mean in the sample. They showed that this is powerful when there is some kind of heterogeneity, e.g., when there are family-wise covariates.

In many applications, however, adjustments are required for family-wise and individual-level covariates. In this case, an appropriate regression is performed that includes the covariates of interest in the first step, and appropriate residuals—the ordinary residual for the HEr method, and the standardized residual for the nHE method—are used for further analyses (see, e.g., Ref. [140]). Wang and Elston [691] made a similar suggestion for independent sib-pairs and showed how a best linear unbiased predictor can be obtained that can then be used for adjusting the mean.

When performing adjustments, some researchers aim at obtaining a normal distribution for the overall distribution of the phenotypes prior to genetic analyses (see, e.g., Ref. [690]). However, as shown in Figure 6.3, the phenotypic distribution in the population may be strongly non-normal. Instead, some models perform better when the phenotypes *given* the genetic effect(s) are normally distributed. Thus, in many situations it is not necessary to have an overall normal distribution.

An interesting alternative is the two-level HE regression [692] in which the variance–covariance structure of the family data is directly modeled. It can therefore make use of all the trait information available in general pedigrees and simultaneously adjust both for individual-level and pedigree-level covariates and for the IBD information; also see Ref. [582] for a similar idea. Wang and Elston have shown that the two-level HE method is asymptotically equivalent to a VC model for detecting linkage under the assumption of normality.

In summary, there is more than one way of doing the mean adjustment or phenotype standardization, and we cannot recommend a single approach.

9.3.4 How can power and required sample size be calculated for the Haseman–Elston method?

Several papers on power and sample size calculations for the standard HE method can be found in the literature, and the first approach was published by Blackwelder and Elston [63]. Their method was extended later [23]. An advantage of the Blackwelder and Elston approach is that power and sample size calculations can be conducted for arbitrary unilineal pairs. A very general approach for sample size and power calculations that can be used for HE, HEr, nHE (see above), and VC methods (see Section 9.4) is given in Ref. [118].

Before we present a different, simpler method for sample size calculations, we summarize the most important finding of Blackwelder and Elston [63]. If a single locus model is considered and if marker and trait locus are identical, about twice as many pairs of second-degree relatives as sib-pairs are required for obtaining the same power. This result is intuitively plausible because second-degree relatives can share

only one allele IBD whereas sib-pairs can share two alleles IBD. For first-cousin pairs, the factor is approximately 3.

In the remaining part of this section, we discuss the approach for power and sample size calculations for the standard HE method of Ref. [272]. It extends and corrects the work of Ref. [566] who considered an additive genetic model and a fully informative marker that is identical to the trait locus. If sib-pairs are randomly sampled from the population, $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ of the sib-pairs should share 0, 1, and 2 alleles IBD at the diallelic trait locus (see Section 8.1).

The denominator of the HE slope coefficient $\hat{\beta}$ (Eq. (9.8)) is given by $\sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})^2$. $\bar{\tau}$ is approximately $\frac{1}{2}$. Furthermore, about half the sib-pairs, i.e., a total of $\frac{n}{2}$, share one allele IBD. For these sib-pairs, $(\hat{\tau}_i - \bar{\tau})^2 \approx 0$. The other half of the n sib-pairs share either zero or two alleles IBD at the trait locus. For these $\frac{n}{2}$ sib-pairs, $(\hat{\tau}_i - \bar{\tau})^2 \approx \frac{1}{4}$ so that $\sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})^2 \approx \frac{n}{2 \cdot 4}$.

The numerator of the HE slope coefficient is given by $\sum_{i=1}^n (\hat{\tau}_i - \bar{\tau})(y_i - \bar{y})$. Without loss of generality, we assume that the data have been centered so that $\bar{y} = 0$. As before, $\hat{\tau}_i - \bar{\tau} \approx 0$ for sib-pairs sharing zero alleles IBD. Similarly, $(\hat{\tau}_i - \bar{\tau})y_i \approx \frac{1}{2}y_{i,2}$ for the approximately $\frac{n}{4}$ sib-pairs sharing two alleles IBD, and $(\hat{\tau}_i - \bar{\tau})y_i \approx -\frac{1}{2}y_{i,0}$ for the approximately $\frac{n}{4}$ sib-pairs sharing zero alleles IBD. Here, $y_{i,j}$ denotes the squared trait difference of a sib-pair with family index i sharing j alleles IBD.

The HE slope coefficient $\hat{\beta}$ is therefore approximately given by

$$\hat{\beta} = \frac{\frac{1}{2} \sum_{i=1}^{n/4} (y_{i,2} - y_{i,0})}{\frac{n}{2 \cdot 4}} = \frac{4}{n} \sum_{i=1}^{n/4} (y_{i,2} - y_{i,0}).$$

For power and sample size calculations, mean and variance of $\hat{\beta}$ are required under both H_0 and H_1 . By using Eq. (9.6), $\mathbb{E}(\hat{\beta}) = -2\sigma_g^2$ if $\theta = 0$, and $\text{Var}(\hat{\beta})$ is approximately given by

$$\text{Var}(\hat{\beta}) = \text{Var}\left(\frac{4}{n} \sum_{i=1}^{n/4} (y_{i,2} - y_{i,0})\right) = \frac{16}{n^2} \text{Var}\left(\sum_{i=1}^{n/4} (y_{i,2} - y_{i,0})\right)$$

for both H_0 and H_1 . The squared trait differences are independent so that $\text{Var}(\hat{\beta})$ reduces to

$$\text{Var}(\hat{\beta}) = \frac{4}{n} (\text{Var}(y_{i,2}) + \text{Var}(y_{i,0})).$$

The null hypothesis we are interested in is that of no linkage $\theta = \frac{1}{2}$ [272]. In other words, we are interested in testing whether a marker is linked to the trait of interest. We thereby assume the existence of a genetic locus, i.e., $\sigma_g^2 > 0$. As a consequence, the hypotheses are $H_0: \theta = \frac{1}{2}$ versus $H_1: \theta < \frac{1}{2}$.

For sample size and power calculations, we need to calculate $\text{Var}(y_{i,2})$ and $\text{Var}(y_{i,0})$ under both the null and the alternative hypotheses. If $\theta = 0$ and $\sigma_g^2 > 0$, we obtain $\text{Var}(y_{i,2}) = 8\sigma_e^4(1 - \varrho_{SS})^2$, and

$$\begin{aligned} \text{Var}(y_{i,0}) &= \lambda + 2\sigma_e^4 + 4\sigma_g^4 + pq(\sigma_e^2 - \sigma_g^2) \cdot \\ &\quad (16a^2 - 32p^2ad + 16p^2d^2 + 32q^2ad + 16q^2d^2), \end{aligned}$$

where $\sigma_e^2 = 2\sigma_e^2(1 - \varrho_{SS})$ and $\lambda = 4pq(p^2(a-d)^4 + q^2(a+d)^4 + 8pqa^4)$. Because these calculations are quite cumbersome, they are left as Problems 9.3 and 9.4 to the reader. Furthermore, we can show $\text{Var}(y_{i,j}) = 8\sigma_e^4(1 - \varrho_{SS})^2 = 2\sigma_e^4$ for $j = 0, 1$, and 2 irrespective of θ if there is no genetic effect, i.e., if $\sigma_g^2 = 0$.

For calculating the variance of the squared trait difference y_i given the sib-pair shares k alleles IBD at a marker locus m , we use the law of total variance and the functional relationship

$$\text{Var}(y_i | \text{IBD}_m = k) = \mathbb{E}(\text{Var}(y_i | \text{IBD}_m = k, \text{IBD}_t)) + \text{Var}(\mathbb{E}(y_i | \text{IBD}_m = k, \text{IBD}_t)),$$

where IBD_t denotes the IBD status at the trait locus. This expression is identical to

$$\begin{aligned} \text{Var}(y_i | \text{IBD}_m = k) &= \sum_{l=0}^2 P(\text{IBD}_t = l | \text{IBD}_m = k) \text{Var}(y_i | \text{IBD}_t = l) \\ &\quad + \sum_{l=0}^2 P(\text{IBD}_t = l | \text{IBD}_m = k) \left(\mathbb{E}(y_i | \text{IBD}_t = l) \right. \\ &\quad \left. - \sum_{l=0}^2 P(\text{IBD}_t = l | \text{IBD}_m = k) \mathbb{E}(y_i | \text{IBD}_t = l) \right)^2, \end{aligned}$$

and the closed formulae of this equation can be obtained with straightforward but cumbersome algebra for $k = 0, 1, 2$. Because the formulae are unwieldy, they are not presented here. Instead, we refer the reader to the simple software tool `POWER.HE` for performing these calculations. Given means and variances under H_0 and H_1 , respectively, sample size and power calculations may be carried out in the usual way: The test is one-sided because $\beta < 0$ under H_1 , and H_0 is rejected if $T = \frac{\hat{\beta}}{\sqrt{\text{Var}_{H_0}(\hat{\beta})}} < z_\alpha$, where z_α denotes the lower α -quantile of the standard normal distribution. Thus, H_0 is rejected if

$$\hat{\beta} < z_\alpha \sqrt{\text{Var}_{H_0}(\hat{\beta})} = \frac{2z_\alpha}{\sqrt{n}} \sqrt{\text{Var}_{H_0}(y_{i,0}) + \text{Var}_{H_0}(y_{i,2})} = CR,$$

where CR is the abbreviation for the point defining the critical region. The power of the HE test is given by

$$1 - \kappa = \Phi\left(\frac{CR - \mathbb{E}_{H_1}(\hat{\beta})\sqrt{n}}{\sqrt{\text{Var}_{H_1}(\hat{\beta})}}\right) = \Phi\left(\frac{z_\alpha \sqrt{\text{Var}_{H_0}(y_{i,0}) + \text{Var}_{H_0}(y_{i,2})} + \sigma_g^2 \sqrt{n}}{\sqrt{\text{Var}_{H_1}(y_{i,0}) + \text{Var}_{H_1}(y_{i,2})}}\right),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. The reader should note that the power is denoted by $1 - \kappa$ in this section to avoid double notation. To obtain a power of $1 - \kappa$ at a significance level α for an arbitrary alternative hypothesis $\theta < \frac{1}{2}$, the required sample size is

$$n = \left(\frac{z_{1-\kappa} \sqrt{\text{Var}_{H_1}(y_{i,0}) + \text{Var}_{H_1}(y_{i,2})} - z_\alpha \sqrt{\text{Var}_{H_0}(y_{i,0}) + \text{Var}_{H_0}(y_{i,2})}}{\sigma_g^2} \right)^2. \quad (9.12)$$

An illustration of this approach is given in Ref. [272].

9.4 WHAT ARE VARIANCE COMPONENTS MODELS FOR LINKAGE ANALYSIS?

Variance components (VC) models have a long tradition in statistics. The first two papers modeling random effects date back to 1861 and 1863 [599]. In genetics, the concept of environmental and genetic correlations and covariances was first introduced by Fisher [213] in 1918 and is nowadays widely applied for many different purposes, including linkage and association analysis. In this section, we sketch only the concept of VC modeling, and the interested reader may refer to the literature for a thorough treatment of this topic [24, 67, 308].

In Section 9.4.1, we formulate the fundamental ideas using a simple univariate VC model. In Section 9.4.2, we sketch the multivariate VC model.

9.4.1 Is there a simple univariate variance components model?

In Section 9.2.1, we noted that the expected value $\mathbb{E}((x_{i1} - x_{i2})^2 | \text{IBD}_i)$ of the squared trait difference given the IBD information can be interpreted as the conditional variance of $x_{i1} - x_{i2}$ given the IBD information. We can therefore test whether the variances $\text{Var}(x_{i1} - x_{i2} | \text{IBD}_i = j)$ differ for $j = 0, 1, 2$. The concept behind this approach is intuitive. If a sib-pair shares two alleles IBD at a trait locus, the siblings have identical genotypes, and no variance is caused in the phenotypic difference due to the trait locus. However, if the sib-pair shares no alleles IBD, the variance in the trait difference is influenced by both the residual term and the difference caused by the genes, and this difference is $2\sigma_g^2$. The test can be carried out as a simple likelihood ratio test (LRT).

More formally, consider a sample of n independent sib-pairs with normally distributed phenotypic differences $x_{i1} - x_{i2}$. The hypotheses of interest are $H_0: \sigma_j^2 = \sigma^2$ and $H_1: \sigma_j^2 \neq \sigma^2$ for $j = 0, 1, 2$. To keep considerations simple, we assume that the IBD values can be determined unambiguously and that n_j sib-pairs share j alleles IBD. The IBD-specific variances σ_j^2 can then be estimated by

$$\hat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j},$$

where $y_{i,j}$ is the squared trait difference of sib-pair i with IBD value j . Similarly, the overall phenotypic variance σ^2 can be estimated via $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n y_i$. To test the equality of the variances, the LRT is given by $T = \sum_{j=0}^2 n_j \hat{\sigma}_j^{-2} \hat{\sigma}^2$, which is asymptotically χ^2 distributed with 2 d.f. [731].

In most applications, the IBD values cannot be determined unambiguously so that the variances need to be estimated by an expectation maximization (EM) algorithm (see Chapter 8).

If the variances are estimated under the assumption of an additive genetic model, i.e., $\sigma_1^2 = (\sigma_0^2 + \sigma_2^2)/2$, the asymptotic distribution under H_0 is that of a χ_1^2 [230]. The Falconer model and Eq. (9.1) imply that a reasonable restriction for the variances is $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2 \geq 0$. If the variances are estimated under this restriction, the limiting distribution is a mixture of χ^2 distributions with different d.f., and the mixture coefficients cannot be determined easily (compare with Section 8.3.1.3).

Sample size calculations can be performed easily for additive genetic models, random sampling of sib-pairs from populations, and no recombination between marker and trait locus, see Ref. [230].

The VC approach discussed in this section has been advocated, e.g., by Kruglyak and Lander [379], p. 449: “Although this [the VC] approach offers conceptual and technical advantages over the regression approach, it has not been adopted . . .” We do not agree with this interpretation because the principle of similarity as depicted in Figure 9.1 explains the Haseman-Elston concept in an intuitive way. Furthermore, we do not see that approaches requiring the EM algorithm have technical advantages over a simple linear regression. Also, the VC method is more dependent on the assumption of normality of the trait difference than the standard HE method and is less valid for selected traits (see Section 9.5). Finally, there is no power gain over the nHE method of Section 9.3.3.

In analogy to the standard HE method, the VC method does not use all available data because it compares only the squared trait differences given the IBD information. To overcome this power loss, the VC approach can be extended to a bivariate VC model (see Ref. [230]).

9.4.2 What is the extension of the simple Falconer model for variance components analysis, and how can the multivariate variance components analysis be performed?

The general VC model is easily obtained using the general variance analytic model of Eq. (6.5) of Section 6.3.2. Specifically, we decompose the phenotype x_{it} , $t = 1, \dots, T_i$, $i = 1, \dots, n$, additively into a general mean μ , a random genetic effect g_{it} due to the trait locus, a random polygenic component G_{it} , p individual fixed covariates $\mathbf{u}_{it}$, and an error term e_{it} :

$$x_{it} = \mu + g_{it} + G_{it} + \mathbf{u}_{it}'\boldsymbol{\beta} + e_{it} \quad (9.13)$$

With the usual assumptions that guarantee identifiability (see Section 6.3.2), including $\mathbb{E}(g_{it}) = \mathbb{E}(G_{it}) = \mathbb{E}(e_{it}) = 0$, the phenotypic mean and variance of individual x_{it} are given by

$$\mathbb{E}(x_{it}) = \mu + \mathbf{u}'_{it}\boldsymbol{\beta} \quad \text{and} \quad \text{Var}(x_{it}) = \sigma_a^2 + \sigma_d^2 + \sigma_G^2 + \sigma_e^2 = \sigma_g^2 + \sigma_G^2 + \sigma_e^2. \quad (9.14)$$

Here, σ_a^2 and σ_d^2 are the additive and dominance variances due to the trait locus, σ_G^2 is the variance due to polygenic components, and σ_e^2 is the residual variance. Model (9.13) does not include residual environment because this is modeled via the covariates $\mathbf{u}_{it}$, which can be defined to be individual-specific and/or constant across specific family members.

For $t \neq t'$, the functional relationship between the phenotypic covariance of unilineally related individuals t and t' of family i is given by (cp. Eq. 6.7)

$$\begin{aligned} \text{Cov}(x_{it}, x_{it'} | \text{IBD}_{i,tt'}) &= f_{i,tt'}(\theta, \tau_{i,tt'})\sigma_a^2 + h_{i,tt'}(\theta, \tau_{i,tt'}, z_{i1,tt'})\sigma_d^2 \\ &\quad + 2c_{i,tt'}\sigma_G^2, \end{aligned} \quad (9.15)$$

where $\tau_{i,tt'}$ denotes the proportion of alleles shared IBD by pair tt' of family i , $z_{i1,tt'}$ is the probability that this pair shares one allele IBD, and $c_{i,tt'}$ is the kinship coefficient of pair tt' (see Section 6.3.3).

$f_{i,tt'}$ and $h_{i,tt'}$ are functions of the recombination function and the IBD probabilities of relative pair tt' . For sib-pairs, they are given by

$$\begin{aligned} f_{i,tt'}(\theta, \tau_{i,tt'}) &= \frac{1}{2} + (1 - 2\theta)^2 \left(\tau_{i,tt'} - \frac{1}{2} \right), \\ h_{i,tt'}(\theta, \tau_{i,tt'}, z_{i1,tt'}) &= 4\theta^2(1 - 2\theta)^2 + (1 - 2\theta)^2 \tau_{i,tt'} + (1 - 2\theta)^4 z_{i1,tt'}. \end{aligned}$$

For other relative pairs, $h = 0$ always, and the functional relationship f has been given, e.g., by Amos ([24], Table 1).

If the aim is to estimate all VC from Eq. (9.14), different relative pairs are required. Estimation with only one specific type of relatives leads to non-identifiability problems. For example, σ_a^2 , σ_d^2 and θ are confounded with σ_G^2 and the residual variance σ_e^2 if only sib-pairs are analyzed. In general, parameters are identifiable if different types of relative pairs are included in the analyses. However, there are still situations in which parameters are confounded. For example, θ cannot be estimated validly if $\sigma_a^2 = 0$ [24]. In these situations, $\theta = 0$ should be used for estimation and testing procedures.

Before we can formulate the estimation problem, we require a little more notation. The phenotypes $x_{i1}, \dots, x_{iT_i}$ of all family members from family i are summarized to the column vector $\mathbf{x}_i$, the corresponding mean vector is denoted by $\boldsymbol{\mu}_i = \mathbb{E}(\mathbf{x}_i)$, and the variance matrix of the pedigree is denoted by $\boldsymbol{\Sigma}_i$. In general, one assumes multivariate normality of the traits, so that the kernel of the joint normed log likelihood function is given by

$$\mathcal{L}(\sigma_a^2, \sigma_d^2, \sigma_G^2, \theta) = \frac{1}{n} \sum_{i=1}^n \left(\ln \det(\boldsymbol{\Sigma}_i) + (\mathbf{x}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_i) \right).$$

Maximization of the kernel of the log likelihood can be done with standard software, and linkage can be tested by LRTs. For example, the hypothesis that the variance components are different from 0 can be tested by comparing the estimate of the unrestricted $-2\mathcal{L}(\sigma_a^2, \sigma_d^2, \sigma_G^2, \theta)$ with a model where $\sigma_a^2 = 0$, i.e., with the estimate of $-2\mathcal{L}(\sigma_a^2|_{\sigma_a^2=0}, \sigma_d^2, \sigma_G^2, \theta)$, and the test statistic is asymptotically χ_1^2 distributed.

A major advantage of this approach is that different types of covariates, i.e., individual-specific and relative-specific covariates can be adjusted for in the analysis. However, the crucial point of this method is the assumption of multivariate normality of the traits. If this assumption is violated, tests can lead to both substantially inflated type I error levels and almost no power for detecting linkage. For example, in Monte-Carlo simulation studies, Allison and colleagues [17] observed a type I error of more than 15% when the nominal significance level was 5%. They have proposed several approaches for making the VC models more robust against deviations from normality. Type I and type II error levels are also substantially influenced if families are ascertained by one or more family members. However, adjustments for single ascertainment, i.e., ascertainment via one individual, can be easily achieved by conditioning on the selected individual in the likelihood.

For more work on VC models, the interested reader may refer to Refs. [18, 24, 67] and the special issue on VC methods in *Behavior Genetics*, 2004, volume 34, issue 2.

9.5 HOW SHOULD SIB-PAIRS BE ASCERTAINED FOR LINKAGE ANALYSIS?

In the previous sections, we discussed several approaches for the analysis of quantitative traits. We focused on regression-based methods and also sketched VC analysis. Both methods can be derived using the Falconer models for randomly recruited families corresponding to a representative random sample of families from the population. However, this does not mirror the reality of many study situations where selected samples are available. These subsequently should lead to a greater cost efficiency in linkage studies. In this section, we discuss both the fundamental ideas and the advantages and disadvantages of selected sampling compared to random sampling of sib-pairs. Basically, three different designs are distinguished for the ascertainment of sib-pairs:

1. The *random sib-pair (RSP)* design, where both siblings are randomly ascertained from the population.
2. The *single proband sib-pair (SPSP)* design, also termed *extreme proband design*, where one offspring has an extreme phenotypic value.
3. The *extreme sib-pair (ESP)* design, where both siblings have extreme phenotypic values.

The RSP design has already been discussed in detail, and the principle of similarity corresponding to this design has also been stated in Section 9.2. A disadvantage of the RSP approach is that most included sib-pairs will have phenotypic values in the

reference range of the quantitative trait. In this case, the probability is high that both parents have only alleles predisposing to these normal values. The IBD value of a sib-pair therefore cannot be derived from the phenotypic values or the phenotypic similarity of the siblings. Furthermore, the compliance for participation in molecular genetic studies will be low in families that are not affected by a disease that is related to the quantitative trait of interest. The only true advantage of this design is that population parameters can be reliably estimated.

An alternative is the SPSP design, which is more powerful than the RSP approach in many situations. The idea of the SPSP design can be formulated as follows:

- If the proband has a high trait value and if the sibs are genotypically similar, then the second sib should have a high trait value.
- If the proband has a high trait value and if the sibs are genotypically dissimilar, then the second sib should have a low trait value.

This principle of similarity directly leads to the following regression approach for a sample of n independent sib-pairs [103, 104]:

$$x_{i2} = \alpha + \beta x_{i1} + \gamma(\tau_i - \frac{1}{2}) + \delta(|\tau_i| - \frac{1}{4}) \quad (9.16)$$

Here, the phenotype of the second sib x_{i2} is regressed on both the phenotype of the selected sib x_{i1} and the IBD information, and the test of linkage is carried out on γ . Instead of regressing x_{i2} on τ_i and z_{i1} , slight modifications are used in Eq. (9.16) because the new parameters can be nicely interpreted. Specifically, $\gamma = \mathbb{E}(x_{i2}|\text{IBD}_i = 2) - \mathbb{E}(x_{i2}|\text{IBD}_i = 0)$ and $\delta = \mathbb{E}(x_{i2}|\text{IBD}_i = 0) + \mathbb{E}(x_{i2}|\text{IBD}_i = 2) - 2\mathbb{E}(x_{i2}|\text{IBD}_i = 1)$.

The possible gain in power of the SPSP design compared to the RSP design is illustrated in Figure 9.2. There is substantial gain in power for the displayed SPSP design compared to the RSP approach regardless of whether the index person has been selected from the top 10% or top 5% of the phenotypic distribution. The gain in power for the selected families can be explained by an enrichment of the alleles leading to different phenotypes in the offspring.

The SPSP design does not only have the advantage of often being more powerful but the recruitment of families may also be easier. In many applications, ascertainment of sib-pairs via a proband with treatment-requiring symptoms for an underlying quantitative trait can be easily achieved if sampling can be performed in a clinic. This approach can also lead to high compliance because family members are often concerned about the treatment-requiring symptoms of the proband.

Sampling by a proband can increase the statistical power. Therefore, it is obvious to ask whether the power to detect linkage can be increased further by selecting sib-pairs where both sibs have extreme phenotypic values. The resulting design is termed the ESP approach. In fact, it can be shown analytically for a single major gene model, that the number of genotypes required to obtain a specific power for detection of linkage may be reduced from 10- to 40-fold compared to conventional studies for mapping quantitative traits using ESP [566]. Specifically, Darvasi and Soller [153] had shown that “almost 100% of the information used in a linkage analysis is obtained from the top 25% and bottom 25% tails of the distribution.” Thus, “in

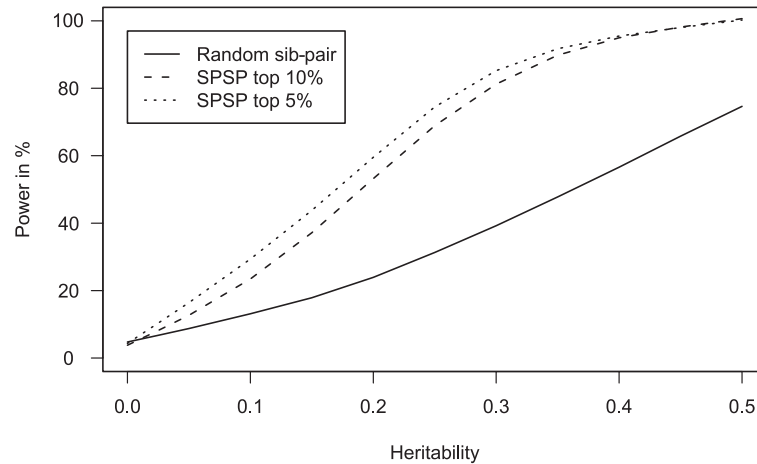


Fig. 9.2 Gain in power for the single proband sib-pair (SPSP) design compared to the random sib-pair (RSP) design against heritability from a Monte–Carlo simulation study of a dominant genetic model with $p = 0.2$, $\mu = 0$, $\sigma_e^2 = 1$, $\varrho_{SS} = 0.4$, $\theta = 0$, $\alpha = 0.05$, and $n = 240$ completely informative sib-pairs. This figure is adapted from Ref. [240] and used with kind permission of *Medizinische Genetik*.

single trait studies, it will almost be never useful to genotype more than the upper and lower 25% of a population.”

The basic idea of the ESP approach can be described as follows:

- If the sib-pair is phenotypically similar, then the sib-pair should also be genotypically similar.
- If the sib-pair is phenotypically dissimilar, then the sib-pair should also be genotypically dissimilar.

It has specifically been shown that *extreme discordant sib-pairs (ED)*, i.e., sibs from opposite tails of the phenotypic distribution, are useful for dominant or additive inheritance models. *Extreme concordant sib-pairs (EC)*, where both sibs are from the same tail of the distribution, are preferable for rare recessive genes. Because EC and ED have the flavor of ASPs and discordant sib-pairs, respectively, the approaches described in Chapter 8 can be applied for linkage analyses in these situations.

There are several arguments against the use of the ED approach based on the power (for a review, see Ref. [761]):

- For oligogenic quantitative traits, i.e., traits that are influenced by more than one locus, the power to detect linkage can decrease by sampling more extreme sib-pairs.
- For oligogenic models with different effects for each locus, the power to detect the one with the weaker displacement effect does not necessarily increase with extremity of sampling.

- If ESPs are ascertained for the most interesting quantitative trait, the total number of sib-pairs that are extreme for a different quantitative phenotype depends on the correlation of the phenotypes and can be drastically reduced so that the power to detect linkage can be low.

Another important issue is that previous arguments were based on reducing the genotyping in a given study to a minimum. However, one should keep in mind that the practical intention should be to minimize the total study cost. For this, it is crucial to note that ostensibly full-sibs may be half-sibs in the ED approach [495, 761]. Estimates for non-paternity vary between 0.3 and 30%, depending on factors such as socioeconomic status of the family, cultural background, or social structure. Generally, population prevalences of such non-paternity may be as high as 12% [109]. The risk that an ED is a half-sib-pair is between 1.5 and 2 times higher than the risk in the general population [761]. At the same time, the a posteriori probability for an EC to be half-sib-pair is approximately 0.7 times as high, thus lower than that in the general population. Thus, extreme sampling and even more extreme sampling is not always better.

In summary, there is not a single ascertainment strategy that is optimal or almost optimal in all situations. Therefore, investigators should collect all information available, determine the optimal study design on the basis of these data, and decide whether the optimal or at least a reasonable recruitment scheme can be realized.

9.6 HOW CAN P-VALUES BE DETERMINED EMPIRICALLY?

All linkage statistics that have been discussed in this book rely on asymptotic properties. The goodness-of-fit of these approximations depends on three factors: the test statistic itself, the sample size, and the significance level of interest. Most rules of thumb for standard approximations have been derived for the 5% test level. In genetic epidemiological applications, however, much lower significance levels are often required. Therefore, the calculation of empirical p -values has been the subject of interest (see, e.g., the letter by North and colleagues and the subsequent discussion [504, 505]). A brief summary of the available approaches is given in Ref. [222].

The following basic procedures can be distinguished. For linkage analysis of quantitative traits in sib-pairs, one may permute either phenotypes or genotypes—for model-free approaches, IBD probabilities—to disperse the dependency of genotypes and phenotypes [686]. To give an example, for the HE approach with n independent sib-pairs, the Monte-Carlo permutation algorithm is as follows.

Algorithm 9.1.

1. Compute the test statistic T of Eq. (9.8) using the original n data pairs $(y_i, \hat{\tau}_i)$
2. For $k = 1$ to a specified number N of permutations,
3. Set counter $C = 0$
 - (a) Permute the squared trait differences y_i , giving $y_i^{(k)}$, and match them to the fixed IBD proportions $\hat{\tau}_i$ of the n pairs
 - (b) Compute $T^{(k)}$ using the permuted pairs $(y_i^{(k)}, \hat{\tau}_i)$

- (c) If $T^{(k)}$ from the permuted data is at least as small as $T^{(k)}$ using the original data, add 1 to C
 - (d) $k = k + 1$
4. The Monte-Carlo p -value is C/N

In ASP studies, however, all sib-pairs have identical phenotypic values, and the approach described in Algorithm 9.1 is not applicable. Furthermore, care is needed when extended pedigrees are considered.

In extended pedigrees, different relative pairs are jointly analyzed. Usually, pedigrees are of different size and thus neither genotypes nor phenotypes can be permuted across pedigrees. Furthermore, genotypes cannot be permuted within pedigrees because this would lead to incompatibilities with Mendelian inheritance. Therefore, one alternative is to permute phenotypes within pedigrees. Although this approach seems to be attractive, it has one major disadvantage, especially for quantitative phenotypes: it destroys the residual correlation between pairs. This residual correlation may vary between different types of relatives. More specifically, sib-pairs generally have more shared environment than first-cousin pairs. If not all factors resulting in shared environment are taken into account in the linkage analysis, the degree of residual correlation is different for these pairs. As a result, a bias may be introduced if phenotypes are permuted within pedigrees.

Therefore, a common approach is to simulate new marker genotypes of each individual in the pedigrees under the null hypothesis of no linkage. Each simulated sample is subject to the same analysis to derive an empirical distribution of the test statistic under H_0 . Although simulation of genotypes is relatively straightforward, it can be time consuming when simulations are done conditional on partial pedigree information [742]. For this approach, marker informativity should be chosen according to the information content (Chapter 3) at the chromosomal position of interest [769].

Zhao and colleagues [742] proposed a randomization procedure for the inheritance vectors obtained from the Lander–Green algorithm (see Appendix A.2). The rationale of this approach is that if a locus is not linked to the trait locus, the grandpaternal and grandmaternal alleles should have an equal chance of transmission to the offspring. Therefore, these alleles have an equal chance to be 0 or 1 in the i th non-founder in the pedigree. Advantages of the randomization procedure compared to other simulation methods include the applicability to both qualitative and quantitative traits; arbitrary pedigree structures, provided that the inheritance vector probabilities can be estimated; and low computational effort, once the inheritance vector probabilities have been estimated in the original sample. A Monte–Carlo permutation procedure which is based on proportionality of marker informativity has been proposed in Ref. [222].

9.7 PROBLEMS

Problem 9.1 Show the validity of $\mathbb{E}(y_i|z_{i1}) = \sigma_\varepsilon^2 + \sigma_a^2 + 2\sigma_d^2$ and of $\mathbb{E}(y_i|z_{i0}) = \sigma_\varepsilon^2 + 2\sigma_a^2 + 2\sigma_d^2$.

Problem 9.2 Consider the 10 nuclear families of Figure 9.3. They have been randomly drawn from the population. A single genetic marker has been genotyped, and a quantitative trait has been assessed that may be linked to the genetic marker.

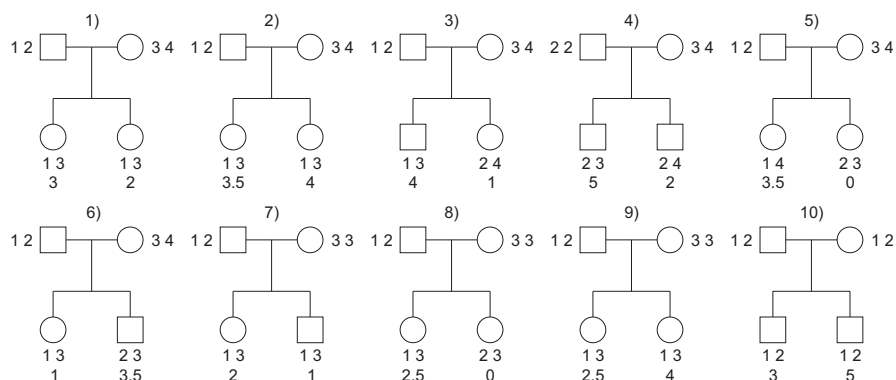


Fig. 9.3 Ten nuclear families for Problem 9.2 that have been randomly drawn from the population. A single genetic marker has been genotyped in both parents and offspring (first row), and a quantitative trait has been assessed in the offspring only (second row).

Problem 9.2.1. Estimate the proportion of alleles shared IBD.

Problem 9.2.2. Calculate the squared trait difference.

Problem 9.2.3. Perform an HE regression using all 10 sib-pairs under the assumption of $\gamma = 0$. Estimate the regression coefficient $\hat{\beta}$ and perform a t -test.

Problem 9.3 Calculate $\text{Var}(y_{i,2})$ with the assumptions from Section 9.3.4 for the case that marker and trait locus are identical, i.e., $\theta = 0$.

Problem 9.4 Calculate $\text{Var}(y_{i,0})$ with the assumptions from Section 9.3.4 for the case that marker and trait locus are identical, i.e., $\theta = 0$.

URLs

MERLIN

<http://www.sph.umich.edu/csg/abecasis/Merlin/download/>

GENEHUNTER

<http://www.fhcrc.org/labs/kruglyak/Downloads/>

POWER.HE

<http://www.imbs-luebeck.de/download/software>

S.A.G.E.

<http://darwin.cwru.edu/sage/>

SOLAR

<http://www.sfbr.org/solar/>

10

Fundamental Concepts of Association Analyses

In Chapters 7–9, we described in detail a number of genetic epidemiological methods that aim at localizing a specific disease locus. The common feature of those methods is that they all utilize the principle of genetic linkage. Consequently, the collection of families in which the disease of interest is inherited is required. In contrast to that, another group of methods has been proposed that do not necessarily need family data. Instead, as in classical epidemiological studies, information from unrelated probands with or without the disease can be used for the localization of a disease locus. The underlying principle here is the concept of genetic association, and this will be the focus of this chapter. More specifically, we will explain the concept of genetic association and point out its differences from genetic linkage in Section 10.1. Because the principle of genetic association strongly depends on linkage disequilibrium, we will discuss this concept and its measures in detail in Section 10.2.

10.1 WHAT IS GENETIC ASSOCIATION?

10.1.1 What is the principle of genetic association?

In general terms, an association exists between any two characteristics if they occur more often than would be expected by chance in the same individual. For a *genetic association*, these two characteristics are a disease, or a phenotype, and a specific allele at a genetic marker. Hence, we see a genetic association if the specific allele is more frequent in a group of affected than in a group of non-affected individuals. From this perspective, genetic association studies resemble classical epidemiological

approaches. Here, the risk factor under investigation is the allele at the genetic marker.

Now, how does this differ from genetic linkage? As we saw in Chapters 7-9, genetic linkage measures the deviation of the independent transmission of a genetic locus and a phenotype or another genetic locus within a family. Association also measures the deviation from independent transmission, but on a population level instead of within families. To pinpoint the differences, two issues distinguish linkage and association:

1. Linkage looks at the transmission of a locus with a disease, whereas association focuses on the relation of an allele with a disease. We repeat the mnemonic from Chapter 7:

Linkage is with Loci—Association is with Alleles

2. Linkage is based on transmission within families, whereas association is within populations.

If an association is observed between an allele and a disease, how can this be interpreted? Basically, there are two possible causal models and hence different interpretations for an association, as depicted in Figure 10.1. In the first case (see Figure 10.1a), the observed association is direct. Here, the analyzed marker locus represents the disease locus itself. Hence, the observed association mirrors the causal relationship between the genetic disease locus and the disease.

Second (Figure 10.1b), in the model of *indirect association*, an association between the disease and a specific marker allele is observed, but this is due to a more complex relationship in the background. Specifically, the marker locus is close to the disease locus, which in turn causes the disease. This dependence between the marker and the disease locus, or, more generally, between two genetic loci, is termed *linkage disequilibrium (LD)* (see also Chapter 5). Hence, it is important to note that LD is the dependence of alleles at two genetic loci, whereas association refers to the dependence of an allele at a genetic locus and a phenotype.

The assumed causal model has consequences for the design of a genetic association study. Studies that are based on the model of direct causality require the selection

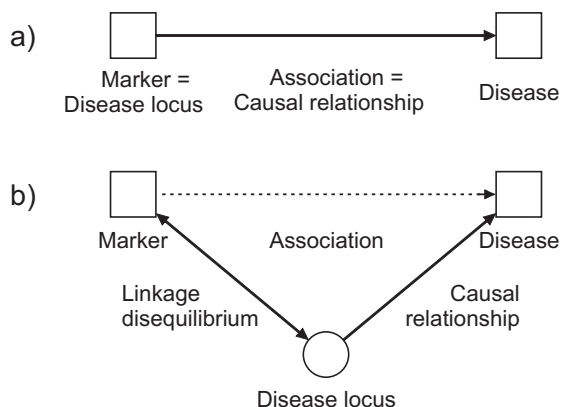


Fig. 10.1 Causal models for genetic association. In direct association a), the marker locus is identical to the disease locus. In indirect association b), the marker locus is in linkage disequilibrium with the disease locus.

of suitable candidate loci for the analysis that might be directly disease relevant. Without this knowledge, a study has to utilize indirect association. Here, many more genetic markers usually have to be genotyped. Also, this concept relies on the specific dependence between the loci to be strong enough to cause an observable association. Hence, the strength of the LD is of essential importance in these studies; therefore, we will discuss this concept in detail in Section 10.2.

10.1.2 Which study designs are used in genetic association analysis?

To study genetic association, a broad distinction can be made between studies using data from unrelated individuals and those based on family data. If unrelated individuals are recruited for the study, it will strongly resemble classical epidemiological studies. Hence, similar study designs can be employed, two of which are depicted in Figure 10.2. For example, in a prospective *cohort study* (Figure 10.2a), a sample of probands—a cohort—is recruited without knowledge of their disease status, ideally prior to anyone developing the disease. The probands are differentiated with regard to the risk factor, which in this case is the genotype. It will then prospectively be observed who develops the disease. An alternative epidemiological design is shown in Figure 10.2b). Here, a sample of affected and unaffected individuals, the cases and the controls, are recruited. These are then analyzed with respect to their genetic status. Hence, the major difference between these two designs is that in a cohort study, the time flow of the study matches the flow of causality, whereas in a *case-control study*, the disease is assessed before the risk factor. Both designs have their advantages and disadvantages, which have been reviewed in detail (see, e.g., Refs. [264, 522, 596] and the standard epidemiological text books [85, 86, 570]). We therefore point to only one specific difference between standard epidemiological studies and genetic epidemiological case-control studies. While the backward-in-time assessment of risk factors usually leads to greater measurement errors in case-control studies than in cohort studies, DNA is stable and generally does not alter over time. Therefore, genetic epidemiological case-control studies suffer from substantial measurement errors only

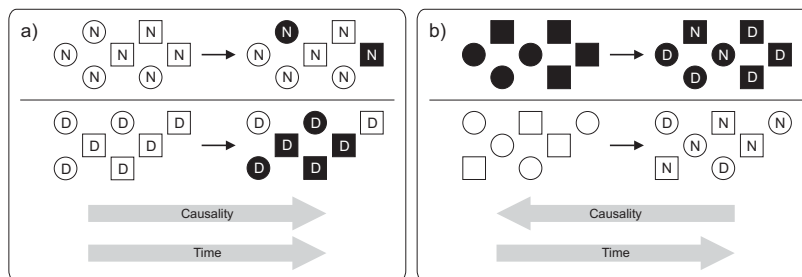


Fig. 10.2 Study designs for association analysis with unrelated individuals. a) Cohort study recruiting probands without knowledge of their disease status. b) Case-control study recruiting probands according to their disease status.

if confounders or supplementary variables are taken into account. In principle, the same statistical methods are used for both study designs, which will be discussed in Chapter 11.

In contrast to these classical epidemiological designs that use unrelated individuals, specific designs have been developed for genetic association studies. Like study designs for linkage analyses, these designs utilize families, and a variety of different pedigree structures can be used. Genetic association study designs that are based on family data will be described in Chapter 12.

10.2 WHAT IS LINKAGE DISEQUILIBRIUM?

As already defined above and in Chapter 5, *linkage disequilibrium (LD)* traditionally describes the deviation from independence between the alleles at two genetic loci in a natural breeding population. Hence, *association* is the relation between an allele and a phenotype, whereas LD refers to the relation between two alleles or to the relation between two genotypes. Synonymous terms for allelic LD that are used less frequently are *allelic disequilibrium*, *gametic disequilibrium*, or *haplotype-based disequilibrium*. If an association study is based on an indirect association, thus on LD between alleles or genotypes at a marker and a disease locus, we need to know how LD can be quantified. Here, two fundamentally different concepts can be distinguished, *allelic LD* and *genotypic LD*. While allelic LD measures rely on the knowledge of the haplotypes, genotypic LD is based on SNP-wise genotypes. Allelic LD measures will be described in the next section, and they are most often used in applications. In many laboratories, only genotypes are observed and haplotypes are ambiguous. In these situations, genotypic LD measures might be preferable (Section 10.2.2). In addition, we need to consider how far two genetic loci might be physically apart while still displaying LD that is strong enough to cause an observable association (Section 10.2.3).

10.2.1 How can allelic linkage disequilibrium be measured?

Table 10.1 Genotype frequencies at markers M_1 and M_2 .

		Marker M_2			
		bb	Bb	BB	
Marker M_1	aa	n_{00}	n_{01}	n_{02}	n_{0+}
	Aa	n_{10}	n_{11}	n_{12}	n_{1+}
	AA	n_{20}	n_{21}	n_{22}	n_{2+}
		n_{+0}	n_{+1}	n_{+2}	n

Before we describe different allelic LD measures, we explain the problem of haplotype ambiguity when only genotypes, and not haplotypes, are measured in the laboratory. Consider the case of two diallelic markers M_1 and M_2 with alleles A and a at marker locus M_1 and B and b at marker locus M_2 having been genotyped in a sample of n subjects (Table 10.1). Here, for all n_{22} probands who have two A alleles and two B alleles, thus who are homozygous A at locus M_1 and homozygous B at locus M_2 , it is clear that they carry two haplotypes $\begin{smallmatrix} A \\ B \end{smallmatrix}$. All other probands who are doubly homozygous can be similarly assigned. The haplotypes of probands who are homozygous at one and heterozygous at the other locus can also be determined unambiguously. For instance, the n_{12} probands heterozygous at marker M_1 and homozygous B at marker M_2 carry exactly one haplotype $\begin{smallmatrix} A \\ B \end{smallmatrix}$ and one haplotype $\begin{smallmatrix} a \\ B \end{smallmatrix}$. However, the haplotypes of the n_{11} doubly heterozygous probands are uncertain. Therefore, if haplotypes have not been resolved in the laboratory, haplotype estimation techniques have to be employed to estimate the allelic LD between the two loci. If the assumptions imposed for haplotype estimation do not hold, estimates may be biased.

An alternative to allelic LD measures therefore are genotype-based LD measures (Section 10.2.2) that dispense with the haplotype estimation step [704, 707, 712]. Haplotype estimation approaches will be described in Chapter 13, and for the time being, we assume that phase is known for each individual if we consider allelic LD measures. Throughout, we will use the terminology from Section 5.3. Specifically, we consider two diallelic loci M_1 and M_2 with allele frequencies p_A and $q_A = 1 - p_A$ at locus M_1 and p_B and $q_B = 1 - p_B$ at marker M_2 (see Table 10.2). In the case of LD between a marker and a disease locus, the second locus might then be considered to be the disease locus, so that the alleles are $B = D$ and $b = N$, and $p_B = p_D$ is the frequency of the disease allele D . The frequencies of specific allele combinations, the haplotypes, are represented by p_{11} , p_{12} , p_{21} , and p_{22} (see Table 10.2). In the following, we do not use the simplification from Section 5.3 but a general form for defining different measures of LD.

Table 10.2 Summary of haplotype probabilities at two diallelic genetic markers.

Marker M_1	Marker M_2		Total
	b	B	
a	$p_{11} = (1 - p_A)(1 - p_B) + \mathcal{D}$	$p_{12} = (1 - p_A)p_B - \mathcal{D}$	$q_A = 1 - p_A$
A	$p_{21} = p_A(1 - p_B) - \mathcal{D}$	$p_{22} = p_A p_B + \mathcal{D}$	p_A
Total	$q_B = 1 - p_B$	p_B	1

If the two genetic markers are not in LD, then the frequency p_{22} of the haplotype AB is equal to $p_A p_B$. However, if there is LD, haplotype frequencies are distorted by

$\mathcal{D} = \mathcal{D}_{AB}$, sometimes in the literature denoted by $\Delta = \Delta_{AB}$, which is the covariance between the two genetic diallelic markers:

$$\begin{aligned}\mathcal{D}_{AB} &= p_{22} - p_A p_B = p_{22} - (p_{21} + p_{22})(p_{12} + p_{22}) = p_{22} - p_{22}(p_{22} + p_{12} + p_{21}) - p_{12}p_{21} \\ &= p_{11}p_{22} - p_{12}p_{21}\end{aligned}$$

$\mathcal{D}_{AB}$ can be directly estimated from the observed haplotype frequencies. The estimator $\widehat{\mathcal{D}}_{AB}$ for $\mathcal{D}_{AB}$ has expected value and large sample variance [705]

$$\begin{aligned}\mathbb{E}(\widehat{\mathcal{D}}_{AB}) &= \frac{2n-1}{2n} \mathcal{D}_{AB}, \\ \mathbb{V}ar(\widehat{\mathcal{D}}_{AB}) &= \frac{1}{2n} (p_A q_A p_B q_B + (1-2p_A)(1-2p_B)\mathcal{D}_{AB} - \mathcal{D}_{AB}^2).\end{aligned}$$

The estimator $\widehat{\mathcal{D}}_{AB}$ for $\mathcal{D}_{AB}$ is asymptotically unbiased, and an estimator $\widehat{\mathbb{V}ar}(\widehat{\mathcal{D}}_{AB})$ for $\mathbb{V}ar(\widehat{\mathcal{D}}_{AB})$ is obtained by replacing allele frequencies p_A and p_B and the allelic $\mathcal{D}_{AB}$ by their estimators. Statistical tests for the null hypothesis $H_0: \mathcal{D}_{AB} = 0$ versus $H_0: \mathcal{D}_{AB} \neq 0$ can be constructed as score test or Wald test:

$$T_S = \frac{\widehat{\mathcal{D}}_{AB} - \mathbb{E}_{H_0}(\widehat{\mathcal{D}}_{AB})}{\sqrt{\widehat{\mathbb{V}ar}_{H_0}(\widehat{\mathcal{D}}_{AB})}} = \frac{\widehat{\mathcal{D}}_{AB}}{\sqrt{\frac{1}{2n}(\hat{p}_A \hat{q}_A \hat{p}_B \hat{q}_B)}}, \quad T_W = \frac{\widehat{\mathcal{D}}_{AB} - \mathbb{E}_{H_0}(\widehat{\mathcal{D}}_{AB})}{\sqrt{\widehat{\mathbb{V}ar}(\widehat{\mathcal{D}}_{AB})}}. \quad (10.1)$$

Both T_S and T_W are asymptotically standard normally distributed under H_0 . An asymptotically valid $1 - \alpha$ confidence interval for $\mathcal{D}_{AB}$ can be constructed as

$$\widehat{\mathcal{D}}_{AB} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathbb{V}ar}(\widehat{\mathcal{D}}_{AB})}. \quad (10.2)$$

The values of $\mathcal{D}_{AB}$ depend on the allele frequencies at both loci. Specifically, one obtains (see Problem 10.1)

$$\max\{-p_A p_B, -q_A q_B\} \leq \mathcal{D}_{AB} \leq \min\{p_A q_B, q_A p_B\},$$

so that

$$\mathcal{D}_{\max} \leq \begin{cases} \min\{p_A q_B, q_A p_B\}, & \text{if } \mathcal{D}_{AB} > 0, \\ \max\{-p_A p_B, -q_A q_B\}, & \text{if } \mathcal{D}_{AB} < 0. \end{cases}$$

Figure 10.3 illustrates the dependency of the maximum LD $\mathcal{D}_{\max}$ on the allele frequencies at both loci. The dependence of $\mathcal{D}_{AB}$ on the allele frequencies means that $\mathcal{D}_{AB}$ between two pairs of genetic loci are comparable only if the allele frequencies are similar, which limits the practical use of this measure. This criticism is identical to the concerns raised against the use of the covariance as a measure of association. To surmount this difficulty, different standardizations of $\mathcal{D}_{AB}$ have been suggested [480] that all take $\mathcal{D}_{AB}$ in the numerator. The resulting measures are displayed in Table 10.3 together with the respective denominators used. For the measures δ and y , different formulae are common, as is shown in Table 10.3. The equality of the formulations is given as Problem 10.2, and further characteristics are worked out in

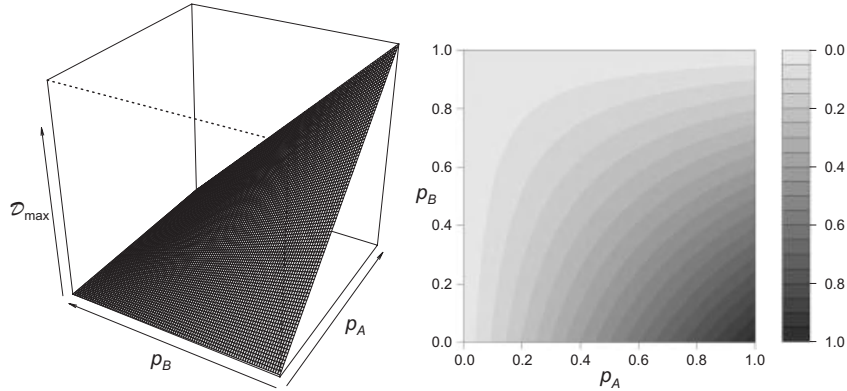


Fig. 10.3 Values of maximum linkage disequilibrium measured by $\mathcal{D}_{\max}$ depend on the allele frequencies p_A and p_B at both loci. In the left part of the figure, the three-dimensional relationship is displayed. Identical gray lines represent identical $\mathcal{D}_{\max}$ values in the right part of the figure. Here, the darker the gray, the higher the maximum LD.

Table 10.3 Linkage disequilibrium measures between two diallelic genetic markers M_1 and M_2 .

Name	Measure	Denominator
Pearson's correlation	r, ϱ, Δ	$\sqrt{p_A q_A p_B q_B}$
Lewontin's $\mathcal{D}'$	$\mathcal{D}'$	$\mathcal{D}_{\max}$
	δ	$p_B p_{11} =$
		$p_B(1 - p_A - p_B + p_A p_B + \mathcal{D}_{AB})$
Frequency difference	f, d	$p_B q_B$
Regression	b, β	$p_A q_A$
Yule's Q	Q, y	$p_{11} p_{22} + p_{12} p_{21} =$
		$2p_A q_A p_B q_B$ $+ \mathcal{D}_{AB}(1 - 2p_A)(1 - 2p_B) + 2\mathcal{D}_{AB}^2$

The numerator is given by $\mathcal{D} = \mathcal{D}_{AB}$ as defined above.

Problem 10.4. If $\mathcal{D}_{\max} = 1$, the LD is said to be *complete*. This is fulfilled, whenever one entry in the 2×2 Table 10.2 of haplotype frequencies is equal to 0.

In addition to these standardizations of the original allele-based $\mathcal{D}_{AB}$, the *coefficient of determination* $\Delta^2 = r^2 = r_{AB}^2$ that is the square of Pearson's product-moment correlation coefficient $\Delta = r = r_{AB}$

$$\Delta_{AB}^2 = r_{AB}^2 = \frac{\mathcal{D}_{AB}^2}{p_A q_A p_B q_B}$$

is often used in applications. Furthermore, classical epidemiological measures have been suggested. For example, the *odds ratio (OR)* and the *population attributable risk (PAR)* have been proposed as measures of LD. They will not be considered here but in the context of association between a phenotype and a marker allele in the next chapter.

A different approach to quantify LD has been taken by Morton and colleagues [439]. As described in Chapter 5, the association probability η is derived from the Malecot model. Under the constraints defined in Chapter 5 and under random sampling, η is identical to $\mathcal{D}'$.

We now illustrate the calculation of different allelic LD measures.

Example 10.1. In the association study by Franjkovic and colleagues [220], the subgroup of 158 subjects with IgE levels ≤ 100 were considered. They were genotyped on a number of SNPs at the IL-4 receptor alpha chain gene *IL4R*. For two SNPs, Q551R (rs1801275) and T1914C, the genotype frequencies are displayed in Table 10.4.

Haplotype estimation is required to determine allelic LD measures if there is at least one subject in the sample that is heterozygous at both SNPs. To avoid this haplotype estimation step, we modify the real data example by ignoring the 21 subjects in the middle of the 3×3 genotype frequency Table 10.4 that have the CT genotype at T1914C and the QR genotype at Q551R. The resulting genotype frequencies are displayed in Table 10.5.

Table 10.4 Genotype frequencies of 158 subjects with IgE levels ≤ 100 from [220] at Q551R (rs1801275) and T1914C from the IL-4 receptor alpha chain gene *IL4R*.

		Q551R			Sum
		RR	QR	QQ	
T1914C	CC	3	21	27	51
	CT	2	21	62	85
	TT	0	2	20	22
	Sum	5	44	109	158

Table 10.5 Genotype frequencies of the 137 subjects with IgE levels ≤ 100 from [220] at Q551R (rs1801275) and T1914C from the IL-4 receptor alpha chain gene *IL4R* after exclusion of the doubly heterozygous subjects.

		Q551R			Sum
		RR	QR	QQ	
T1914C	CC	3	21	27	51
	CT	2	0	62	64
	TT	0	2	20	22
	Sum	5	23	109	158

To establish the allelic LD between the two SNPs, both the allele frequencies p_T at T1914C and p_Q at Q551R, which are the allele frequencies of the T allele and the amino acid Q, respectively, and the haplotype frequencies need to be estimated. Given Hardy–Weinberg equilibrium, allele frequencies can be easily estimated from the genotype frequencies of Table 10.5 and are given by

$$\begin{aligned}\hat{p}_T &= \frac{2 \cdot 22 + 1 \cdot 64}{2 \cdot 137} \approx 0.3942, & \hat{q}_T &= \frac{2 \cdot 51 + 1 \cdot 64}{2 \cdot 137} \approx 0.6058, \\ \hat{p}_Q &= \frac{2 \cdot 109 + 1 \cdot 23}{2 \cdot 137} \approx 0.8796, & \hat{q}_Q &= \frac{2 \cdot 5 + 1 \cdot 23}{2 \cdot 137} \approx 0.1204.\end{aligned}$$

Determining haplotype frequencies from Table 10.5 is simple because the doubly heterozygous subjects are omitted. For example, the haplotype frequency of the TQ haplotype is obtained from the 20 subjects homozygous at both SNPs—they have two copies of this haplotype; the two subjects that are homozygous TT at $T1914C$ and heterozygous QR at $Q551R$, and, finally, the 62 subjects that are heterozygous at $T1914C$ and homozygous QQ at $Q551R$. Thus, $\hat{p}_{TQ} = \frac{2 \cdot 20 + 1 \cdot 2 + 1 \cdot 62}{2 \cdot 137} \approx 0.3796$. Given these frequencies, different LD statistics can now be calculated, for example,

$$\begin{aligned}\hat{D}_{TQ} &\approx 0.3796 - 0.3942 \cdot 0.8796 \approx 0.0329, \\ \hat{D}'_{TQ} &\approx \frac{0.0329}{0.3942 \cdot 0.1204} \approx 0.6926, \\ \hat{\Delta} &\approx \frac{0.0329}{\sqrt{0.3942 \cdot 0.6058 \cdot 0.8796 \cdot 0.1204}} \approx 0.2069.\end{aligned}$$

As a result, $\hat{\Delta}^2 \approx 0.0428$.

10.2.2 How can genotypic linkage disequilibrium be measured?

We now turn to the description of genotype-based measures of LD. Although approaches for estimating genotypic LD were proposed about 30 years ago, they have only rarely been employed in practice. One reason probably is that many genotypic LD coefficients can be defined [704, 705, 707] because the data are not presented in a 2×2 but a 3×3 table. Here, we present two ways of defining genotypic LD. We first consider the composite LD approach of Weir and Cockerham [704, 705, 707]. Second, we describe our recent r^2 method to genotypic LD [712].

Consider the number of A alleles and B alleles at markers M_1 and M_2 , and let π_{ij} , $i, j = 0, 1, 2$, denote the population frequencies of the 9 possible genotype combinations of the number of M_1 and M_2 alleles (Table 10.6). Analogously, let n_{ij} denote the observed frequencies of the genotype combinations.

Table 10.6 Joint genotype probabilities at two diallelic markers with, respectively, X and Y denoting the number of A alleles at M_1 and B alleles at M_2 .

		Y			Total
		0	1	2	
X	0	π_{00}	π_{01}	π_{02}	π_{0+}
	1	π_{10}	π_{11}	π_{12}	π_{1+}
	2	π_{20}	π_{21}	π_{22}	π_{2+}
	Total	π_{+0}	π_{+1}	π_{+2}	1

When SNPs at two loci are considered, the two double heterozygotes $\frac{A}{B} \mid \frac{a}{b}$ and $\frac{A}{b} \mid \frac{a}{B}$ cannot be distinguished so that allelic LD cannot be inferred directly. Weir and Cockerham [704, 707] therefore proposed the *composite LD* as

$$\mathcal{D}_{XY} = 2P\left(\frac{A}{B} \mid \frac{A}{B}\right) + P\left(\frac{A}{B} \mid \frac{A}{b}\right) + P\left(\frac{A}{B} \mid \frac{a}{B}\right) + \frac{1}{2}\left(P\left(\frac{A}{B} \mid \frac{a}{b}\right) + P\left(\frac{A}{b} \mid \frac{a}{B}\right)\right) - 2p_A p_B. \quad (10.3)$$

$\mathcal{D}_{XY}$ of eq. (10.3) is also termed *digenic LD*. It can be estimated by $\widehat{\mathcal{D}}_{XY} = \frac{1}{n}(2n_{00} + n_{01} + n_{10} + \frac{1}{2}n_{11}) - 2\hat{p}_A \hat{p}_B$. Weir and Cockerham [704, 705, 707] propose a series of other higher order genotypic LD measures, including a quadrigenic LD coefficient. They also provide the variance of the composite LD although the variance is not of a standard form. Finally, the authors propose a score test using the composite LD. Although these genotypic LD measures are well founded using population genetic theory, their use in practice is limited because no relation has been established to other LD measures (Table 10.3).

We [712] have therefore proposed a different measure of genotypic LD using the following arguments. A reasonable assumption for a measure of genotypic LD is that it should remain invariant under the transition from haplotypes to diploid genotypes if Hardy–Weinberg equilibrium (HWE) is fulfilled. Phrased differently, if HWE can be taken for granted, the LD measure should not change its value when computed from genotype rather than haplotype frequencies. We consider this assumption to be reasonable because most haplotype estimation approaches assume HWE (Chapter 13), and haplotype estimation is straightforward when this assumption holds.

In fact, Lewontin's $\mathcal{D}'$ fails to remain invariant if haplotype frequencies are replaced with HWE consistent genotype frequencies (see Problem 10.5). However, the Pearson product-moment correlation coefficient $\Delta = \Delta_{XY}$ based on the 3×3 table of genotypes from the two SNPs (Table 10.6) fulfills this property (see Problem 10.6).

The asymptotic variance of the estimated correlation coefficient $\widehat{\Delta}$ can be estimated by

$$\widehat{\text{Var}}(\widehat{\Delta}) = \frac{1}{n} \sum_{i=0}^2 \sum_{j=0}^2 \hat{\zeta}_{ij}^2 \hat{\pi}_{ij} - \frac{1}{n} \left(\sum_{i=0}^2 \sum_{j=0}^2 \hat{\zeta}_{ij} \hat{\pi}_{ij} \right)^2,$$

which has the clear structure of a variance. This variance can be derived using the multivariate delta method, i.e., the same technique as used in Solution 4.9.

The terms ζ_{ij} can be estimated by

$$\hat{\zeta}_{ij} = \frac{ij - i\bar{y} - j\bar{x}}{s_x s_y} - \frac{1}{2}\widehat{\Delta} \left(\frac{i(i - 2\bar{x})}{s_x^2} + \frac{j(j - 2\bar{y})}{s_y^2} \right),$$

where $\bar{x}$ and $\bar{y}$ are the means of the two SNPs, and s_x^2 and s_y^2 denote their variances.

Because the correlation coefficient is limited to the interval $[-1; 1]$, an asymptotically valid confidence interval should preferably be derived using Fisher's z transformation. To this end, let

$$\hat{\gamma} = \frac{1}{2} \ln \frac{1 + \widehat{\Delta}}{1 - \widehat{\Delta}}$$

be the estimated Fisher z transformed correlation coefficient. Using the delta method, the variance of $\hat{\gamma}$ can be estimated by

$$\widehat{\text{Var}}(\hat{\gamma}) = \frac{\widehat{\text{Var}}(\hat{\Delta})}{(1 - \hat{\Delta}^2)^2}.$$

An asymptotically valid confidence interval for γ at confidence level $1 - \alpha$ is therefore given by

$$\underline{\gamma}_{(1-\alpha)} = \hat{\gamma} - z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{\gamma})}, \quad \bar{\gamma}_{(1-\alpha)} = \hat{\gamma} + z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{\gamma})},$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $\alpha/2$ quantile of the standard normal distribution. The inverse Fisher z transformation results in an asymptotically valid $1 - \alpha$ confidence interval for Δ

$$\underline{\Delta}_{(1-\alpha)} = \tanh \underline{\gamma}_{(1-\alpha)}, \quad \bar{\Delta}_{(1-\alpha)} = \tanh \bar{\gamma}_{(1-\alpha)}.$$

A series of pathological cases need to be considered separately, and solutions to these are available [712].

An alternative to the described asymptotic approach is to use the exact distribution [705, 712]. However, the variance approximation to the genotypic LD as measured by the conventional correlation coefficient Δ is sufficiently accurate at a sample size of $n = 60$, and it may be used for constructing asymptotically valid confidence intervals [712]. A corresponding Problem 10.9 is given below.

We illustrate the calculation of the genotypic LD Δ in the following example.

Example 10.2. Consider the subgroup of 158 subjects with IgE levels ≤ 100 from the association study by Franjkovic and colleagues [220]. For the two SNPs Q551R (rs-1801275) and T1914C from the IL4R gene, the genotype frequencies are displayed in Table 10.4.

We estimate the joint genotype frequencies and obtain for $i, j = 0$ to 2

$$[\hat{\pi}_{ij}]_{i,j=0}^2 = \begin{pmatrix} 0.01899 & 0.1329 & 0.1709 \\ 0.01266 & 0.1329 & 0.3924 \\ 0.00000 & 0.0127 & 0.1266 \end{pmatrix}.$$

The marginal genotype frequencies are

$$\begin{aligned} [\hat{\pi}_{i+}]_{i=0}^2 &= (0.3228, 0.5380, 0.1392)', \\ [\hat{\pi}_{+j}]_{j=0}^2 &= (0.0316, 0.2785, 0.6899)'. \end{aligned}$$

From these, we obtain

$$\begin{aligned} \bar{x} &= \sum_{i=0}^2 i \hat{\pi}_{i+} = 0.8165, & s_x^2 &= \left(\sum_{i=0}^2 i^2 \hat{\pi}_{i+} \right) - \bar{x}^2 = 0.4283, \\ \bar{y} &= \sum_{j=0}^2 j \hat{\pi}_{+j} = 1.6582, & s_y^2 &= \left(\sum_{j=0}^2 j^2 \hat{\pi}_{+j} \right) - \bar{y}^2 = 0.2883. \end{aligned}$$

The covariance s_{xy} and Pearson correlation coefficient $\hat{\Delta}$ are estimated to be

$$s_{xy} = \left(\sum_{i=0}^2 \sum_{j=0}^2 i j \hat{\pi}_{ij} \right) - \bar{x} \bar{y} = 0.0955 \quad \hat{\Delta} = \frac{s_{xy}}{s_x s_y} = 0.2718.$$

Next, we calculate $\hat{\zeta}_{ij}$ as

$$\left[\hat{\zeta}_{ij} \right]_{i,j=0}^2 = \begin{pmatrix} 0.0000 & -1.2315 & -3.4059 \\ -4.5184 & -2.9040 & -2.2325 \\ -9.6712 & -5.2109 & -1.6935 \end{pmatrix},$$

so that we obtain $\widehat{\text{Var}}(\hat{\Delta}) = 0.0046$. The estimated z transformed correlation coefficient $\hat{\gamma}$ and its estimated variance can be obtained as $\hat{\gamma} = 0.2788$ and $\widehat{\text{Var}}(\hat{\gamma}) = 0.0054$. An asymptotically valid 95% confidence interval for γ is obtained as $[0.1354; 0.4222]$. With the inverse Fisher z transformation, we finally get the asymptotically valid 95% confidence interval for Δ as $[0.1346; 0.3988]$.

In summary, we have $\hat{\Delta} = 0.2718$ with an asymptotically valid 95% confidence interval for Δ of $[0.1346; 0.3988]$.

Given the different proposed measures for allelic and genotype-based LD, an obvious question arises which measure to use in practice. Different studies have been carried for allelic LD measures. In general, the conclusions were in favor of η , $\mathcal{D}'$, and Δ . More specifically, Devlin and Risch [161] have shown that $\mathcal{D}'$ and Δ have a unimodal and symmetric distribution, with the maximum at the disease locus itself, that is not relevantly affected by population genetic effects. Under random sampling as in a cohort study, both measures are directly related to the recombination fraction. In addition, if the disease is rare, Δ approaches $\mathcal{D}'$ as with $p_B \rightarrow 0$, $p_{11} \approx q_A$. Furthermore, in a cohort study, $\mathcal{D}'$ is equivalent to η [137]. Comparing the error variance of the different measures, Morton and colleagues [482] have shown that η is the most efficient measure, with δ coming second.

All other allelic LD measures suffer from the disadvantage of depending on allele frequencies. In addition, the comparisons by Devlin and Risch [161] showed that in a model population with a disease and a number of marker loci segregating for 100 generations, the maxima of these statistics are not at the disease locus, but that their distributions are multimodal; this drawback even worsens under population genetic effects.

The next natural question is whether Δ or $\mathcal{D}'$ should be used. $\mathcal{D}'$ has two major shortcomings. First, it indicates complete LD, i.e., $\mathcal{D}' = 1$, whenever a single cell in the 2×2 table of haplotype frequencies is 0 (see Problem 10.4.2), irrespective of the amount of probability mass assigned to the other diagonal of the table. $\mathcal{D}'$ also fails to be invariant when haplotype frequencies are replaced by HWE consistent genotype frequencies. The composite LD estimated of Eq. (10.3) fails to have a natural statistical interpretation. Therefore, Δ is the only estimator with appropriate statistical properties. Nevertheless, $\mathcal{D}'$ can be adequately interpreted in specific genetic applications, e.g., the Malecot model (see Chapter 5).

The last question that needs to be addressed is whether the allelic or genotypic Δ should be used. The answer is simple if haplotypes have been determined in the laboratory. In this case, a loss in precision of the Δ estimates may occur, and the haplotype-based estimator is double asymptotic efficient in the extreme situation when there is no LD. The difference between the two estimators is, however, low when $\Delta \geq 0.8$ [712]. This means that for fine-mapping purposes the genotypic LD is appropriate even if gametic information is available. However, haplotypes are usually not determined in the laboratory but need to be inferred from genotype data. In this case, the corresponding allelic LD estimator of Δ has roughly the same standard error as the genotypic LD estimator that is directly calculated from the joint distribution of genotypes. Thus, if only genotypes are available, the use of the genotypic LD Δ is recommended for use in routine instead of the allelic LD.

Different implementations are available for estimating LD measures from a genotyped sample. For example, the package `GENASSOC` is based on `STATA/SE`, can be downloaded, and provides estimates for a variety of allelic LD measures. Also, the package `GENETICS` for `R` estimates allelic $\mathcal{D}$, $\mathcal{D}'$, and Δ from genotype data. Once LD has been estimated, it can be plotted in heat maps using, e.g., `GOLD` or `HAPLOVIEW` (for an example, see Chapter 5). So far, we have described how LD can be measured between a pair of alleles or genotypes at two SNPs. Extensions of allelic LD measures to multiallelic markers and multiple loci have been proposed [426, 500, 507, 705, 706]. The interested reader may refer to the literature for details.

10.2.3 How far does linkage disequilibrium extend?

Thinking back to the causal models for association in Figure 10.1, we recall that in a study based on indirect association, our success depends on the LD between the marker and the disease locus. When markers are selected for the study, it has to be ensured that these will be close enough to the disease-causing locus to be in LD with it. But how close is close enough? Or, more elaborately, how far, on a physical distance scale, does LD extend?

The LD in a given region depends on a number of factors. This is best described by envisaging the introduction of a novel mutation in a population. When a new mutation occurs, it does so necessarily on the background of a specific haplotype. Because this is the only haplotype with the mutation, there is perfect LD between the mutation and the other alleles. In successive generations, however, the mutation is separated from the other alleles by recombination events. Hence, the LD between the mutation and a certain other allele on the haplotype will decay with time, and the speed of decay depends on the recombination fraction between the two loci (see Problem 10.7). If the recombination fraction θ is close to 0, the decline will be slow; if it is close to 0.5, LD will rapidly decrease to 0, indicating a linkage equilibrium.

To illustrate this, let t be the number of generations after the mutation has entered the population and $\mathcal{D}_0$ the original level of LD at the time point of introduction; further, the level of LD after t generations $\mathcal{D}_t$ is given by $\mathcal{D}_t = (1 - \theta)^t \cdot \mathcal{D}_0$. For example, if there were a low recombination fraction of 0.05, the LD would be halved after about 15 generations. This is depicted in Figure 10.4 for different values of θ . As

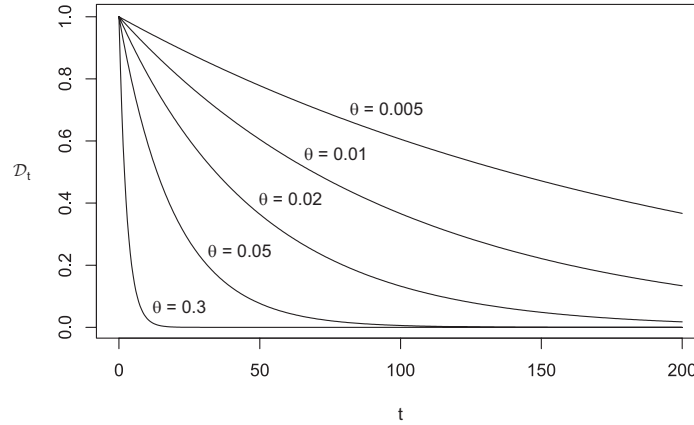


Fig. 10.4 Decay of linkage disequilibrium measured by $\mathcal{D}_t$ in successive generations t depending on the recombination fraction θ .

a result of this process, high LD between a mutation and another allele might be either due to a low recombination fraction between them or due to the mutation being rather young. In addition, other population genetic effects influence LD, such as admixture between populations, selection, and the overall population history [30, 385].

Based on the relationship between distance and time in generations, it might be expected that, in empirical investigations, there should be a straightforward relationship between LD and genetic or physical distance (also see Chapter 5). Indeed, the traditional rule of thumb resulting from this was that an association is detectable as long as the distance is not greater than 1 megabase (Mb) $\approx \theta = 0.01 \approx 1$ centiMorgan (cM).

However, Figure 10.5 illustrates that in practice, this is not the case. In this example, 12 SNPs were genotyped in the interleukin-21 receptor (*IL21R*) gene and in the interleukin-4 receptor α chain gene (*IL4RA*) on chromosome 16p12 in a sample of 304 healthy blood donors with Caucasian origin [290]. The relationship between pairwise LD measured as $\mathcal{D}'$ and physical distance in this case is clearly not monotonic. This result coincides with other studies on different chromosomal regions in the human genome. Here, the general pattern has been that with an increasing physical distance, LD also decreases as expected. However, in some cases the LD was very weak, even over very short distances; in others, it stretched for more than 100 kilobases (Kb) or even 1Mb.

In addition, it has become clear that the extent of LD strongly depends on the chromosomal region. Specifically, because of higher recombination fractions near telomeres, LD decays more rapidly with physical distance in telomeric regions compared to others [30, 694]. Also, LD strongly depends on a number of population genetic factors, and the influence of these varies across the genome. With regard to different populations, it is usually found that LD extends for much shorter distances in African populations than in Caucasian populations (see, e.g., Ref. [556]).

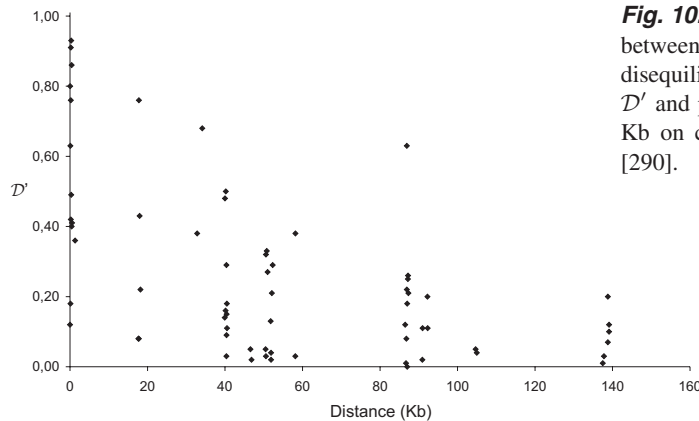


Fig. 10.5 Relationship between pairwise linkage disequilibrium measured as D' and physical distance in Kb on chromosome 16p12 [290].

The general consequence of these empirical results is that LD and, through this, association are not a monotonic function of the distance. Similarly, the significance of LD is irregular, and LD is not predictable from one region or one population to another.

Once it was established that the relationship between LD and distance is not as straightforward as previously thought, the natural next step was to study the empirical pattern in more detail. Consequently, there was a shift from trying to quantify levels of LD to studying the structure of LD instead. After several studies had been performed in different genomic regions, a specific pattern emerged. In the investigated areas, there were regions of strong LD that were bordered by regions with almost equilibrium [152, 156, 233, 520, 556]. The regions with strong LD, and hence only a few recombinations, were henceforth termed *haplotype blocks*, and the interspersed areas with many recombinations were called *recombination hot spots*. Within the haplotype blocks, it was found that there was a limited diversity of haplotypes, with often only four to six haplotypes making up most of the variability. The length of the haplotype blocks was variable, with some extending more than several hundred Kb. The recombination fraction in the recombination hot spots was estimated to be up to 1000 times higher than within other regions.

Following these initial observations, further analyses found that the length of the haplotype blocks depends on the study design. Specifically, with a higher density of genotyped markers, longer blocks are split and the average block length becomes shorter [673]. Also, regions that previously showed no block structure do become blocked. In addition, the sample size is a crucial issue: if too few individuals are analyzed, rare haplotypes might not be detected. Therefore, it has been suggested as a rule of thumb that at least 100 chromosomes need to be typed. Finally, haplotype blocks are usually defined via pairwise LD measures, and as these are affected by genotyping errors, mutations, and population genetic factors, as will the resulting haplotype blocks. We want to stress that several problems exist with the definition

of haplotype blocks, and the reader may refer to a thorough review by Collins and colleagues [138].

10.3 PROBLEMS

Problem 10.1 Show $\max(-p_A p_B, -q_A q_B) \leq \mathcal{D}_{AB} \leq \min(p_A q_B, q_A p_B)$.

Problem 10.2 Confirm the following equalities stated in Table 10.3:

Problem 10.2.1. $\delta = \frac{\mathcal{D}_{AB}}{p_B p_{22}} = \frac{\mathcal{D}_{AB}}{p_B(1 - p_A - p_B + p_A p_B + \mathcal{D}_{AB})}$

Problem 10.2.2. $y = \frac{\mathcal{D}_{AB}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\mathcal{D}_{AB}}{2p_A q_A p_B q_B + \mathcal{D}_{AB}(1 - 2p_A)(1 - 2p_B) + 2\mathcal{D}_{AB}^2}$

Problem 10.3 We use the notation of Section 10.2.1. The most common measure of association for cohort and case-control studies is the *odds ratio (OR)* that can be defined via the *odds*. Specifically, the *odds* for the disease allele $B = D$ if one carries allele A or allele a at locus M_1 is defined by

$$\text{odds}_A = \frac{p_{22}}{p_A} \bigg/ \left(1 - \frac{p_{22}}{p_A}\right) = \frac{p_{22}}{p_{21}}, \quad \text{odds}_a = \frac{p_{12}}{q_A} \bigg/ \left(1 - \frac{p_{12}}{q_A}\right) = \frac{p_{12}}{p_{11}},$$

and the *odds ratio (OR)* is defined as

$$OR = \frac{\text{odds}_A}{\text{odds}_a} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Show that Yule's Q , as given in Table 10.3, can also be expressed in terms of the OR :

$$y = (OR - 1) / (OR + 1).$$

Problem 10.4 Confirm the following characteristics of the allele-based $\mathcal{D}'_{AB}$ and Δ_{AB}^2 :

Problem 10.4.1. Show that even if $\mathcal{D}'_{AB} = 1$, Δ_{AB}^2 can be equal to 1 only if the allele frequencies at the markers M_1 and M_2 , p_A and p_B , are equal.

Problem 10.4.2. Show that $\mathcal{D}'_{AB}$ can be equal to 1 with different allele frequencies p_A and p_B if the disease allele A occurs only in the presence of the marker allele B .

Problem 10.5 Give an example that Lewontin's $\mathcal{D}'$ is not identical when computed from genotype frequencies or from haplotype frequencies if HWE holds.

Problem 10.6 Show that the Pearson product-moment correlation coefficient Δ is identical when computed from genotype frequencies or from haplotype frequencies if HWE holds. For convenience, the joint genotype distribution at markers M_1 and M_2 is given in Table 10.7.

Problem 10.7 Show that the LD decays as a function of the recombination fraction. Specifically, show $\mathcal{D}_{t+1} = (1 - \theta)\mathcal{D}_t$.

Table 10.7 Joint genotype distribution obtained from the haplotype frequencies shown in Table 10.2 under Hardy–Weinberg equilibrium with, respectively, X and Y denoting the number of A alleles at M_1 and B alleles at M_2 .

		Y		
		0	1	2
X	0	p_{11}^2	$2p_{11}p_{12}$	p_{12}^2
	1	$2p_{11}p_{21}$	$2(p_{11}p_{22} + p_{12}p_{21})$	$2p_{12}p_{22}$
	2	p_{21}^2	$2p_{21}p_{22}$	p_{22}^2

Problem 10.8 In a small study investigating the genetic background of asthma, a group of 8 affected subjects were recruited and genotyped for two SNPs in a candidate gene on chromosome 5. As a comparison, data from 7 healthy controls were utilized, who were also genotyped at the two SNPs.

Table 10.8 Genotyping results from 8 asthma cases and seven healthy controls genotyped at two single nucleotide polymorphisms (SNPs).

Case individuals	SNP 1	SNP 2	Control individuals	SNP 1	SNP 2
1	AA	CT	9	GG	CT
2	GA	CT	10	GA	CC
3	GA	TT	11	GA	CC
4	GG	CT	12	GG	CT
5	AA	CT	13	AA	TT
6	GA	CC	14	GA	CT
7	GA	CT	15	GG	CT
8	GG	CC			

The results are shown in Table 10.8. At SNP 1, allele G is the wild type, and at SNP 2, allele C is the wild type.

Problem 10.8.1. Determine the genotype and allele frequencies at both SNPs for the entire group. For this, assume that Hardy–Weinberg equilibrium holds.

Problem 10.8.2. As described in Section 10.2.1, allelic LD can be estimated only if the haplotypes are known. For which individuals can the haplotypes be determined unambiguously from the genotypes? If all others were simply excluded from the calculations, what would be the haplotype frequencies?

Problem 10.8.3. Disregard the double heterozygous subjects, create the 2×2 table of gametes, and estimate the allelic LD measures $\mathcal{D}$, $\mathcal{D}'$, and Δ^2 between SNPs 1 and 2.

Problem 10.8.4. Disregard the double heterozygous subjects, use the asymptotic score and Wald test statistics of Eq. (10.1), and test the null hypothesis of no LD.

Problem 10.8.5. Disregard the double heterozygous subjects and estimate the asymptotic 95% confidence interval for $\mathcal{D}_{AB}$ of Eq. (10.2).

Problem 10.9 This problem is based on the freely available data of the Cardiogenomics project. Specifically, we took SNPs rs8077276 and rs4343 of the

angiotensin converting enzyme *ACE* gene, a candidate gene on chromosome 17q23 for cardiovascular diseases. Detailed descriptions of the project, the data set and the *ACE* gene can be found on the project web page. The joint frequency table of the genotypes at the two SNPs is given in Table 10.9.

Problem 10.9.1. Estimate the genotypic LD Δ .

Problem 10.9.2. Estimate its variance $\text{Var}(\hat{\Delta})$.

Problem 10.9.3. Construct an asymptotically valid 95% confidence interval for the genotypic LD Δ .

Table 10.9 Genotyping results at rs8077276 and rs4343 of the angiotensin converting enzyme *ACE* gene from the Cardiogenomics project.

rs8077276	rs4343			Sum
	AA	AG	GG	
AA	29	0	0	29
AG	10	18	0	28
GG	3	9	8	20
Sum	42	27	8	77

URLs

Cardiogenomics

<http://cardiogenomics.med.harvard.edu>

GENASSOC for STATA/SE

<http://www-gene.cimr.cam.ac.uk/clayton/software/stata>

GENETIC for R

<http://cran.r-project.org/web/packages/genetics/index.html>

GOLD

<http://www.sph.umich.edu/csg/abecasis/GOLD/>

R

<http://www.r-project.org>

SNP_CORR

http://zihost5.zi-mannheim.de/SNP_CORR/corr_count.sas

STATA

<http://www.stata.com>

11

Association Analysis with Unrelated Individuals

In the previous chapter, we described the principle of genetic association and the concepts and measures of linkage disequilibrium (LD). We have thus concerned ourselves with the relationship between the alleles at two genetic loci. The starting point for this had been that in the model of indirect association, alleles at the disease and the marker locus have to be in LD for an association between the marker allele and the disease status to be observed.

We will now turn to the analysis of the association itself and focus on studies utilizing data from unrelated individuals in a case-control or a cohort study design. Before the methods for comparing individuals with and without the respective phenotype are described in Section 11.2, we will consider the more practical issue of which individuals should be recruited for the study in Section 11.1. Although the definition of the cohort for a cohort study is as crucial and complex as the definition of the case and the control group(s) in case-control studies, most investigators prefer using the cheaper and simpler case-control design. Therefore, Section 11.2 relates only to case-control studies.

Section 11.3 deals with the calculation of required sample sizes in case-control studies and discusses the effect of the different causal models leading to genetic association on the required sample size.

The ascertainment of cases and controls for an association study can result in substantial bias. An often discussed bias, known as *population stratification*, has long led to case-control studies being shunned in favor of family-based association studies (see Chapter 12). Modern approaches for dealing with the problem of population stratification will be discussed in Section 11.4.

At the end of this chapter, we describe gene-gene and gene-environment interactions in Section 11.5.

11.1 HOW SHOULD CASES AND CONTROLS BE SELECTED?

The selection of individuals for an association study should be guided by the overall aim of the study. Basically, case-control and cohort studies are performed in three stages that can be distinguished. In the first stage, one is interested in identifying genetic variants that are associated with the disease. Hence, one is interested in selecting cases and controls so as to maximize the effect of the genetic variant. In the second stage, investigators may be interested in how different genetic and/or environmental factors act together and form the basis for a prognosis. Here, population-based studies may be preferred to the analysis of extreme groups. The final stage is concerned with an exact quantification of the genetic risk of a genetic locus or a specific variant. Because adjustments for sample selection are problematic, cohort studies are the design of choice. In the following, we focus on the first step, where a sample selection is performed to enrich the study groups for the disease-causing variant.

This can be achieved for variants with a major effect, for instance, by going for patients with more affected relatives, especially if the specific allele is rare in the overall population [480, 730]. Other strategies to enrich for the susceptibility allele include predominantly sampling for patients with unusual characteristics such as an early age of onset, more severe affection, the generally less affected gender, or with a reduced environmental risk. On the other hand, a control phenotype might be selected that is somewhat the antithesis of the affection status. For example, if one is looking for genes with major effects, it might be prudent to select cases with a positive family history and compare these to healthy controls with a negative family history. Alternatively, if the interest is in genes acting protectively, affected cases might be contrasted with controls who are healthy but have been heavily exposed. More examples for this have been given by Morton and Collins [480]. Lacking such a specific hypothesis, the controls should be population-based matched to the cases by the most important confounding factors.

It has been suggested that for the detection of genes underlying complex diseases, sampling of probands from specific populations might be advantageous. Here, recent genetic founder populations or isolated populations have been heralded. For arguments on this, the interested reader may refer to Ref. [730].

11.2 HOW CAN GENETIC ASSOCIATION BE TESTED?

Because even in a cohort study, affected individuals eventually will be compared with the unaffected, for simplicity we will use the terms cases and controls regardless of the study design employed in this section and Section 11.3. In those instances where the study design makes a difference with regard to the analysis, we will point this out. In this section we introduce a series of different test statistics for association analysis (Section 11.2.1). When asking two different genetic epidemiologists about the test that should be chosen for association analysis, you might end up with five different recommendations. We therefore discuss the choice of an appropriate test statistic in Section 11.2.2. No statistical test should be reported in clinical trials without a

corresponding effect estimate and an appropriate confidence interval for the underlying effect measure. And this should be identical for genetic epidemiological studies. We therefore describe different impact measures, i.e., classical epidemiological effect measures in Section 11.2.3. In many cases, the underlying genetic model is unknown, and several approaches for selecting the most likely genetic model have been proposed. These are presented in Section 11.2.4. Finally, we discuss in Section 11.2.5 how association on the X chromosome can be tested.

11.2.1 What are common association tests?

The starting point for tests on genetic association is that a sample of cases and controls has been genotyped at a genetic marker. When considering a diallelic marker locus with alleles A and a , with a being the wild-type and A the mutated variant, the notation that will be used is shown in Table 11.1. Furthermore, we denote the allele frequency of the A allele by $p_{A,a}$ and $p_{A,u}$ in affected cases and unaffected controls, respectively (see Table 11.2).

Table 11.1 Observed genotype frequencies and theoretical probabilities in parenthesis for affected cases (a) and unaffected controls (u) at a diallelic marker with alleles a and A .

	aa	aA	AA	Total
Cases	r_0 ($p_{0,a}$)	r_1 ($p_{1,a}$)	r_2 ($p_{2,a}$)	r (p_a)
Controls	s_0 ($p_{0,u}$)	s_1 ($p_{1,u}$)	s_2 ($p_{2,u}$)	s (p_u)
Total	n_0	n_1	n_2	n (1)

Table 11.2 Observed allele frequencies and theoretical probabilities in parenthesis for cases and controls at a diallelic marker.

	a	A	Total
Cases	$2r_0 + r_1$	$2r_2 + r_1$ ($p_{A,a}$)	$2r$
Controls	$2s_0 + s_1$	$2s_2 + s_1$ ($p_{A,u}$)	$2s$
Total	$2n_0 + n_1$	$2n_2 + n_1$	$2n$

In testing whether a marker allele is associated with a phenotype, one can test independence of the phenotype either from the genotypes or from the alleles. The test for alleles is based on the 2×2 table of observed allele frequencies shown in Table 11.2.

The standard χ^2 statistic is of the form $\sum \frac{(o-e)^2}{e}$, where o denotes “observed,” e “expected,” and $\sum$ is the sum over all table entries. The expected absolute frequencies e can be calculated from the marginal frequencies if these are assumed to be fixed. If the alleles are independent from the case-control status, the proportion of cases in the A column should be identical to the proportion of cases across the two alleles. For example, e_2 is the proportion of cases $2r/2n$ in the column of allele A that has

total frequency $2n_2 + n_1$. Subsequently, $e_2 = (2n_2 + n_1)\frac{2r}{2n}$. The other expected frequencies are obtained analogously and displayed in Table 11.3.

Table 11.3 Observed (o) and expected (e) absolute frequencies in cases and controls for the allelic test at a diallelic marker.

	Observed		Expected	
	a	A	a	A
Cases	$o_1 = 2r_0 + r_1$	$o_2 = 2r_2 + r_1$	$e_1 = \frac{(2n_0+n_1)2r}{2n}$	$e_2 = \frac{(2n_2+n_1)2r}{2n}$
Controls	$o_3 = 2s_0 + s_1$	$o_4 = 2s_2 + s_1$	$e_3 = \frac{(2n_0+n_1)2s}{2n}$	$e_4 = \frac{(2n_2+n_1)2s}{2n}$

The test of association can be performed by using the classical χ^2 test statistic (c for classical)

$$\begin{aligned}\chi_c^2 &= \sum_{k=1}^4 \frac{(o_k - e_k)^2}{e_k} \\ &= \frac{(2n(2r_0 + r_1) - (2n_0 + n_1)2r)^2}{2n(2n_0 + n_1)2r} + \frac{(2n(2r_2 + r_1) - (2n_2 + n_1)2r)^2}{2n(2n_2 + n_1)2r} \\ &\quad + \frac{(2n(2s_0 + s_1) - (2n_0 + n_1)2s)^2}{2n(2n_0 + n_1)2s} + \frac{(2n(2s_2 + s_1) - (2n_2 + n_1)2s)^2}{2n(2n_2 + n_1)2s},\end{aligned}\quad (11.1)$$

which is asymptotically χ^2 distributed with one degree of freedom (d.f.) under the null hypothesis of independence (Problem 11.1). The null hypothesis of no association is rejected for large values of the χ^2 test statistic. χ_c^2 does not reflect the strength of association as measured, e.g., by the *odds ratio* (OR) (see Problem 10.2.2) or the covariance in the 2×2 table. The covariance is well reflected by the determinant of the frequencies (Table 11.3)

$$o_1 o_4 - o_2 o_3 = (2r_0 + r_1)(2s_2 + s_1) - (2r_2 + r_1)(2s_0 + s_1). \quad (11.2)$$

Subsequently, tests for independence that resemble the degree of association via Eq. (11.2) are appealing. In fact, it is possible to show (Problem 11.2) that χ_c^2 is equivalent to

$$\chi_A^2 = 2n \cdot \frac{[(2r_0 + r_1)(2s_2 + s_1) - (2r_2 + r_1)(2s_0 + s_1)]^2}{2r \cdot 2s \cdot (2n_0 + n_1) \cdot (2n_2 + n_1)}, \quad (11.3)$$

which takes the sample size $2n$ multiplied by the determinant from Table 11.2 as numerator and the product of the marginal frequencies as denominator.

If the interest is in a comparison of genotype instead of allele frequencies based on the originally observed frequencies in Table 11.1, several alternatives exist. A standard χ^2 test for independence again might be employed. In this case, the observed genotype frequencies are compared with the expected genotype frequencies under the null hypothesis of no association. These are calculated as follows. If genotypes are

independent from the case-control status, the proportion of cases with the aa genotype should be identical to the proportion of cases across the different genotypes. With the notation from Table 11.1, the expected number of cases with aa genotype therefore is $r_0^e = n_0 \cdot p_a$, which is identical to $n_0 \cdot \frac{r}{n}$ if all marginal frequencies are fixed. Utilizing these and the superscript e for *expected*, the χ^2 test statistic is given by

$$\chi_g^2 = \frac{(r_0 - r_0^e)^2}{r_0^e} + \frac{(r_1 - r_1^e)^2}{r_1^e} + \frac{(r_2 - r_2^e)^2}{r_2^e} + \frac{(s_0 - s_0^e)^2}{s_0^e} + \frac{(s_1 - s_1^e)^2}{s_1^e} + \frac{(s_2 - s_2^e)^2}{s_2^e},$$

which is, under the null hypothesis, asymptotically χ^2 distributed with 2 d.f. Again, the null hypothesis of no association is rejected for large values of the χ^2 test statistic. One drawback of this simple approach is that it tests the null hypothesis that there is no difference in the genotype frequencies against the alternative that there are genotype frequency differences with a 2 d.f. test. However, in most situations, more specific alternative hypotheses can be formulated that are based on different genetic models. If a model with a dominant or a recessive mode of inheritance is assumed, or if there is heterozygote disadvantage/advantage, i.e., excess heterozygosity or heterozygote deficiency, the frequencies from Table 11.1 can be collapsed to 2×2 tables as shown in Table 11.4. Here, for a dominant model with regard to allele A , genotypes with one or two A alleles are subsumed (Table 11.4, left), whereas for a recessive model, we combine genotypes with zero or one A allele (Table 11.4, right). Finally, for diseases where selection of either homozygotes or heterozygotes plays an important role, we single out heterozygous subjects and combine genotypes with zero or two A alleles (Table 11.4, middle).

Table 11.4 Genotype frequencies in cases and controls for the dominant, the recessive, and the heterozygote disadvantage/advantage model with regard to allele A .

	Dominant		Heterozygote		Recessive	
	aa	aA or AA	aa or AA	aA	aa or aA	AA
Cases	r_0	$r_1 + r_2$	$r_0 + r_2$	r_1	$r_0 + r_1$	r_2
Controls	s_0	$s_1 + s_2$	$s_0 + s_2$	s_1	$s_0 + s_1$	s_2
Total	n_0	$n_1 + n_2$	$n_0 + n_2$	n_1	$n_0 + n_1$	n_2

Test statistics for these specific genetic models are analogously defined to the χ^2 statistic of Eq. (11.3). Thus, χ_{dom}^2 , χ_{rec}^2 , and χ_{het}^2 can be calculated for the dominant model, the recessive model, and the model with heterozygote disadvantage/advantage, respectively, by

$$\begin{aligned} \chi_{dom}^2 &= n \cdot \frac{(r_0(s_1 + s_2) - (r_1 + r_2)s_0)^2}{r \cdot s \cdot n_0 \cdot (n_1 + n_2)}, \\ \chi_{rec}^2 &= n \cdot \frac{((r_0 + r_1)s_2 - r_2(s_0 + s_1))^2}{r \cdot s \cdot (n_0 + n_1) \cdot n_2}, \\ \chi_{het}^2 &= n \cdot \frac{(r_1(s_0 + s_2) - (r_0 + r_2)s_1)^2}{r \cdot s \cdot n_1 \cdot (n_0 + n_2)}. \end{aligned} \quad (11.4)$$

All three tests are asymptotically χ^2 distributed with one d.f. under the null hypothesis of no association, and the null hypothesis of no association is rejected for large values of the χ^2 test statistic.

Another possibility to reduce the 2 d.f. goodness-of-fit test to a 1 d.f. test is to assume an additive or multiplicative mode of inheritance. For this, the data from Table 11.1 can be utilized and tested with the Cochran–Armitage trend test [34, 135]. The test for trend measures a linear trend in proportions weighted by a general measure of exposure dosage $x_i, i = 1, \dots, n$. When measuring the effects of alleles, we let the dosages be identical to the number of A alleles a person has, thus, $x_i = 0, 1$, or 2 . The resulting trend test statistic with additive scores

$$\chi_{trend}^2 = \frac{n}{rs} \cdot \frac{(2r_2s - 2rs_2 + r_1s - s_1r)^2}{2n_2n + (2n_2 + n_1)(n_0 - n_2)} \quad (11.5)$$

is asymptotically χ^2 distributed with 1 d.f. As before, the null hypothesis of no association is rejected for large values of the χ^2 test statistic. This approach is also analogous to a linear regression model where the genotype is coded $0, 1, 2$ and used as independent variable, whereas the case-control status is the dependent variable (see Problem 11.3). If only small samples are investigated, the approximations of the test statistic to the χ^2 statistic may not be valid. Hence, it is preferable to use exact tests or Monte–Carlo tests instead.

We can also formulate the trend test statistic using general scores x_i for groups $i = 0, 1, 2$ with

$$\chi_{gtrend}^2 = \frac{n^2}{rs} \cdot \frac{((\sum_{i=0}^2 x_i r_i) - r\bar{x})^2}{\sum_{i=0}^2 n_i (x_i - \bar{x})^2}, \quad (11.6)$$

where $\bar{x} = \frac{1}{n} \sum_{i=0}^2 n_i x_i$. Note that Eq. (11.6) reduces to Eq. (11.5) if scores $0, 1, 2$ are used. For scores $0, 0, 2$ and $0, 2, 2$ we obtain the test statistics for the recessive and the dominant modes of inheritance, respectively.

When the underlying genetic model, i.e., the choice of the scores is unclear, a substantial loss of power may incur when a test suitable for a specific mode of inheritance is chosen but a different mode is the true one. To this end, Freidlin and colleagues [226] proposed the MAX test as an alternative, which is the maximum of the standardized optimal tests with dominant, recessive and addi-

tive scores. Their procedure is Monte–Carlo based and therefore time consuming. The general idea is as follows (see Algorithm 11.1): To make the test statistics $\chi_{dom}^2, \chi_{rec}^2$, and χ_{trend}^2 comparable with respect to location and variability, we first need to estimate the mean and the standard deviation of each statistic under the null hypothesis from permutations. Second, we standardize the test statistics using the estimated mean and standard deviation. The MAX test statistic then selects the max-

Table 11.5 Observed genotype frequencies in patients and controls at IL1B-511 (rs16944).

	CC	CT	TT	Total
Patients	126	87	18	231
Controls	151	162	32	345
Total	277	249	50	576

imum of the standardized statistics for the additive, dominant, and recessive scores, and the p -values are obtained from the permutations.

Algorithm 11.1.

1. Compute χ_{dom}^2 , χ_{rec}^2 , and χ_{trend}^2 using the original n data.
2. For a specified number N of permutations,
 - (a) Permute the case-control status and arrange the data into n new subjects.
 - (b) Compute $\chi_{dom,(j)}^2$, $\chi_{rec,(j)}^2$ and $\chi_{trend,(j)}^2$ using the j th permuted data and keep all permuted test statistics.
3. Compute the means m and standard deviations s of the permuted test statistics. For example, mean and standard deviation of the permuted test statistic under a dominant mode of inheritance are

$$m_{dom} = \frac{1}{N} \sum_{j=1}^N \chi_{dom,(j)}^2 \text{ and } s_{dom} = \frac{1}{N-1} \sum_{j=1}^N (\chi_{dom,(j)}^2 - m_{dom})^2.$$
4. Compute the standardized test statistics S_{dom} , S_{rec} , and S_{trend} of the original test statistic. For example, the standardized dominant test statistic is obtained as $S_{dom} = \frac{\chi_{dom}^2 - m_{dom}}{s_{dom}}$.
5. Compute the MAX test as $MAX = \max\{S_{dom}, S_{rec}, S_{trend}\}$.
6. Set counter $C = 0$.
7. For the already existing N permutations,
 - (a) Compute the permuted MAX test statistic

$$MAX_{(j)} = \max\{S_{dom,(j)}, S_{rec,(j)}, S_{trend,(j)}\}$$
 for the j th permutation,
 where, e.g., $S_{dom,(j)} = \frac{\chi_{dom,(j)}^2 - m_{dom}}{s_{dom}}$.
 - (b) If the test statistic $MAX_{(j)}$ from the permuted data j is at least as large as MAX using the original data, add 1 to C .
8. The Monte--Carlo p -value is C/N .

It is quite clear that this permutation procedure is quite time consuming, and an asymptotic test would be preferable, and this will be discussed in Section 11.2.4, where a combined approach for testing and selecting the most appropriate genetic model is discussed.

Example 11.1. *In this example, we consider the data of Reich and colleagues [557]. Here, 231 psoriasis patients and 345 healthy controls were investigated with regard to their genotype distribution in the candidate genes encoding for tumor necrosis factor- α (TNFA) and interleukin-1 β (IL1B). To analyze a possible association between psoriasis and the SNP rs16944 (NG_008851.1:g.4490T>C), also known as IL1B-511T>C, we can choose one of the tests described above. Of course, the power of the chosen test will strongly rely on the underlying genetic model. However, for didactical purposes we will calculate all test statistics introduced previously, irrespective of a suspected mode of inheritance. The data used for calculations are displayed in Table 11.5.*

The first test used is the classical χ_c^2 test statistic χ_c^2 , and for this we obtain

$$\begin{aligned}
 \chi_c^2 &= \frac{(2 \cdot 576 \cdot (2 \cdot 126 + 87) - (2 \cdot 277 + 249) \cdot 2 \cdot 231)^2}{2 \cdot 576 \cdot (2 \cdot 277 + 249) \cdot 2 \cdot 231} + \dots \\
 &\quad + \frac{(2 \cdot 576 \cdot (2 \cdot 32 + 162) - (2 \cdot 50 + 249) \cdot 2 \cdot 345)^2}{2 \cdot 576 \cdot (2 \cdot 50 + 249) \cdot 2 \cdot 345} \\
 &= 4.9245.
 \end{aligned}$$

This corresponds to a p -value of 0.0265. As the χ_A^2 test statistic is equivalent to χ_c^2 , we receive the same result calculating

$$\begin{aligned}\chi_A^2 &= 2 \cdot 576 \cdot \frac{[(2 \cdot 126 + 87)(2 \cdot 32 + 162) - (2 \cdot 18 + 87)(2 \cdot 151 + 162)]^2}{2 \cdot 231 \cdot 2 \cdot 345 \cdot (2 \cdot 277 + 249) \cdot (2 \cdot 50 + 249)} \\ &= 4.9245.\end{aligned}$$

The test statistic χ_g^2 that utilizes the genotype and not the allele frequencies is given by

$$\chi_g^2 = \frac{(126 - 277 \cdot \frac{231}{576})^2}{277 \cdot \frac{231}{576}} + \dots + \frac{(32 - 50 \cdot \frac{345}{576})^2}{50 \cdot \frac{345}{576}} = 6.4571.$$

Because this test statistic is asymptotically χ^2 distributed with 2 d.f. under the null hypothesis of no association, the test has less power than the allele-based test χ_A^2 . Subsequently, the corresponding p -value is somewhat higher ($p = 0.0396$).

We next calculate the test statistics χ_{dom}^2 , χ_{rec}^2 and χ_{het}^2 . Given the notation of Table 11.4, we get

$$\begin{aligned}\chi_{dom}^2 &= 576 \cdot \frac{(126 \cdot (162 + 32) - (87 + 18) \cdot 151)^2}{231 \cdot 345 \cdot 277 \cdot (249 + 50)} = 6.4376, \\ \chi_{rec}^2 &= 576 \cdot \frac{((126 + 87) \cdot 32 - 18 \cdot (151 + 162))^2}{231 \cdot 345 \cdot (277 + 249) \cdot 50} = 0.3839, \\ \chi_{het}^2 &= 576 \cdot \frac{(87 \cdot (151 + 32) - (126 + 18) \cdot 162)^2}{231 \cdot 345 \cdot 249 \cdot (277 + 50)} = 4.8700.\end{aligned}$$

Here, the resulting test statistics differ substantially. While the test for a model of dominant inheritance shows a significant result ($p = 0.0112$), the test for the recessive model does not ($p = 0.5355$). The χ_{het}^2 statistic is significant as well, but with a slightly higher p -value ($p = 0.0273$). This is a hint that at least the recessive inheritance model may be invalid.

Finally, we will consider the Cochran–Armitage trend test. We have

$$\begin{aligned}\chi_{trend}^2 &= \frac{576}{231 \cdot 345} \cdot \frac{(2 \cdot 18 \cdot 345 - 2 \cdot 231 \cdot 32 + 87 \cdot 345 - 162 \cdot 231)^2}{2 \cdot 50 \cdot 576 + (2 \cdot 50 + 249)(277 - 50)} \\ &= 5.0433\end{aligned}$$

with a corresponding p -value of 0.0247.

We refrain from exercising the MAX test in its form proposed by Freidlin and colleagues [226], and we refer to the example in Section 11.2.4.

In sum, all tests except for the recessive inheritance model χ_{rec}^2 indicate a significant association of IL1B-511T>C with psoriasis.

11.2.2 Which association test should be used in applications?

After describing a number of different variants for testing genetic association, the obvious question will be which variant to use in a given setting. If an a priori hy-

pothesis of the genetic model is available, the appropriate model-specific association tests should be used. Zheng and colleagues [749] have shown that the tests for the dominant and the recessive models are optimal for true dominant and recessive modes of inheritance in the sense that they minimize the required sample size for the trend test at type error level α to achieve a pre-specified power. Furthermore, the scores 0, 1, 2 are locally optimal for the additive model, and the multiplicative model is asymptotically equivalent to the additive model.

Lacking a specific biological hypothesis, Sasieni [577] has advised assuming a linear trend and hence the application of the trend test. This test is based on genotype frequencies and not on allele frequencies. Test statistics based on allele frequencies depend on the assumption of Hardy–Weinberg equilibrium (HWE), and they have been shown not to be robust against deviations from it (compare Section 4.3.4, Eqs. (4.4) and (4.5)). In contrast, the Cochran–Armitage trend test inherently takes deviations from the HWE into account using $\chi_{trend}^2 = \chi_c^2 / (1 + \hat{\rho})$, where ρ is the intra-subject correlation or inbreeding coefficient (Chapter 4) and, with the notation of Table 11.1, is estimated by $\frac{4n_0n_2 - n_1^2}{(2n_0 + n_1)(2n_2 + n_1)}$ [265, 577, 771]. In addition, it seems biologically more plausible to utilize genotypes, as this is how the genetic information in humans is naturally available. We therefore absolutely agree with this recommendation and use the Cochran–Armitage trend test for primary analysis in most instances. Alternatively, one can use the asymptotic version of the MAX test approach (Section 11.2.4 and Ref. [315]) or the computationally intensive original permutation-based MAX test approach of Freidlin and colleagues [225, 226] that has been shown to be generally powerful [752].

One of the shortcomings of the approaches discussed is that both cases and controls are required to test for association. Therefore, Lee [403] has proposed searching for associations utilizing a test for detecting deviation from HWE in affected individuals only (see Chapter 4). This proposal has been strongly criticized (see, e.g., Ref. [703]), with arguments similar to those described in Section 8.9.4 for the situation of affected sib-pairs.

So far, we have considered only the case of testing association with a diallelic marker. If instead a multiallelic marker is genotyped, a number of variants exist:

1. According to the number of observed genotypes k , a $2 \times k$ table is set up. A general χ^2 test of independence with $(k - 1)$ d.f. can be performed to test for association. A pitfall of this approach, however, is that with a higher number of alleles and observed genotypes, the number of observations per cell will become smaller; hence, the approximation may not be valid.
2. The alleles can be grouped into presumed risk and no-risk alleles. Based on this, a 2×3 table can again be constructed with the frequencies for cases and controls of carrying zero, one, or two risk alleles. Then, the same strategies as for genotypes of a diallelic marker can be employed.
3. Instead of summarizing alleles based on function, in short tandem repeat (STR) markers a grouping might be based on allele lengths and frequencies. For this, alleles with a higher number of repeat units and those with fewer repeat units can be grouped. Put simply, from a larger number of alleles, a binary variable

indicating *short allele* or *long allele* is constructed. Again, a 2×3 table can be constructed giving the frequencies of carrying zero, one, or two long alleles.

4. In some situations, it might be possible to rank the observed genotypes according to their associated presumed risk. In this case, the trend test by Armitage [34] as described above, or extensions of it, may be utilized.

11.2.3 How can the genetic effect be measured?

In this section, we consider different effect measures used in association analyses. The different variants of statistical tests so far described can be applied regardless of how the unrelated individuals have been recruited—in a case-control or a cohort approach. However, to estimate the size of the genetic effect, the case-control study design has to be differentiated from the cohort study. Specifically, if probands were recruited prospectively and independent of the disease status as in a cohort study, risk estimates and relative risks can be utilized. In case-control studies, however, where the affection status is known before a subject is recruited, risks generally cannot be estimated. However, a comparison of the odds of disease is possible between two groups, and this leads to the odds ratio.

The 2×2 frequency tables of the different genetic models were depicted in Table 11.4. For a general view, we use the notation given in Table 11.6. Specifically, we speak of a variant that might be present $G = 1$ or absent $G = 0$. In the general epidemiological context, the variant is replaced by an exposure E ; exposed subjects are denoted by $E = 1$ or just E and unexposed subjects by $E = 0$. For the application to specific 2×2 tables see Problem 11.4.

The *incidence* is a measure of the probability of a subject for developing disease, or, generally speaking, a specific novel phenotype, within a specified period of time when belonging to a pre-specified population. We often use the term *risk* instead of incidence and denote it by $p = P(\text{aff})$ for the general population. Similarly, $p_1 = P(\text{aff}|G = 1)$ and $p_0 = P(\text{aff}|G = 0)$ denote the probability of becoming affected when carrying and not carrying the variant, respectively. The two risks can be estimated by

$$\hat{p}_1 = \frac{x_1}{x_1 + y_1} = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_0 = \frac{x_0}{x_0 + y_0} = \frac{x_0}{n_0}.$$

An approximate $100(1 - \alpha)\%$ confidence interval for the risk p_1 with fairly good coverage probabilities is given by

$$CI(p_1) = \frac{\hat{p}_1 + c^2/2}{1 + c^2} \pm c \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1) + c^2/4}{1 + c^2}},$$

Table 11.6 2×2 frequency table in cases and controls. It corresponds to any of the three tables provided in Table 11.4.

	Variant	
	$G = 1$	$G = 0$
Cases	x_1	x_0
Controls	y_1	y_0
Total	n_1	n_0

where $c = z_{1-\frac{\alpha}{2}}/\sqrt{n_1}$, and $z_{1-\frac{\alpha}{2}}$ is the upper $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. The confidence interval for p_0 is analogously given. For a review of many different confidence intervals for a single proportion, see Ref. [497].

The main aim of a study often is to compare the risks for becoming affected in variant carriers and non-carriers. To contrast two risks, one can either consider the ratio or the difference of the two risks. In the first case, we obtain the *relative risk* (RR) or *risk ratio*

$$RR_G = \frac{P(\text{aff}|G=1)}{P(\text{aff}|G=0)} = \frac{p_1}{p_0},$$

which can be estimated by

$$\widehat{RR}_G = \frac{\hat{p}_1}{\hat{p}_0} = \frac{x_1/n_1}{x_0/n_0} = \frac{x_1 n_0}{x_0 n_1}.$$

This estimator cannot be calculated when either $x_1 = 0$ or $x_0 = 0$. Therefore, we may use the slightly biased estimator of $\log p_1/p_0$ [685]:

$$\log \widehat{RR}_{G,0.5} = \log \frac{\hat{p}_1}{\hat{p}_0} = \log \frac{x_1 + 0.5}{n_1 + 0.5} - \log \frac{x_0 + 0.5}{n_0 + 0.5}.$$

Many different approaches for constructing asymptotic confidence intervals for RR have been proposed, and an overview is given in Ref. [542]. As a consequence of the delta method, the log of the relative risk has a sampling distribution that is approximately normal with variance that can be estimated by a formula involving the number of subjects in each group and the event rates in each group. This permits the construction of an approximate confidence interval that is symmetric around $\log RR$. The variance estimator of $\log \widehat{RR}_{G,0.5}$ is [531]

$$\text{Var}(\log \widehat{RR}_{G,0.5}) = \frac{1}{x_1 + 0.5} + \frac{1}{x_0 + 0.5} - \frac{1}{n_1 + 0.5} - \frac{1}{n_0 + 0.5},$$

and an approximate $100(1 - \alpha)\%$ confidence interval for RR is then given by

$$\begin{aligned} CI(RR_{G,0.5}) &= \widehat{RR}_{G,0.5} \cdot \exp \left(\pm z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\log \widehat{RR}_{G,0.5})} \right) \\ &= \widehat{RR}_{G,0.5} \cdot \exp \left(\pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{x_1+0.5} + \frac{1}{x_0+0.5} - \frac{1}{n_1+0.5} - \frac{1}{n_0+0.5}} \right). \end{aligned}$$

Using a Bayesian approach, Price and Bonett [542] derived a general formula for approximate $100(1 - \alpha)\%$ credible intervals for the RR . In an extensive simulation study, they investigated the coverage probabilities of their approach and recommended the following approximate credible interval:

$$CI(RR_{PB}) = \widetilde{RR}_{G,PB} \cdot \exp \left(\pm z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\log \widetilde{RR}_{G,PB})} \right),$$

where

$$\log \widetilde{RR}_{G,PB} = \log \frac{x_1 + 0.25}{n_1 + 1.75} - \log \frac{x_0 + 0.25}{n_0 + 1.75}$$

and

$$\mathbb{V}\text{ar}(\log \widehat{RR}_{G,PB}) = \frac{1}{(x_1 + 0.25) + \frac{(x_1 + 0.25)^2}{y_1 + 1.5}} + \frac{1}{(x_0 + 0.25) + \frac{(x_0 + 0.25)^2}{y_0 + 1.5}}.$$

Instead of looking at the ratio of the two risks, one may consider their difference, and this leads to the *risk difference (RD)*. In the context of clinical trials, the RD is also termed *absolute risk reduction (ARR)*. It is calculated by subtracting the risk in the carriers p_1 from the risk in the non-variant carriers p_0 :

$$\text{RD} = p_0 - p_1$$

The negative value of RD, i.e., $-\text{RD}$, is termed *excess incidence*.

The standard approximate confidence interval for the RD is based on the assumption both proportions are independent and asymptotically normally distributed, yielding

$$CI(\text{RD}) = \widehat{\text{RD}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}. \quad (11.7)$$

Newcombe [496] compared this and 10 other confidence intervals for the RD showing that Eq. (11.7) cannot be recommended as approximate confidence interval for RD. Instead, one should use the simple to calculate approximate confidence interval of Newcombe given by

$$CI(\text{RD}_N) = [\widehat{\text{RD}} - \delta; \widehat{\text{RD}} + \varepsilon],$$

where

$$\begin{aligned} \delta &= \sqrt{(\hat{p}_1 - l_*)^2 + (u_* - \hat{p}_0)^2} = z_{1-\frac{\alpha}{2}} \sqrt{\frac{l_*(1-l_*)}{n_1} + \frac{u_*(1-u_*)}{n_0}}, \\ \varepsilon &= \sqrt{(u_* - \hat{p}_1)^2 + (\hat{p}_0 - l^*)^2} = z_{1-\frac{\alpha}{2}} \sqrt{\frac{u_*(1-u_*)}{n_1} + \frac{l^*(1-l^*)}{n_0}}, \end{aligned} \quad (11.8)$$

with

$$\begin{aligned} l_* &= \frac{2x_0 + z^2 - \xi}{2(n_0 + z^2)}, & l^* &= \frac{2x_0 + z^2 + \xi}{2(n_0 + z^2)}, \\ u_* &= \frac{2x_1 + z^2 - \xi}{2(n_1 + z^2)}, & u^* &= \frac{2x_1 + z^2 + \xi}{2(n_1 + z^2)}, \end{aligned}$$

for $\xi = z\sqrt{z^2 + 4x_0(1-\hat{p}_0)}$ and $z = z_{1-\frac{\alpha}{2}}$. This confidence interval is available as EXCEL tool EECI, and its calculation is illustrated in Example 11.2.

An important measure is the *attributable risk (AR)* that is the proportion of cases in the population attributable to the variant. It is the proportion of cases in the population that could be prevented if the variant had normal function. The AR is given by

$$\text{AR} = \frac{P(\text{aff}) - P(\text{aff}|G=0)}{P(\text{aff})} = \frac{p - p_0}{p}.$$

Equivalent definitions use the RR . Specifically, with $P(G = 1)$ denoting the probability for being a variant carrier, the AR may also be defined as [52]

$$AR = \frac{P(G = 1)(RR - 1)}{1 + P(G = 1)(RR - 1)} = P(G = 1|aff) \frac{RR - 1}{RR}. \quad (11.9)$$

The first alternative definition given in Eq. (11.9) shows that AR depends not only on the strength of the association through RR but also on the probability of being a variant carrier. Some confusion in the terminology exists, and the most commonly used terms for the attributable risk are *population attributable risk (PAR)*, *etiologic fraction*, *population attributable fraction (PAF)*, or just *attributable fraction*. Even worse, some authors call the risk difference “attributable risk,” and the difference between these terms can only be detected from the context.

In case-control studies, the AR is estimated by $\widehat{AR} = \frac{x_1 y_0 - x_0 y_1}{(x_0 + x_1) y_0}$ [432]. Many different variance estimators for the AR have been derived, depending on the study design, and, for case-control studies, the standard error S.E. of AR is estimated by

$$S.E.(\widehat{AR}) = (1 - \widehat{AR}) \sqrt{\frac{x_1}{x_0(x_0 + x_1)} + \frac{y_1}{y_0(y_0 + y_1)}}.$$

Three approximate confidence intervals are commonly used for the AR. The standard confidence interval is based on the untransformed AR point estimate and given by

$$CI(AR) = \widehat{AR} \pm z_{1-\frac{\alpha}{2}} S.E.(\widehat{AR}).$$

The second confidence interval uses the delta method and is based on the log-transformed variable $\log(1 - AR)$. It is given by

$$CI(AR_{\log}) = 1 - (1 - \widehat{AR}) \exp\left(\pm z_{1-\frac{\alpha}{2}} \frac{S.E.(\widehat{AR})}{(1 - \widehat{AR})}\right)$$

and is wider than the standard confidence interval. The third confidence interval also uses the delta method and is based on the logit-transformed variable $\log \frac{AR}{1-AR}$. It is computed from [52, 405]

$$CI(AR_{\logit}) = \left(1 + \frac{1 - \widehat{AR}}{\widehat{AR}} \exp\left(\pm \frac{z_{1-\frac{\alpha}{2}} S.E.(\widehat{AR})}{\widehat{AR}(1 - \widehat{AR})}\right)\right)^{-1}.$$

This confidence interval is not applicable if $\widehat{AR} < 0$ because the logit transformation needs to be positive. In this case, the second confidence interval is preferable.

A comparison between the three confidence intervals is difficult, and no recommendation can be given for all situations and all models. As a theoretical result, the logit-transformed interval is narrower than the standard interval when $|\widehat{AR} - \frac{1}{2}| \leq \sqrt{3}/6$, i.e., for values of AR strictly between 0.21 and 0.79 [405], and wider otherwise. Leung and Kupper [405] compared the standard untransformed

confidence interval with their logit-transformed confidence interval. They recommend the use of the logit-transformed confidence interval when the estimate of AR is in the interval from 0.21 to 0.79 and the standard untransformed confidence interval for AR estimates outside of this interval. Further, $CI(AR_{\text{logit}})$ is recommended as confidence interval for AR if RR is large, i.e., ≥ 4 ; it is inappropriate when $RR \approx 1$ [431]. When the frequency of the variant is low, the log-transformed confidence interval is generally recommended. It also performs well if the frequency of the variant is moderate or large ≥ 0.2 [431]. However, it may still be inefficient in some situations.

Another measure for the impact of a gene is the *attributable risk in the variant carriers* (AR_G) that is termed *attributable risk in the exposed* in the general epidemiological context. It is defined as [52]

$$AR_G = \frac{P(\text{aff}|G=1) - P(\text{aff}|G=0)}{P(\text{aff}|G=1)} = \frac{p_1 - p_0}{p_1}.$$

An equivalent representation is

$$AR_G = \frac{RR - 1}{RR}.$$

AR_G is also termed *attributable fraction in the exposed* or *excess fraction*. Similar to AR, AR_G has an interpretation as a measure of the impact of the gene and could be used to see the relative difference if a variant would not be present in the population. An approximate confidence interval for AR_G is simple to calculate given the approximate for RR . In this case, the transformation rule for the confidence interval is used. Specifically, given the lower and the upper bounds of the confidence interval for the RR , say RR_L and RR_U , the lower and upper bounds of the confidence interval for AR_G are given by

$$CI(AR_G) = \left[\frac{RR_L - 1}{RR_L}; \frac{RR_U - 1}{RR_U} \right].$$

Independent from the assumption that risk estimates are available, and hence an alternative to the more commonly used case-control designs, is the definition of the genetic effect in terms of the OR as defined in Problem 10.2.2. Specifically, the odds₁ for affection given a variant and the odds₀ for affection if one carries no variant are estimated by

$$\widehat{\text{odds}}_1 = \frac{x_1}{y_1} \quad \text{and} \quad \widehat{\text{odds}}_0 = \frac{x_0}{y_0}.$$

Subsequently, an estimator for the odds ratio OR_G is

$$\widehat{OR}_G = \frac{\widehat{\text{odds}}_1}{\widehat{\text{odds}}_0} = \frac{x_1 y_0}{y_1 x_0}.$$

To construct a confidence interval for the OR , different methods are available. For instance, an asymptotically valid $(1 - \alpha)$ confidence interval by Woolf [728] is

$$CI(OR_G) = \widehat{OR}_G \cdot \exp\left(\pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{x_0} + \frac{1}{x_1} + \frac{1}{y_0} + \frac{1}{y_1}}\right).$$

The coverage of this confidence interval is still good in case of small samples and it is therefore generally recommended [6].

If the disease is rare in the general population, $y_1 \approx n_1$ and $y_0 \approx n_0$ so that the OR can be viewed as an approximation of the RR for diseases with a low prevalence. In these situations, some measures of impact can be estimated using the OR instead of the RR . For example, the odds ratio-based AR_G is given by $(OR - 1)/OR$.

We close this section by illustrating the calculation of risks, RR , RD , AR and related measures. To this end, we need to switch to an example where data come from a cohort study.

Example 11.2. *In this example, we consider the study by Menges et al. [462]. Patients encountering severe trauma are at risk of developing sepsis syndrome. In this study, 159 severely traumatized patients from a single center were included, and the replication sample consisted in $n = 76$ patients (Table 11.7). It was investigated whether carriage of the A allele of the $TNF-308A>G$ SNP (rs1800629) A allele was associated with the development of sepsis syndrome. In this example, we consider the initial sample. Calculations involving the replication sample are left as Problem 11.8. We do not provide detailed intermediate steps in the calculation. The Excel tool EECI is available from our web site.*

Table 11.7 Distribution of $TNF-308A>G$ (rs1800629) variants among patients with and without sepsis syndrome in the initial and replication samples.

Sepsis	Initial sample		Replication sample	
	AA or GA	GG	AA or GA	GG
Yes	33	38	9	8
No	9	74	15	44
Total	42	112	24	52

In the initial sample, the proportion of patients developing sepsis is $\hat{p} = 71/154 = 0.4610$. To calculate an approximate 95% confidence interval for p we first calculate c to be 0.1579. The first term for this confidence interval gives 0.4620 and the second term yields 0.078 so that the confidence interval is given by $[0.3833; 0.5407]$. The risk for a sepsis is not low, and therefore we cannot expect that RR and OR are similar in this study.

In variant carriers, the proportion of patients developing a sepsis syndrome is $\hat{p}_1 = 0.7858 = 33/42$, and it is $\hat{p}_0 = 0.3393 = 38/112$ in non carriers. The RD is estimated as $\widehat{RD} = \hat{p}_0 - \hat{p}_1 = -0.4464$. For calculating the 95% CI of Eq. (11.8), we use a few intermediate steps. First, we need to calculate l_* to u^* . The first terms in the numerator are $2x_0 + z^2 = 79.8415$ and $2x_1 + z^2 = 69.8415$. For the denominator, we obtain $2(n_0 + z^2) = 231.6829$ and $2(n_1 + z^2) = 91.6829$. The second terms in the numerator are given by $z\sqrt{z^2 + 4x_0(1 - \hat{p}_0)} = 20.0137$ and $z\sqrt{z^2 + 4x_1(1 - \hat{p}_1)} = 11.1092$ so that $l_* = 0.2582$, $u_* = 0.6406$, $l^* = 0.4310$, and $u^* = 0.8829$. The 95% CI finally is obtained as $[-0.5730, -0.2748]$.

Using the Price and Bonett formula, we first estimate RR as $\widehat{RR}_{G,PB} = 2.2601$, and with the standard approach we obtain as estimates $\widehat{RR}_G = 2.3158$ and $\widehat{RR}_{G,0.5} = 2.3033$. The 95% credible interval according to Price and Bonett is computed to be $CI(RR_{PB}) = [1.6623; 3.0730]$, and the standard 95% confidence interval using $RR_{G,0.5}$ is $CI(RR_{G,0.5}) = [1.7066; 3.1087]$.

Using the estimator of the RR and the credible interval according to Price and Bonett, we obtain $\widehat{AR}_{G,PB} = 0.5474$ as estimator of AR and $CI(AR_{G,PB}) = [0.3803; 0.6695]$ as the 95% credible interval for AR_G . For the standard estimator of AR_G , we similarly get $\widehat{AR}_G = 0.5682$ as an estimator of AR and $CI(AR_G) = [0.4172; 0.6801]$ as 95% confidence interval for AR_G .

The study is a cohort study, and we should not use the estimator of the AR and the confidence intervals given above.

Finally, for the odds ratio we yield $\widehat{OR} = 7.1404$, and the 95% confidence interval is obtained as $CI(OR) = [3.0998; 16.4475]$.

11.2.4 How can the genetic model be selected?

We have already seen in Section 11.2.1 that most of the test statistics used to test for association assume a specific genetic model, and typically the three tests dominant, recessive, and additive are used. The misspecification of the genetic model can lead to a considerable loss in power. An easy but unsatisfactory solution is to perform multiple tests for every genetic model and subsequently adjust for the multiple testing. Alternatively, we can utilize an approach that simultaneously tests for association and selects the most likely model. One of them, the MAX test proposed by Freidlin et al. [226], was already described in Section 11.2.1. As pointed out, because the asymptotic distribution of the test statistic is unknown, p -values are obtained from simulation, which makes this procedure computationally too intensive to be feasible in many settings. Similar approaches were investigated for the special case of matched case-control data [755].

A solution to this drawback has recently been suggested by Hothorn and Hothorn [315]. They used the general framework of conditional independence tests [640] in which the MAX test is a special case. One important advantage is that the distribution can thus be easily approximated using a three-dimensional normal distribution, and the most likely underlying genetic model is simultaneously estimated.

To formulate the test statistic, we use the grouping of Table 11.1 as starting point. For the six different genotype groups i , Table 11.8 provides the case-control status y_i with values 1 for cases and 0 for controls. $x_i^{(model)}$ denotes the genotype scores for the corresponding cells given the specific genetic model. The genotype frequencies are given in column 4 of the table, and they will be used as weights w_i . Testing for association can be formulated as testing the null hypothesis of independence of the case-control status y from the genotype score x for a distribution $\mathcal{L}$, thus $H_0: \mathcal{L}(y|x) = \mathcal{L}(y)$.

Based on the scores, a three-dimensional test statistic T is derived with each dimension representing one of the model-specific test statistics. To this end, we

Table 11.8 Reformulation of genotype distribution for the conditional test.

i	Disease status	Genotype	w_i	y_i	$x_i^{(dom)}$	$x_i^{(add)}$	$x_i^{(rec)}$
1	Case	aa	r_0	1	0	0	0
2	Case	aA	r_1	1	1	1	0
3	Case	AA	r_2	1	1	2	1
4	Control	aa	s_0	0	0	0	0
5	Control	aA	s_1	0	1	1	0
6	Control	AA	s_2	0	1	2	1

i denotes the genotype group, w_i the weight for the conditional test, y_i the dichotomous case-control status, and $x_i^{(mod)}$ denotes the scores for the different genetic models with dom for dominant, add for additive, and rec for recessive.

collect the scores $\mathbf{x}_i = (x_i^{(dom)}, x_i^{(add)}, x_i^{(rec)})'$ to a 3×1 column vector $\mathbf{x}_i$, and we let

$$\mathbf{T} = \begin{pmatrix} T_{dom} \\ T_{add} \\ T_{rec} \end{pmatrix} = \sum_{i=1}^6 w_i y_i \mathbf{x}_i.$$

The linear statistic $\mathbf{T}$ is the vector of the unstandardized trend statistics. We then use the so-called influences y that has the conditional expectation $\mathbb{E}(y) = \frac{1}{n} \sum_i w_i y_i$ and variance $\mathbb{V}\text{ar}(y) = \frac{1}{n} \sum_i w_i (y_i - \mathbb{E}(y))^2$. These are utilized for obtaining the conditional expectation and variance of $\mathbf{T}$ under the null hypothesis, which are given by

$$\begin{aligned} \mathbb{E}(\mathbf{T}) &= \mathbb{E}(y) \sum_{i=1}^6 w_i \mathbf{x}_i, \\ \mathbb{V}\text{ar}(\mathbf{T}) &= \frac{n}{n-1} \mathbb{V}\text{ar}(y) \cdot \left(\sum_{i=1}^6 w_i \mathbf{x}_i \mathbf{x}_i' \right) - \frac{1}{n-1} \mathbb{V}\text{ar}(y) \cdot \left(\sum_{i=1}^6 w_i \mathbf{x}_i \right) \left(\sum_{i=1}^6 w_i \mathbf{x}_i \right)'. \end{aligned}$$

$\mathbf{T}$ is asymptotically distributed as a three-dimensional normal distribution. Under H_0 , it has expectation $\mathbb{E}(\mathbf{T})$ and covariance matrix $\mathbb{V}\text{ar}(\mathbf{T})$.

The MAX test can now be defined as

$$T_{max} = \max \left| \frac{\mathbf{T} - \mathbb{E}(\mathbf{T})}{\sqrt{\text{diag}(\mathbb{V}\text{ar}(\mathbf{T}))}} \right|,$$

and its distribution can be evaluated using three-dimensional normal probabilities; for a description of calculating the maximum of independent random variables, see, e.g., Ref. [195]. Here, independence can be established through standardization. In itself, the MAX test does not inform on the underlying genetic model. For this, multiplicity adjusted p -values for each test are used that indicate which genetic model is most likely. This approach can also be used for deriving asymptotically valid confidence intervals [314]. A number of extensions are possible for this approach such as

comparing more than two groups or using a stratified design as in a meta-analysis [597]. A worked-out example is given below.

Other procedures to evaluate the genetic model within an association study have been proposed that are not detailed here. For instance, a different approximation of the distribution of the MAX test statistic was derived by Gonzalez and colleagues [255]. In an alternative approach, Zheng and Ng [754] suggested a two-stage procedure with the purpose of the first stage being the estimation of the genetic model, which is followed by the association test in the second.

In addition to procedures that select a genetic model from the data of a single study, some approaches have been suggested for the application in meta-analyses. To simultaneously estimate the genetic effect and the genetic model, the two log OR s corresponding to the heterozygous and homozygous effects can be modeled together in a bivariate response. Let OR_{1k} denote the odds of being affected carrying the aA genotype over the odds of being affected with the aa genotype and OR_{2k} be the odds of being affected carrying the AA genotype over the odds of being affected with the aa genotype, with $k = 1, \dots, K$ indicating the respective study. In the following, the corresponding log OR s are termed θ_{1k} and θ_{2k} , the approximate variances $\text{Var}(\hat{\theta}_{\cdot k})$ of $\hat{\theta}_{\cdot k}$ are denoted by $\sigma_{\cdot k}^2$, and the covariances by $\text{Cov}(\hat{\theta}_{1k}, \hat{\theta}_{2k}) = \sigma_{12,k}$.

For considering the joint distribution of $\hat{\theta}_{1k}$ and $\hat{\theta}_{2k}$, two different approaches are possible within a meta-analytical framework. In the so-called *fixed effects model* (*FE model*), we assume that the true effects θ_1 and θ_2 are identical for every study, so that $\hat{\theta}_{1k}$ and $\hat{\theta}_{2k}$ are jointly normally distributed. In contrast, the *random effect model* (*RE model*) allows heterogeneity between studies with regard to the true effect, and we therefore include a random component of the between studies variance τ^2 . In the unrealistic case of no between-study heterogeneity, both models produce the same results because of $\tau^2 = 0$.

Utilizing an RE model, we consider the joint distribution of $\hat{\theta}_{1k}$ and $\hat{\theta}_{2k}$ as

$$\begin{pmatrix} \hat{\theta}_{1k} \\ \hat{\theta}_{2k} \end{pmatrix} \stackrel{a}{\sim} N \left(\begin{bmatrix} \theta_{1k} \\ \theta_{2k} \end{bmatrix}, \begin{bmatrix} \sigma_{1k}^2 & \sigma_{12,k} \\ \sigma_{12,k} & \sigma_{2k}^2 \end{bmatrix} \right),$$

with the mean vector $(\theta_{1k}, \theta_{2k})'$ distributed as

$$\begin{pmatrix} \theta_{1k} \\ \theta_{2k} \end{pmatrix} \stackrel{a}{\sim} N \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \begin{bmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{bmatrix} \right).$$

These expressions can be combined to

$$\begin{pmatrix} \hat{\theta}_{1k} \\ \hat{\theta}_{2k} \end{pmatrix} \stackrel{a}{\sim} N \left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \begin{bmatrix} \sigma_{1k}^2 + \tau_1^2 & \sigma_{12,k} + \tau_{12} \\ \sigma_{12,k} + \tau_{12} & \sigma_{2k}^2 + \tau_2^2 \end{bmatrix} \right). \quad (11.10)$$

The empirical variances of Eq. (11.10) can be estimated by Woolf-type [728] variance estimators (cf. Section 11.2.3)

$$\hat{\sigma}_{1k}^2 + \hat{\tau}_1^2 = \frac{1}{r_{0k}} + \frac{1}{s_{0k}} + \frac{1}{r_{1k}} + \frac{1}{s_{1k}} \quad \text{and} \quad \hat{\sigma}_{2k}^2 + \hat{\tau}_2^2 = \frac{1}{r_{0k}} + \frac{1}{s_{0k}} + \frac{1}{r_{2k}} + \frac{1}{s_{2k}},$$

where $r_{.k}$ and $s_{.k}$ are the genotype frequencies of Table 11.1 extended to K studies. Similarly, the empirical covariances can be estimated by (see Ref. [41])

$$\hat{\sigma}_{12,k} + \hat{\tau}_{12} = \frac{1}{s_{0k}} + \frac{1}{r_{0k}}.$$

The underlying parameters can be estimated from the study-specific log OR s and the estimated covariance matrices using a maximum likelihood procedure as described in detail by Bagos [41], Appendix III, using STATA code.

Based on the estimated model parameters $\hat{\theta}_1$ and $\hat{\theta}_2$, we can now proceed to the evaluation of the genetic model. For this, we need to test the hypotheses $H_0: \theta_1 = 0$ and $H_0: \theta_2 = 0$. In addition, we need to compare θ_1 and θ_2 with each other and therefore define the null hypothesis as $H_0: \delta = \theta_1 - \theta_2 = 0$ [42, 41]. As δ is normally distributed with mean $\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_2$ and variance $\mathbb{V}\text{ar}(\hat{\delta}) = \mathbb{V}\text{ar}(\hat{\theta}_1 - \hat{\theta}_2) = \mathbb{V}\text{ar}(\hat{\theta}_1) + \mathbb{V}\text{ar}(\hat{\theta}_2) - 2\text{Cov}(\hat{\theta}_1, \hat{\theta}_2)$, we can use the Wald-type test statistic

$$T = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\widehat{\mathbb{V}\text{ar}}(\hat{\theta}_1 - \hat{\theta}_2)}},$$

which is asymptotically standard normally distributed under H_0 . A $1 - \alpha\%$ confidence interval for the difference $\delta = (\theta_1 - \theta_2)$ can be derived from

$$\hat{\delta} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathbb{V}\text{ar}}(\hat{\delta})}.$$

We can now apply the following decision rules, which are depicted in Figure 11.1. If $\theta_1 = 0$ and $\theta_2 > 0$, a recessive model is assumed. If $\theta_1 > 0$ and $\theta_2 > 0$ but $\theta_1 = \theta_2$, evidence is in favor of a dominant model. However, if $\theta_1 > 0$ and $\theta_2 > 0$ but $\theta_1 < \theta_2$, an additive model is likely. Finally, excess heterozygosity or heterozygote deficiency can also be deduced.

This model can be adopted to apply the approach by Minelli et al. [466]. These authors suggested to use the effect estimates from a meta-analysis to simultaneously estimate the genetic effect and the genetic model. Specifically, the magnitude of the genetic effect is estimated by OR_2 , whereas the ratio

$$\lambda = \frac{\log OR_1}{\log OR_2}$$

is estimated to describe the mode of inheritance. Here, values for λ of 0, 0.5, and 1 correspond to a recessive, an additive, and a dominant genetic model, respectively. Values of > 1 and < 1 suggest excess heterozygosity and heterozygote deficiency, respectively. However, if these can be excluded a priori, better estimates for λ are obtained by constraining it to lie between 0 and 1. To estimate λ , Eq. (11.10) can be modified to jointly estimate λ and θ_2 with $\theta_1 = \lambda\theta_2$. The model is then given by

$$\begin{pmatrix} \hat{\theta}_{1k} \\ \hat{\theta}_{2k} \end{pmatrix} \overset{a}{\sim} N \left(\begin{bmatrix} \lambda\beta_2 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} \sigma_{1k}^2 + \lambda^2\tau^2 & \sigma_{12,k} + \lambda\tau^2 \\ \sigma_{12,k} + \lambda\tau^2 & \sigma_{2k}^2 + \tau^2 \end{bmatrix} \right).$$

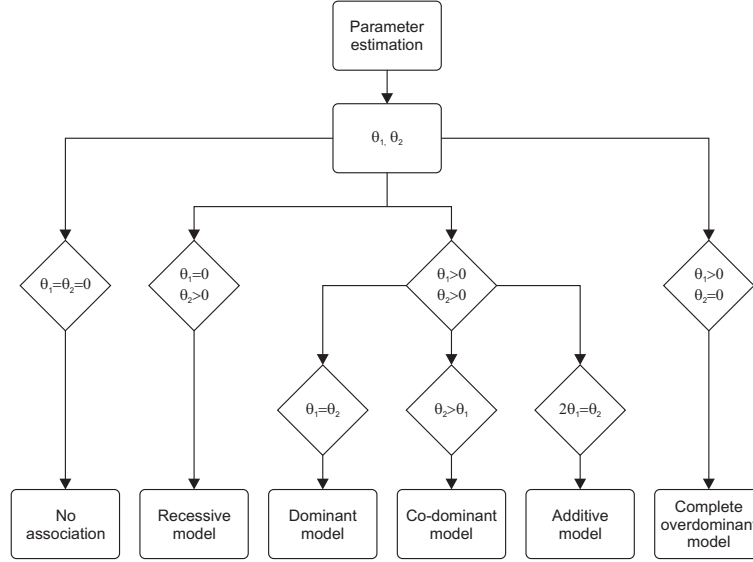


Fig. 11.1 Graphical representation of steps required for selecting the most plausible genetic model using logistic regression.

In this model, a single parameter τ is assumed, so that λ is estimated as a fixed effect. Using a Taylor series expansion and the delta method, Bagos [41] derived the approximate variance of $\hat{\lambda}$ as a function of θ_1 and θ_2 . This approach can be used for testing the null hypothesis $H_0: \lambda = 0$ using a Wald-type test statistic and for deriving an approximate confidence interval for λ . A Stata program METAGEN is available to perform the calculations.

Example 11.3. To illustrate the evaluation of a genetic model in a single study based on the approach by Hothorn and Hothorn [315], we use part of the data from the study by Erdmann et al. [191]. In this genome-wide association study on coronary artery disease, a three-stage approach was utilized to identify putative SNPs in the first genome-wide stage in one sample, test these in the second stage in three independent genome-wide samples, and finally validate findings in the third stage. Data from the first two stages are shown in Table 11.9 for the newly identified SNP rs9818870. For now, we focus on the data from the German Myocardial Infarction Family Study II (GerMIFS II) that served as the first stage.

Because T is the presumed risk allele, the three unstandardized test statistics for the dominant, additive, and recessive genetic models are given by

$$\begin{aligned}
 T_{dom} &= 796 \cdot 0.1 + 373 \cdot 1.1 + 52 \cdot 1.1 + 937 \cdot 0.0 + 333 \cdot 1.0 + 28 \cdot 1.0 = 425, \\
 T_{add} &= 796 \cdot 0.1 + 373 \cdot 1.1 + 52 \cdot 2.1 + 937 \cdot 0.0 + 333 \cdot 1.0 + 28 \cdot 2.0 = 477, \\
 T_{rec} &= 796 \cdot 0.1 + 373 \cdot 0.1 + 52 \cdot 1.1 + 937 \cdot 0.0 + 333 \cdot 0.0 + 28 \cdot 1.0 = 52.
 \end{aligned}$$

Table 11.9 Observed data from stages one and two of the association between SNP rs9818870 and coronary artery disease from Ref. [191].

Study	Cases			Controls		
	CC	CT	TT	CC	CT	TT
GerMIFS II	796	373	52	937	333	28
WTCCC	1328	538	59	2120	744	70
MIGen/IATVB	2159	738	70	2287	735	53
GerMIFS I	506	224	29	1148	432	54

The expected value and the variance of the influence are given by

$$\begin{aligned}
\mathbb{E}(y) &= \frac{1}{2519} \cdot (796 \cdot 1 + 373 \cdot 1 + 52 \cdot 1 + 937 \cdot 0 + 333 \cdot 0 + 28 \cdot 0) \\
&= 0.4847, \\
\mathbb{V}ar(y) &= \frac{1}{2519} \cdot (796 \cdot (1 - 0.4847)^2 + 373 \cdot (1 - 0.4847)^2 + 52 \cdot (1 - 0.4847)^2 \\
&\quad + 937 \cdot (0 - 0.4847)^2 + 333 \cdot (0 - 0.4847)^2 + 28 \cdot (0 - 0.4847)^2) \\
&= 0.2498.
\end{aligned}$$

From this, we can calculate the expected values of the three linear statistics to be

$$\begin{aligned}
\mathbb{E}(T_{dom}) &= 0.4847 \cdot (796 \cdot 0 + 373 \cdot 1 + 52 \cdot 1 + 937 \cdot 0 + 333 \cdot 1 + 28 \cdot 1) = 380.9869, \\
\mathbb{E}(T_{add}) &= 0.4847 \cdot (796 \cdot 0 + 373 \cdot 1 + 52 \cdot 2 + 937 \cdot 0 + 333 \cdot 1 + 28 \cdot 2) = 419.7642, \\
\mathbb{E}(T_{rec}) &= 0.4847 \cdot (796 \cdot 0 + 373 \cdot 0 + 52 \cdot 1 + 937 \cdot 0 + 333 \cdot 0 + 28 \cdot 1) = 38.7773.
\end{aligned}$$

Finally, with $\mathbf{x}_i$ being $\mathbf{x}_1 = \mathbf{x}_4 = (0, 0, 0)'$, $\mathbf{x}_2 = \mathbf{x}_5 = (1, 1, 0)'$, and $\mathbf{x}_3 = \mathbf{x}_6 = (1, 2, 1)'$, we can compute the covariance as

$$\begin{aligned}
\mathbb{V}ar(\mathbf{T}) &= \frac{2519}{2519-1} \cdot 0.2498 \cdot \\
&\quad \left(796 \cdot \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + 373 \cdot \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + 52 \cdot \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} + 937 \cdot \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + 333 \cdot \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + 28 \cdot \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} \right) \\
&\quad - \frac{1}{2519-1} \cdot 0.2498 \cdot \\
&\quad \left(796 \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}' + 373 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}' + 52 \cdot \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}' + 937 \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}' + 333 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}' + 28 \cdot \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}' \right) \\
&\quad \cdot \left(796 \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}' + 373 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}' + 52 \cdot \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}' + 937 \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}' + 333 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}' + 28 \cdot \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}' \right)' \\
&= \begin{pmatrix} 135.1137 & 148.8657 & 13.7520 \\ 148.8657 & 181.9722 & 33.1065 \\ 13.7520 & 33.1065 & 19.3544 \end{pmatrix}.
\end{aligned}$$

Thus, the test statistics for the dominant, the additive, and the recessive models are

$$\begin{aligned} T_{dom} &= \left| \frac{425 - 380.9869}{\sqrt{135.1137}} \right| = 3.7865, \\ T_{add} &= \left| \frac{477 - 419.7642}{\sqrt{181.9722}} \right| = 4.2429, \\ T_{rec} &= \left| \frac{52 - 38.7773}{\sqrt{19.3544}} \right| = 3.0056. \end{aligned}$$

Based on a three-dimensional normal distribution, the associated p -values are $1.5175 \cdot 10^{-4}$, $2.2233 \cdot 10^{-5}$, and $2.6234 \cdot 10^{-3}$ for the dominant, the additive, and the recessive models, respectively. In detail, the p -value for the dominant model is obtained by evaluating

$$P \left(\begin{pmatrix} X_1 \leq 425 \\ X_2 \leq 425 \\ X_3 \leq 425 \end{pmatrix} \middle| \mu = \begin{pmatrix} 380.9869 \\ 380.9869 \\ 380.9869 \end{pmatrix}, \Sigma = \begin{pmatrix} 135.1137 & 148.8657 & 13.7520 \\ 148.8657 & 181.9722 & 33.1065 \\ 13.7520 & 33.1065 & 19.3544 \end{pmatrix} \right)$$

for a three-dimensional normal distribution. The comparison of the p -values makes the additive model the most plausible for this study, although the other models are also significant at a 5% level. The respective results for the other three studies from stage 2 are given in Table 11.10. These show that for both WTCCC and MIGen, the additive model is the most likely, whereas GerMIFS I hints at a dominant mode of inheritance.

Table 11.10 Results for genetic models in the study on the association between SNP rs9818870 and coronary artery disease with p -values for the dominant (p_{dom}), the additive (p_{add}), and the recessive (p_{rec}) genetic models [191].

Study	p_{dom}	p_{add}	p_{rec}
GerMIFS II	$1.52 \cdot 10^{-4}$	$2.22 \cdot 10^{-5}$	0.0026
WTCCC	0.0149	0.0094	0.1430
MIGen/IATVB	0.1500	0.0786	0.0809
GerMIFS I	0.0775	0.0851	0.4249

Included are the German Myocardial Infarction Family Study (GerMIFS) I and II, the Wellcome Trust Case Control Consortium (WTCCC), and the Myocardial Infarction Genetics Consortium with the Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group (MIGen/IATVB).

Estimating the genetic model in the meta-analytic framework, the approaches by Minelli et al. [466] and by Bagos and Nikolopoulos [42] have been described above. Details of their application are not presented here but, for comparison, their results from combining the data across all four studies are. The estimated log ORs with their respective 95% confidence intervals from the model are $\hat{\theta}_1 = 0.2440$ [0.0208; 0.4672] and $\hat{\theta}_2 = 0.3913$ [0.1437; 0.6389]. Since both θ_1 and θ_2 are greater than 0, a recessive model can be excluded. Furthermore, the difference is estimated to be

Table 11.11 Observed genotype frequencies affected cases and unaffected controls at an X-linked diallelic marker with alleles a and A.

	Female				Male		
	aa	aA	AA	Total	a	A	Total
Cases	r_{f0}	r_{f1}	r_{f2}	r_f	r_{m0}	r_{m1}	r_m
Controls	s_{f0}	s_{f1}	s_{f2}	s_f	s_{m0}	s_{m1}	s_m
Total	n_{f0}	n_{f1}	n_{f2}	n_f	n_{m0}	n_{m1}	n_m

$\hat{\delta} = 0.1473 [0.0491; 0.2456]$ showing that $\theta_2 > \theta_1$. Thus, this approach is in favor of an additive mode of inheritance. Similarly, $\hat{\lambda} = 0.6235 [0.3631; 0.8839]$, therefore excluding both recessive and dominant genetic models and making the additive model most likely.

11.2.5 How can association on the X chromosome be tested?

SNPs on the pseudo-autosomal regions PAR1 and PAR2 of the X chromosome can be treated identical as SNPs on an autosome (see Chapter 1). However, they need to be analyzed differently for other regions because males carry only one allele at an X chromosomal SNP, while females carry two copies. The possible case-control data for an X-linked SNP are displayed in Table 11.11. For autosomal SNPs, we have argued in Section 11.2.2 that the Cochran–Armitage trend test should be the standard association test. Here, we formulate a similar test statistic for SNPs on the X chromosome. However, we need to discuss how males should be scored because they carry only one allele.

The most intuitive scoring for males would be 0 and 1. However, as discussed in detail by Clayton [127], most loci are subject to X inactivation in females [121], so that a female has approximately half her cells with one copy active while the remainder of the cells have the other activated. Thus, in the absence of interaction with other loci or environmental factors, males should be equivalent to homozygous females for these SNPs. This suggests that males should be scored 0 and 2. Thus, the numerator of the trend test statistic of Eq. (11.6) can be used for X-chromosomal markers. However, a different denominator has to be employed because of different variances in males and females. Specifically, three genotypes are possible in females, while only two alleles can be present in males.

In detail, in the first step we assume that the allele frequency p is identical in males and females. The genotype distribution in females may deviate from HWE. The total sample size is n , phenotypes are denoted by y_i , and genotypes x_i of female subjects are scored 0, 1, and 2 according to the number of variants. Male genotypes x_i are accordingly scored 0 and 2. Since allele frequencies are assumed to be identical between males and females, $\hat{p}$ is estimated from the entire sample, i.e., $\hat{p} = (n_{f1} + 2n_{f2} + n_{m1}) / (2n_f + n_m)$. Similarly, we consider the overall phenotypic sample mean, which is estimated by $\bar{y} = (r_f + r_m) / (n_f + n_m)$.

The numerator of the trend test statistic (Problem 11.5) is given by $S = \sum_{i=1}^n (y_i - \bar{y})x_i$. This expression cannot be substantially simplified because the overall sample mean $\bar{y}$ is considered. The variance $\mathbb{V}ar(S)$ of S can be estimated by

$$\widehat{\mathbb{V}ar}(S) = \frac{1}{n_f} \sum_{i=1}^{n_f} (y_i - \bar{y})^2 \sum_{i=1}^{n_f} (x_i - \bar{x})^2 + 4\hat{p}(1 - \hat{p}) \sum_{i=1}^{n_m} (x_i - \bar{x})^2.$$

The trend test statistic finally is given by $T = \frac{S^2}{\widehat{\mathbb{V}ar}(S)}$ (Problem 11.5), which is asymptotically χ_1^2 distributed under the null hypothesis of no association [127].

If allele frequencies are gender-specific, i.e., if they differ between males and females, the approach described above is invalid. In this case, the numerator of the trend test statistic for females is $S_f = \sum_{i=1}^{n_f} (y_i - \bar{y}_f)x_i$ with variance estimated by $v_f = \widehat{\mathbb{V}ar}(S_f) = \frac{1}{n_f} \sum_{i=1}^{n_f} (y_i - \bar{y}_f)^2 \sum_{i=1}^{n_f} (x_i - \bar{x})^2$. For males it is given by $S_m = \sum_{i=1}^{n_m} (y_i - \bar{y}_m)x_i$ with variance estimated by $v_m = \widehat{\mathbb{V}ar}(S_m) = 4\hat{p}_m(1 - \hat{p}_m)$. Here, $\bar{y}_f = r_f/n_f$ is the proportion of cases in females and $\bar{y}_m = r_m/n_m$ is the proportion of cases in males. Furthermore, $\hat{p}_m = r_{m1}/n_{m1}$ is the allele frequency estimated only in males.

Two approaches for combining males and females in a single test statistic have been proposed by Thompson [662]. The first is similar to Hotelling's T^2 that has less power than the *directional test* we consider here. In the directional test, males and females are combined inversely proportional to their variance. Specifically, let $w_f = 1/v_f$ and $w_m = 1/v_m$. Then, under the null hypothesis of no association, $w_f S_f$ and $w_m S_m$ are both asymptotically normally distributed with variance w_f and w_m , respectively. Hence,

$$T_{f,m} = \frac{(w_f S_f + w_m S_m)^2}{w_f + w_m}$$

is asymptotically χ_1^2 distributed under H_0 . We stress that this test statistic is required only if the gender ratio differs between cases and controls. This stratification of the analysis by gender can lead to a considerable loss of power. When the allele frequency does not differ between females and males, this stratification by gender, with its loss in power, is unnecessary.

This approach is different from the combination proposed in Ref. [753] in which the authors do not weight females and males inversely proportional to their variance. They first form two separate asymptotically normally distributed test statistics, i.e., one for females Z_f and one for males Z_m . In one approach they use $Z_f^2 + Z_m^2$ that is asymptotically χ_2^2 distributed. Here, neither the different variability of female and males nor the different sample size is taken into account. In the second approach, they suggest weighting Z_f for females and Z_m for males inversely proportional to their sample size. However, this does not reflect the different variability between females and males. They also propose test statistics when the risk allele is different in males and females. However, we do not see any biological reason for such an effect and therefore refrain from describing this approach.

11.3 WHAT SAMPLE SIZE IS REQUIRED TO TEST FOR ASSOCIATION?

For calculating required sample sizes, the different causal models leading to genetic association have to be taken into account. Specifically, in the case of direct association, we assume that the genetic marker under investigation is identical to the disease locus and that the risk allele at the marker is directly causative for the disease status. In this situation, the sample size calculation is straightforward, as usual formulae can be utilized. Consider, for instance, the situation of comparing allele frequencies in a case-control study. To detect a pre-specified OR_A based on an allele frequency $p_{A,u}$ in the unaffected controls with power $1 - \beta$ at a significance level α , the required total sample size is approximately given by [592]

$$n \geq 2 \cdot \frac{\left(z_{1-\alpha/2} \sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta} \sqrt{p_{A,a}(1-p_{A,a}) + p_{A,u}(1-p_{A,u})} \right)^2}{(p_{A,a} - p_{A,u})^2},$$

with the allele frequency in the affected cases $p_{A,a}$ and the mean allele frequency $\bar{p}$:

$$p_{A,a} = \frac{p_{A,u} \cdot OR_A}{1 + p_{A,u} \cdot (OR_A - 1)} \quad \text{and} \quad \bar{p} = \frac{1}{2}(p_{A,a} + p_{A,u})$$

An example using the above formula is given at the end of this section. If sample size calculations should be based on the genotype distribution and not on OR_A , they can be done within the framework of the Cochran–Armitage trend test. This has been described in detail by Slager and Schaid [619].

In the case of indirect association, the approach from above needs to be slightly extended. Here, it is assumed that the association with a marker allele is observed because the marker locus is in LD with the disease locus, which is, in turn, the causal determinant for the disease. Therefore, the sample size depends on the effects at both the marker and the disease locus and on the LD between the two loci. Coming back to the sample size calculation for the allelic odds ratio OR_A , the OR at the marker locus OR_A can be calculated from the OR at the disease locus OR_D , if the allele frequencies at both loci and the LD are known, by

$$OR_A = 1 + \frac{\mathcal{D}(OR_D - 1)}{p_{A,u} \left((1 - p_{A,u}) + (p_{D,u}(1 - p_{A,u}) - \mathcal{D})(OR_D - 1) \right)},$$

with $\mathcal{D}$ giving the LD between the loci and $p_{A,u}$ and $p_{D,u}$ as the allele frequencies at the marker and the disease locus, respectively. Utilizing this, the sample size may then be estimated as before based on the OR and allele frequencies at the marker locus. Resulting sample sizes for this situation are presented in Table 11.12. For estimating the required sample size based on the trend test by Armitage, an extension of the methodology by Slager and Schaid [619] has been proposed by Pfeiffer and Gail [532]. We conclude this section by illustrating the sample size calculation approach. Another example is given in Problem 11.6.

Table 11.12 Required sample sizes to detect an odds ratio OR_D for the disease allele D with an allele frequency $p_{A,u}$ and disease allele frequency $p_{D,u}$ in unaffected controls at $\mathcal{D} = \mathcal{D}_{\max}$, a significance level $\alpha = 0.0001$, and a power of 80%.

$p_{A,u}$	OR_D	$p_{D,u}$				
		0.05	0.1	0.2	0.3	0.4
0.05	1.2	52,354	59,196	77,124	103,648	145,092
	1.5	9,639	11,122	15,062	20,998	30,431
	2	2,944	3,502	5,013	7,342	11,122
	3	84	85	87	88	90
0.1	1.2	105,528	27,893	36,366	48,904	68,502
	1.5	18,338	5,200	7,055	9,855	14,307
	2	5,200	1,616	2,325	3,420	5,200
	3	109	105	106	109	111
0.2	1.2	231,816	59,936	15,986	21,532	30,207
	1.5	38,997	10,556	3,053	4,284	6,245
	2	10,556	3,053	982	1,459	2,238
	3	173	167	167	166	168
0.3	1.2	394,186	101,134	26,592	12,409	17,442
	1.5	65,559	17,442	4,901	2,427	3,557
	2	17,442	4,901	1,510	806	1,251
	3	294	287	289	301	290
0.4	1.2	610,679	156,065	40,733	18,877	11,060
	1.5	100,974	26,624	7,364	3,599	2,214
	2	26,624	7,364	2,214	1,160	758
	3	620	618	644	691	750

Example 11.4. Consider the case of comparing allele frequencies between cases and controls. When the disease-causing variant is investigated directly, an $OR_A = 2$ is assumed for an allele with a frequency of $p_{A,u} = 0.1$ in the control population. Hence, the expected allele frequency in the cases and the mean allele frequency are

$$p_{A,a} = \frac{0.1 \cdot 2}{1 + 0.1 \cdot (2-1)} = 0.18 \quad \text{and} \quad \bar{p} = \frac{1}{2}(0.18+0.1) = 0.14.$$

To detect this effect in a two-sided test with a power of $(1 - \beta) = 0.8$ at a significance level of $\alpha = 0.05$,

$$n \geq 2 \cdot \frac{\left(1.96\sqrt{2 \cdot 0.14(1-0.14)} + 0.84\sqrt{0.18(1-0.18) + 0.10(1-0.10)}\right)^2}{(0.18-0.10)^2} \geq 566$$

case and control alleles are required. It should be noted that because allele frequencies are tested in this case, this refers to the number of alleles, not individuals. Hence, 283 case and 283 control alleles are needed, leading to the recruitment of 142 cases and 142 controls.

What would be the consequence if, instead of the disease locus, a neighboring marker were genotyped? Assuming that the allele frequency at the marker locus $p_{A,u} = 0.3$, the maximal LD between the disease and the marker locus is given by

$\max(\mathcal{D}) = \mathcal{D}_{max} = \min \{0.1(1 - 0.3), 0.3(1 - 0.1)\} = 0.07$. In the situation that the LD is maximal, the OR at the marker locus would therefore be reduced to

$$OR_A = 1 + \frac{0.07(2-1)}{0.3[(1-0.3) + (0.1(1-0.3) - 0.07)(2-1)]} = 1.33.$$

Fitting OR_A and $p_{A,u}$ into the sample size equation demonstrates the effect: instead of 566 case and control alleles, 1720 are required in this setting of indirect association, corresponding to 430 cases and 430 controls.

11.4 WHAT IS POPULATION STRATIFICATION, AND WHAT CAN BE DONE?

In describing the selection of cases and controls for a study on genetic association (see Section 11.1), it has already been emphasized that the possibilities for bias need to be minimized. A prominent bias, known as *population stratification*, has long led to case-control studies being shunned in favor of other study designs. To understand this problem, let us turn back to reconsider the models for association to be observed as shown in Figure 10.1. In the ideal case of direct association, the marker locus is identical to the disease locus; hence, an association reflects the causal relationship. In the model of indirect association, the association between marker allele and phenotype is observed because the marker is in LD with the disease locus, which in turn causes the phenotype. However, there is a third situation in which association might be observed, which is if the marker allele is associated with a confounding variable that in turn is causative for the disease. In general, a *confounder* is a variable that is not the focus of the study and that is causative for the phenotype while at the same time being associated with the variable under investigation, in this case, the marker allele. If the confounder is the ethnic affiliation of the individuals in the study, the effect is called *population stratification*, also termed *confounding by ethnicity*. This can occur if two conditions are met: 1) cases and controls come from different ethnical populations, and 2) the investigated marker alleles are unequally distributed in these populations. It should be noted that in addition to yielding a false-positive result, this confounding might also mask a true association, thus reducing the power to detect a genetic effect.

An illustrative fictitious example is given by the story of the SUSHI gene that has been told by Hamer and Sirota [281]. Imagine that you are investigating the genetic basis for eating with chopsticks. As a researcher in New York, you are collecting cases who eat with chopsticks to compare with controls who do not. Indeed, you find a significant association of the case-control status with marker alleles from a candidate region that has previously been shown to be associated with other behavioral phenotypes. You tell a colleague from San Francisco about your exciting result, and he also sets up a sample of cases and controls, validating your results. After this success, you term the locus the SUSHI (successful-use-of-selected-hand-instruments) gene. However, after some time, you have to withdraw your conclusion based on two facets of your study. First, your New York and San

Francisco cases were primarily of Asian origin, whereas the controls were mostly Caucasians. Second, your SUSHI locus is nothing more than the human leukocyte antigen (HLA) locus. As such, it is distributed differently in Asian and Caucasian individuals.

A classical real-life example of population stratification in case-control studies has been given by Knowler and colleagues [368]. In a study on Pima Amerindians, an association was reported between type II diabetes and the *Gm* locus. However, type II diabetes is generally more frequent in Pima Amerindians than in Caucasians, whereas the allegedly protective allele at the *Gm* locus is less frequent in Pima Amerindians than in Caucasians. The important point was that a genetically heterogeneous group of Pima Amerindians was investigated: individuals in the case and the control groups differed with regard to their degree of Caucasian ancestry.

Example 11.5. A numerical example of population stratification is used for illustration. To understand how it can lead to erroneous results, consider the data given in Table 11.13. Here, for analysis of some disease, cases and controls were collected from a large population 1 and a smaller population 2. Both were genotyped for a polymorphism at a specific candidate locus. For the statistical analysis, the frequencies of mutation carriers are compared between cases and controls, and the odds ratio is calculated to give an estimate of the strength of the association.

Table 11.13 Illustrative example from a case-control study in two different populations.

	Population 1		Population 2	
	≥ 1 mutation	No mutation	≥ 1 mutation	No mutation
Cases	10	90	40	20
Controls	90	810	20	10

In both populations, the OR is estimated to be

$$\widehat{OR}_1 = \frac{10 \cdot 810}{90 \cdot 90} = 1.0 \quad \text{and} \quad \widehat{OR}_2 = \frac{40 \cdot 10}{20 \cdot 20} = 1.0.$$

Hence, no association of the allele with the disease is visible. However, if both populations are pooled for analysis, the OR becomes $\widehat{OR} = \frac{(10+40) \cdot (810+10)}{(90+20) \cdot (90+20)} \approx 3.3884$.

Why has this happened? The answer is generally summarized under the term Simpson's paradox and can be explained by first noting that the frequency of mutations substantially differs in populations 1 and 2, being $(10+90)/(90+10+810+90) = 0.1$ in population 1 and $(40+20)/(20+40+10+20) = 0.7$ in population 2. Second, the populations differ with regard to the frequency of the disease: the prevalence is $(90+10)/(90+10+810+90) = 0.1$ in population 1 but $(20+40)/(20+40+10+20) = 0.7$ in population 2.

The actual extent of population stratification in typical case-control studies is the topic of an ongoing debate. Instead of reiterating the arguments here, the interested

reader is referred to the literature [661, 681]. In planning a specific case-control study, however, it will be important to know what extent of population differentiation can be expected in the area from which cases and controls are to be drawn. For this, recent works that have investigated the genetic structure within German [630], European [400], and European American populations [602] are helpful. These also produced sets of ancestry informative markers that can be specifically used to evaluate the ancestral structure in a given sample.

The classical epidemiological approach to deal with possible population stratification is to use self-reported or other measures of ancestry to match cases and controls accordingly. This, however, can be undermined by two factors. First, the reported ethnicity may be culturally rather than genetically defined. Second, there might be a *cryptic relatedness* that is difficult to detect using merely visible characters. Specifically, it may be the case that there is more hidden relatedness among the cases than among the controls, simply because they share a genetically determined disorder. This renders the assumption of independent probands invalid and may bias statistics [162].

To solve the problem of possible population stratification, three general strategies are possible. First, one might refrain from conducting association studies with unrelated probands in the first place. Instead, for some time family-based association study designs have been favored. These will be described in Chapter 12. Second, investigators might be able to recruit a genetically homogeneous sample in a certain area. In analogy to a multicenter study, samples from different centers can be jointly analyzed by taking the center information into account. However, instead of the center information, information about the ethnic group might be more relevant. Again, this information can be utilized. In standard statistical analysis, this may lead to the use, e.g., of the Cochran–Mantel–Haenszel approach or a logistic regression with adjustment for center or ethnic group.

Thirdly, and more recently, specific suggestions have been made on how to analyze the extent of and directly deal with population stratification in a given study. We will briefly describe these in the remaining part of this chapter. Typically, these approaches require the genotyping of additional genetic markers, the so-called *neutral* or *null markers*. These are markers that are not themselves associated with the disease of interest and are not in LD with predisposing alleles. Also, they should be in linkage equilibrium with each other. The underlying idea is that if there is population stratification, it should be visible in these null markers as well. Based on this principle, three now classical approaches have been proposed. More recently, however, these methods have also been applied without the separate genotyping of null markers but within a genome-wide context. Specifically for this situation, a fourth procedure has been suggested that is also described below.

11.4.1 How can we test for population stratification?

The first approach was suggested by Pritchard and Rosenberg [545] and is based on the following rationale: if cases and controls in the sample differ with regard to their ethnic background, there should be a consistent pattern of allele frequency

differences at many loci in the genome. In contrast, if they are ethnically similar, differences should occur only at markers close to the actual disease susceptibility loci. Hence, the suggestion is to first test for genetic association at the candidate loci as before. If association is found, additional null markers are genotyped; simulations have shown that about 15 to 20 STR markers should be sufficient. If SNPs are typed, this number needs to be higher. These are then tested for indications of population stratification. Consider that such unlinked L markers have been genotyped. For diallelic markers, differences in genotype or allele frequencies between cases and controls can be investigated for each of the L markers as described in Section 11.2.1. Then, to test for stratification across all null loci, the sum of the test statistics

$$\chi_P^2 = \sum_{i=1}^L \chi_i^2$$

is formed, with χ_i^2 being the single χ^2 statistics. The resulting statistic is asymptotically χ^2 distributed with L d.f. If there is no indication of population stratification by this test, the association with the candidate loci can be assumed to be valid. However, if population stratification is present, the results need to be interpreted with caution, and further measures need to be taken.

Applying this method to different data sets with matched control groups, Ardlie and colleagues [31] found that a careful matching eliminated the observed population substructure. In contrast, Freedman and colleagues [224] reported in their analysis of previously assembled case-control data sets that although no statistically significant evidence for stratification could be found, a few null markers were not enough to rule out even moderate levels of stratification. Increasing the number of markers and individuals led to the detection of stratification; therefore, they suggested always monitoring for stratification.

11.4.2 What is the approach of structured association?

What should be done after population stratification has been detected? One possibility is to employ the approach of *structured association* (SA) that was suggested by Pritchard and colleagues [547]. Basically, this is a two-step procedure. The first step includes inferring details of the population structure in the sample based on genotype correlations among unlinked markers. For this, a Bayesian clustering method is employed [546] that uses the individuals' genotypes to assign every person to a subpopulation. About 100 STRs or even more SNPs need to be genotyped for this purpose. Within these subpopulations, the marker loci have to be unlinked, be in linkage equilibrium with each other and be in HWE. Hence, within a subpopulation, no further structure is present.

To estimate the number of subpopulations, the allele frequencies in each subpopulation, and the population genetic background for each individual, a Markov chain Monte-Carlo method is used. The accuracy of the derived population structure depends on the sample size, the number of unlinked loci, the degree of admixture in the sample, and the extent of allele frequency differences between the populations. To perform the analysis, the software STRUCTURE is available.

In the second step of the SA approach, the derived information is used to test for association within subpopulations. The underlying idea is that if the sampled individuals have been assigned to unstructured subpopulations, any association between the candidate allele and the phenotype within a subpopulation could not be due to population structure. Consequently, the resulting test for association is valid even in the presence of population stratification. Specifically, for this test the standard null hypothesis of no overall association is replaced with a null hypothesis of no association within the subpopulations. As statistical test, a likelihood ratio test can be employed that contrasts the likelihood of the observed genotype data given the allele frequencies in the subpopulations under the alternative hypothesis and the population genetic background of the individuals with the respective likelihood under the null hypothesis. Here, the population genetic background as estimated in the previous step is included, and the allele frequencies in the subpopulations are estimated using an EM algorithm. The significance of the statistic is approximated via simulations, and the authors termed the test STRAT for STRUCTURED population Association Test. The respective software is also available.

Following the first suggestion by Pritchard and colleagues, more similar approaches were proposed as extensions or improvements. For example, Satten and colleagues [579] utilized a latent-class method aimed at simultaneously detecting and accounting for population stratification. Instead of a Bayesian framework, they used a maximum likelihood method. Their work was extended by Purcell and Sham [549] for more general situations (software L-POP). A slightly different procedure with the software ADMIXMAP was suggested by Hoggart and colleagues [300], who specifically allowed for the uncertainty in the estimations. How, in practice, these methods compare with each other is still unresolved.

11.4.3 What is the concept of genomic control?

The third approach to deal with population stratification relies on the genotyping of the so-called *genomic controls* (GC) [162]. Here, it is assumed that in the case of population stratification, the false-positive rate will be increased when the substructure in the sample population is ignored. Instead of aiming at directly reconstructing these substructures as in SA, this inflation is corrected for in the test statistic. For this, a number L of diallelic null loci are genotyped in addition to the markers at candidate loci. At each of these markers, a test statistic that is approximately χ^2_1 distributed as in Eqs. (11.4) and (11.5) is calculated. Under the null hypothesis of no population stratification, the expected value of the respective statistics will be 1. From this, an inflation factor λ is estimated that mirrors the deviation from the expectation. Different estimates for λ have been suggested by Devlin and Roeder [162] for an additive genetic model and by Zheng and colleagues [750] for dominant and recessive genetic models. With χ^2_i as the test statistic of the i th null locus, a simple and robust estimate for the additive model is given by [38]

$$\hat{\lambda} = \frac{\text{median}(\chi^2_1, \chi^2_2, \dots, \chi^2_L)}{0.456},$$

where the value 0.456 is the 50% quantile of the χ_1^2 distribution. The test statistic for any candidate locus c is then corrected using $\hat{\lambda}$ with $\chi_{GC}^2 = \frac{\chi_c^2}{\hat{\lambda}}$ and is again asymptotically χ_1^2 distributed. To account for sampling variability, it has also been suggested to bound $\hat{\lambda}$ by using $\max(\hat{\lambda}, 1)$ instead.

Since λ scales with sample size [162], it has been suggested to report λ_{1000} , which is the value that would be expected in a study of 1000 cases and 1000 controls [224]. This can be easily computed from

$$\lambda_{1000} = 1 + (\hat{\lambda} - 1) \cdot \left(\frac{1/n_{cases} + 1/n_{controls}}{1/2000} \right).$$

The procedure relies on the important assumption that the inflation factor is constant across the genome, and Devlin and Roeder [162] have demonstrated that this is usually the case. Simulation studies have suggested that for estimating $\hat{\lambda}$ reliably, about 20–50 additional SNPs are required [38]. GC often performs well but may become overly conservative if population structure is more extreme. However, even modest levels of population structure cannot be ignored [440]. For applications, software has been made available, and GC may be used for both SNPs and STRs [163].

If many genotypes are available as from a genome-wide association study, it is helpful to create a *quantile-quantile (Q-Q) plot*. This is obtained from ranking the observed test statistics from smallest to largest and to plot these against their expected values under the null hypothesis of no association [131]. A typical example for this is shown in Figure 11.2. In this graph, it becomes clear that only a few SNPs extremely deviate from the expected line, and these might represent markers truly associated with the investigated phenotypes. However, it can also be seen that, apart from those SNPs, the general trend of the observed values is slightly above the

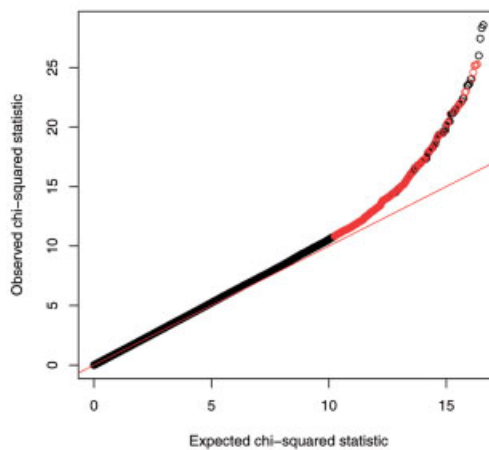


Fig. 11.2 Quantile-quantile (Q-Q) plot depicting observed versus expected test statistics under the null hypothesis of no association. Red dots denote single nucleotide polymorphisms in regions known to be associated. The overall inflation factor λ estimated from the linear regression is 1.04 (95% confidence interval [1.0398; 1.0402]). Data are taken from Ref. [191], which is described in more detail in Example 11.3.

line of the expected values. This is therefore an inflation of the test statistics that might, in addition to other factors, be due to population stratification. Based on this observation, an alternative method to estimate the overall inflation factor λ obviously is to linearly regress the observed on the expected values and use the slope of the regression line as an estimate. Because this will be biased upward when all SNPs are included, i.e., also those that are truly associated, it has been recommended to exclude the top 10% for this estimation [131].

11.4.4 How do genomic control and structured association compare with each other?

In comparing the approaches, there seem to be two crucial points that determine which might be more powerful. For one, if the population substructure is intricate and thus difficult to infer, SA might yield invalid tests, and GC fares better. In contrast, if the substructure can be estimated well, SA may be more powerful. This issue depends on the number of neutral loci genotyped because with a higher number, the structure will be easier to reconstruct. The second point to consider is the distribution of the genetic effect. If this is more or less consistent across the different subpopulations, GC will usually be better. However, if the effects vary from population to population, this can be detected more reliably using SA [544].

These points were substantiated by an empirical comparison using genotype data of more than 10,000 SNPs in four populations [285]. They found that if substantial population stratification was present, both SA and GC were sensitive enough to detect it based on only 20–50 null markers, rendering the SA approach more attractive because it yields more information. If stratification was less extreme, both approaches had difficulties in detection, but GC had the advantage of controlling for it.

11.4.5 How can principal components analysis be used to adjust for population stratification?

In the previous sections, approaches were described that had been developed in the context of candidate gene association studies. Thus, they require that, in addition to the markers of interest, null markers are genotyped that are subsequently used to estimate and adjust for population stratification. In recent years, however, these methods have also been applied in genome-wide association (GWA) studies using all available genotype data at hand. For this specific application, a different procedure is used that explicitly aims at detecting and correcting for population stratification on a genome-wide scale [541]. Specifically, it incorporates principal components analysis (PCA) to infer continuous axes of genetic variation, which more adequately mirrors the variation within a continent than distinct clusters.

This approach consists of the following steps:

1. Inference of variation: PCA is applied to genotype data to identify axes of genetic variation. If there are ancestry differences between the samples, these often can be geographically interpreted.

2. Adjustment for ancestry variation: Residuals of linear regressions are computed to adjust for the variation along each axis. This renders a virtual set of matched cases and controls.
3. Association analysis: Association statistics are computed using the adjusted data.

For applications, it is important to know how many probands and markers are required for an accurate identification of the axes. This crucially depends on the extent of divergence between the populations as measured by F_{ST} . Complementary to the inbreeding coefficient F_{IS} that was defined in Section 4.3.4, this coefficient gives the effect of subpopulations S compared to the total population T . In our context of PCA, it was shown that with a sample size n and a number of markers m , statistical evidence for $F_{ST} = 1/\sqrt{mn}$ can be reliably obtained [521]. Thus, if there is a sample comprising 1000 cases and 1000 controls so that $n = 2000$, $m = 1715$ SNPs need to be genotyped to detect an F_{ST} of $5.4 \cdot 10^{-4}$ as plausible for a German population [630]. Looked at in a different way, using 100,000 SNPs from a genome-wide genotyping chip, an F_{ST} of as small as $7.1 \cdot 10^{-5}$ could be detected with the same sample size.

Using the data from a GWA study theoretically introduces the problem of dependence between the markers, and high LD could lead to an over-representation of some genetic regions that in turn would bias the identified structure, as shown in Ref. [521]. On the other hand, Price and colleagues [541] investigated the impact of the LD structure on the inferred axes and found it to be only marginal. Another challenge from basing the approach on all SNPs from a GWA study is that some of these SNPs are not random but truly associated with the phenotype. But again, it was shown that the method is robust with regard to including or excluding these SNPs.

The PCA approach is implemented within the EIGENSOFT package. An alternative implementation termed SHELLFISH has been developed by Dan Davison, which allows the computations to be performed in parallel.

How does the PCA compare with the other approaches? In extensive simulations, Price et al. found that it yields a lower frequency of false-positive results than the genomic control procedure while achieving greater power [541]. Compared to structured association, it needs to be pointed out that the aims are different: Structured association classifies all individuals into discrete populations or linear combinations of these populations, whereas the PCA describes every proband with regard to the coordinates on the inferred axes. An interesting possibility would be to combine these methods by using the PCA to identify the number of significant eigenvalues and then to use this to define the number of clusters for structured association.

At the end of this section, we point out that many other methods to evaluate population stratification have been proposed lately that are not detailed here, including multidimensional scaling (MDS) [409] or non-metric multidimensional scaling (NMDS) [758]. Still other procedures include a logistic regression approach [606], a δ -centralization [258], or a stratification-score method [188]. A final evaluation of all approaches is still pending.

11.5 WHAT IS INTERACTION?

In Section 2.3 it was emphasized that complex phenotypes are expected to be determined not by single genetic variants but by an intricate interplay of a number of genetic and environmental factors. Consequently, this section is devoted to gene-gene (G-G) and gene-environment (G-E) interactions.

To describe G-G interaction, often the term *epistasis* is used. Epistasis is a composition of “epi” (Greek *επι*, upon, above, besides, among) and “stasis” (Greek *στασις*, standing, stop). Stasis refers to a state of stability in which all forces are equal and opposing, therefore they cancel out each other. Together with the prefix epi, epistasis means a masking effect whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect. Indeed, this was the original definition used by Bateson [48], and an example is the coat color of Section 11.5.2. Here, the effect of the brown locus is suppressed by the extension locus if the dog is homozygous ee. However, for other models of G-G interaction, the term epistasis is inappropriate, and therefore we recommend using the terms G-G interaction and epistasis interchangeably. Finally, G-G interaction often refers to interaction between different loci, i.e., loci on different chromosomes or, at least, with great physical distance. Interaction can, however, also occur between alleles at the same locus. In this case, haplotypes are considered, and this within-gene interaction is typically not termed G-G interaction.

11.5.1 What are classical examples for gene-gene and gene-environment interactions in humans?

In the simplest case of G-G or G-E interaction, only the presence or absence of a variant or an exposure needs to be considered. Three classical examples for G-E interaction meeting this simple representation are displayed in Table 11.14 that has been adapted from Refs. [325, 352]. The first example is xeroderma pigmentosum (XP). Exposure to ultraviolet light increases the risk of developing skin cancer in non-carriers of XP mutations, but the combination of these mutations and exposure

Table 11.14 Classical examples for gene-environment (G-E) interaction in humans.

Genetic variant	Environmental exposure	Relative risk		
		XP	PKU	Emphysema
Absent	Absent	1.0	1.0	1.0
Present	Absent	1.0	1.0	Modest
Absent	Present	Modest	1.0	Modest
Present	Present	Very high	Very high	High

XP: xeroderma pigmentosum; PKU: phenylketonuria.

Examples have been given in Ref. [352]; for the table, see Ref. [325].

to ultraviolet light vastly increases the risk of skin cancer. Phenylketonuria (PKU) is the second example. Only subjects with recessive mutations in the functionally relevant gene (phenylalanine hydroxylase) that are exposed to phenylalanine in the diet are susceptible to PKU. The third example is on emphysema. Individuals with a deficiency in the α -1 antitrypsin gene are at an increased risk of developing emphysema. Smokers also have an increased risk, but smokers with a deficiency in the α -1 antitrypsin gene are at a high risk of developing emphysema.

Table 11.15 Selected examples of gene-environment interactions observed in at least two studies.

Gene symbol	Variant(s)	Environmental exposure	Outcome and nature of interaction
<i>CCR5</i>	Δ -32 deletion	HIV	Lower rates of HIV infection and disease progression in carriers of the receptor deletion
<i>MTHFR</i>	Ala222Val	Folic acid intake	Different risk of colorectal cancer and adenomas in Ala222Val homozygotes if nutritional folate status is low
<i>UGT1A6</i>	Slow-metabolism SNPs	Aspirin	Increased benefit of prophylactic aspirin use in carriers of the slow metabolism variants
<i>APOE</i>	E4 allele	Cholesterol intake	Exaggerated changes in serum cholesterol in response to dietary cholesterol changes in APOE4 carriers
<i>ADH1C</i>	γ -2 alleles	Alcohol intake	Negative association between ethanol intake and myocardial infarction; risk is higher in carriers of slow-oxidizing γ -2 alleles
<i>ADRB2</i>	Arg16Gly	Asthma drugs	Greater response in the airway to albuterol in Arg16Gly homozygotes
<i>MDR1</i>	c.3435C/T (Ile1140Ile)	Pesticides	Increased risk of developing Morbus Parkinson when exposed to pesticides 3435C in homozygotes, see Refs. [91, 772]
<i>CYP2C19</i>	CYP2C19*2	Treatment with clopidogrel	Higher risk of subsequent cardiovascular events for patients with an acute myocardial infarction who were receiving clopidogrel when they are homozygous for the CYP2C19*2 loss-of-function allele, see Refs. [136, 617]

Table adapted from Ref. [325]. *ADH1C*, alcohol dehydrogenase 1C (class I), γ -polypeptide; *ADRB2*, adrenergic, β -2-, receptor, surface; *CCR5*, chemokine (C-C motif) receptor 5; *APOE*, apolipoprotein E; *MTHFR*, 5, 10-methylenetetrahydrofolate reductase (NADPH); *UGT1A6*, UDP glycosyltransferase 1 family, polypeptide A6; *MDR1*, multidrug resistance protein 1; *CYP2C19*, cytochrome P450, family 2, subfamily C, polypeptide 19.

In addition to these classical examples for G-E interaction, Table 11.15 gives a series of examples for G-E interactions observed in at least two studies.

A series of examples for G-G interactions has been given in Ref. [535]. However, most of them suffer from a lack of replication. This means that the G-G interaction was identified in one study but either not studied or not detected in at least another independent sample. The few G-G interactions that have been observed in at least two studies include the interactions between the *DLG5* (discs, large homolog 5) and the *NOD2* gene (nucleotide-binding oligomerization domain containing 2) with Morbus Crohn [635]. Other examples for Crohn's disease are the interactions between SNPs in the autophagy-related 16-like 1 *ATG16L1* and SNPs in the *NOD2* gene [282] and between SNPs in the toll-like receptor 9 gene (*TLR9*) and variants in *NOD2* and *IL23R* (interleukin 23 receptor) genes [665].

A complex example where the functional basis of a potential G-G interaction has been solved is for multiple sclerosis (MS). The HLA-DR2 haplotype in the major histocompatibility complex (MHC) contains the DR alleles DRB1*1501 (*DR2b*) and DRB5*0101 (*DR2a*) at the respective DRB1 and DRB5 loci, located 85 Kb apart. The inseparability of *DR2a* from *DR2b* that are known to be associated with MS could imply the persistence of a founder haplotype. Another possibility might be an advantageous epistatic interaction between these alleles. Gregersen et al. [262] found evidence that natural selection causes the strong LD between *DR2a* and *DR2b*. To this end, they generated genetically engineered mice that synthesize the corresponding human immune proteins and found that mice producing the protein product of *DR2b* were highly susceptible to disease, whereas those producing the *DR2a* product did not progress toward disease. Then, in a crucial test, mice expressing both alleles had an overall reduced susceptibility to disease, suggesting that *DR2a* modulates the impact of *DR2b* [535].

The descriptions from above indicate that more examples have been reported in the literature for G-E than for G-G interaction, and that G-G interaction can be quite complicated. We therefore consider well-studied examples for G-G interaction in the next section.

11.5.2 Which genes interplay in determining the coat color in the Labrador retriever?

Coat color variation in mammals is one of the best studied examples for the effect of gene-gene interactions on phenotypes. For mice alone, more than 120 different loci are known [53]. For domestic animals, the biggest variation in size and body type is seen in dogs, and they also represent a fantastic example for studying coat color genetics. The first books postulating a number of genes that could explain the inheritance of coat color and pattern in dogs were published in the 1950s, and the hypotheses were based on breeding data [594]. In this section, we consider the interaction between the brown locus and the extension locus and their effect on the coat color of the Labrador retriever [594, 656].

The first gene of interest is the melanocortin receptor (*MC1R*) gene. It causes the yellow color in Labrador retrievers. A loss-of-function mutation at 914C>T in *MC1R* causes an arginine to be replaced by a premature stop codon, and this variant is present in a wide range of dog breeds [594]. The locus is called *extension locus* or *E locus*. The arginine represents the wild-type allele and is termed E, and the premature stop codon is termed e.

The second involved gene is the tyrosinase-related protein 1 (*TYRP1*) gene. Three different new variants have been detected in *TYRP1*, a combination of any two of which will cause brown coat color. One variant contains a premature stop codon in exon 5 (Q331ter) (c.991C>T). The second variant has a deleted proline residue in exon 5 (345delP) (c.1033–6 deleted), and the third variant is a base-pair substitution in exon 2 that causes a serine to be changed to a cysteine (S41C) (c.121T>A) [595]. In the following, any variant at this locus, termed *brown locus* or *B locus* is denoted by b, and B represents a dog with wild-type variants at all positions.

The three main colors of the Labrador are black, brown (chocolate), and yellow, and they are determined by the variants at the brown and the extension locus (Figure 11.3). Specifically, the Labrador is yellow if it is homozygous for the e allele, irrespective of the genotype at the brown locus. If, however, the dog carries at least one E allele at the *E* locus, the *B* locus determines the coat color. If the Labrador is homozygous for the b allele, it is brown, and black, otherwise. Thus, the effect of the *B* locus is suppressed if the dog is homozygous for the e allele. Furthermore, brown is recessive to black.

Genotypes	E E	E e	e e
B B			
B b			
b b			

Fig. 11.3 Interaction of brown and extension loci and its effect on the coat color in the Labrador retriever. There is a limited phenotypic expression of the brown locus in homozygous ee dogs in that this locus is still responsible for pigmentation of the iris, nose, and lips. The bottom right figure is reprinted from Ref. [656] with kind permission from Oxford University Press, the late Joe W. Templeton, and William S. Fletcher. All other figures have been kindly provided by Yvonne Eckel-Fiedler, president of the Retriever Club Europa.

Table 11.16 Interaction of brown and extension loci in the Labrador retriever.

Genotypes	EE	Ee	ee
BB	Black coat	Black coat	Yellow coat
	Black nose	Black nose	Black nose
Bb	Black coat	Black coat	Yellow coat
	Black nose	Black nose	Black nose
bb	Brown coat	Brown coat	Yellow coat
	Brown nose	Brown nose	Brown nose

The effect of the brown locus is suppressed if dogs are homozygous ee at the extension locus. However, the *B* locus has a limited expression in homozygous ee dogs in that it is responsible for pigmentation of the iris, nose and lips.

The story about the color in the Labrador retriever using the *B* and *E* loci is more interesting than just the simple interaction as described so far. Figure 11.3 also shows that the *B* locus has a limited expression in homozygous ee dogs in that it is responsible for pigmentation of the iris, nose, and lips (Table 11.16; also see Ref. [656]). Specifically, the yellow coat-colored dog in the last row has the normal pigment of a chocolate Labrador in the nose and lips and has genotypes bb and ee, while a yellow coat-colored Labrador with black nose and lips has at least one *B* allele.

11.5.3 What are commonly used concepts of interaction? And what is epistasis?

Primarily in the late 1970s, early 1980s, there has been a debate in the epidemiologic literature about the meaning of the term *interaction* (see Refs. [571, 616]). This discussion has been settled, and in the modern literature four different concepts of interaction need to be distinguished, which are

- Statistical interaction
- Biological interaction
- Interaction in public health
- Interaction in individual decision making.

In the following, we will define these different interaction terms and will illustrate the differences between the terms with examples for the situation of 2×2 contingency tables. We will heavily borrow from Refs. [571, 616] and will take their definitions.

The term well known to statisticians is *statistical interaction*. A gene *G* and an environmental factor *E* are considered to be statistically independent if the risk of disease given exposure to both *G*, i.e., carriers of the risk variant, and *E* can be modeled as a function of the separate effects of being exposed to *G* and being exposed to *E*. If an additional parameter is required to adequately describe the risk of joint exposure, we say there is statistical interaction.

It is important to note that statistical interaction depends on the employed statistical model. We consider the simplest case of G-E interaction, i.e., the presence or the absence of a variant or an exposure. For example, we can model the absolute risk of disease $p = P(\text{aff})$ as a linear function of a genetic main effect G , an environmental main effect E , and a cross-product interaction term $GE = G \cdot E$

$$p = P(\text{aff}) = b_0 + b_G G + b_E E + b_{GE} GE.$$

Here, the interaction is interpreted as the difference across the two environmental strata of the risk difference caused by the genetic effect [375]. If the parameter b_{GE} is 0, we say there is no statistical interaction, or in short no interaction.

The alternative standard model is the logistic regression model. Here, the log odds of disease is modeled as a linear function of main effects plus the cross-product term:

$$\text{logit}(p) = \ln \frac{p}{1-p} = \ln \frac{P(\text{aff})}{1-P(\text{aff})} = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE. \quad (11.11)$$

The coefficient β_{GE} is the log *OR* of the interaction effect, and $\exp(\beta_{GE})$ is the ratio of the *OR* comparing exposed carriers and exposed non-carriers to the *OR* comparing unexposed carriers and unexposed non-carriers [376]. Similar to above, if $\beta_{GE} = 0$, we conclude that there is no statistical interaction. What is important is that what does or does not look like an interaction on one scale may or may not look like an interaction on the other scale [375]. This is illustrated in Figure 11.4 where different scalings are used for the plots in the left and right columns. For 11.4a and b, parameters have been chosen in a way that there is no statistical interaction in the absolute risk model. At the same time, there is interaction on the log odds scale, i.e., when modeling $\text{logit}(p) = \ln \left(\frac{p}{1-p} \right)$. The model chosen for Figures 11.4c and d reflects just the opposite: There is no interaction on the log odds scale, but there is interaction for the absolute risk model.

For *biological interaction*, Siemiatycki and Thomas [616] gave the following definition: Two factors are biologically independent if the qualitative nature of the mechanism of action of each is not affected by the presence or absence of the other. Otherwise, they are said to interact. To illustrate this definition, consider the left part of Table 11.17, where a disease is displayed that affects 1/1000 unexposed non-gene carriers. The relative risk is $RR_E = 2$ for exposed subjects, and it is also $RR_G = 2$ for gene carriers. If a subject is exposed and is a gene carrier, the relative risk is $RR_{GE} = 3$ compared to unexposed non-gene carriers. This example represents a situation of biological independence because the relative risk increases in both columns with the same value on the additive scale. It also increases with the same value on the additive scale in both rows. On the additive scale, this model is simply $RR_{GE} = RR_G + RR_E - 1$. Of course, E and G do not act synergistically in this example.

In the right part of the table, an example is given for biological independence on the multiplicative scale. Specifically, the relative risk is 5 row-wise, and it is 2 column-wise. The relative risk is 10 in exposed carriers, and it is obtained as $RR_{GE} = RR_G \cdot RR_E$. Thus, there is biological independence on the multiplicative

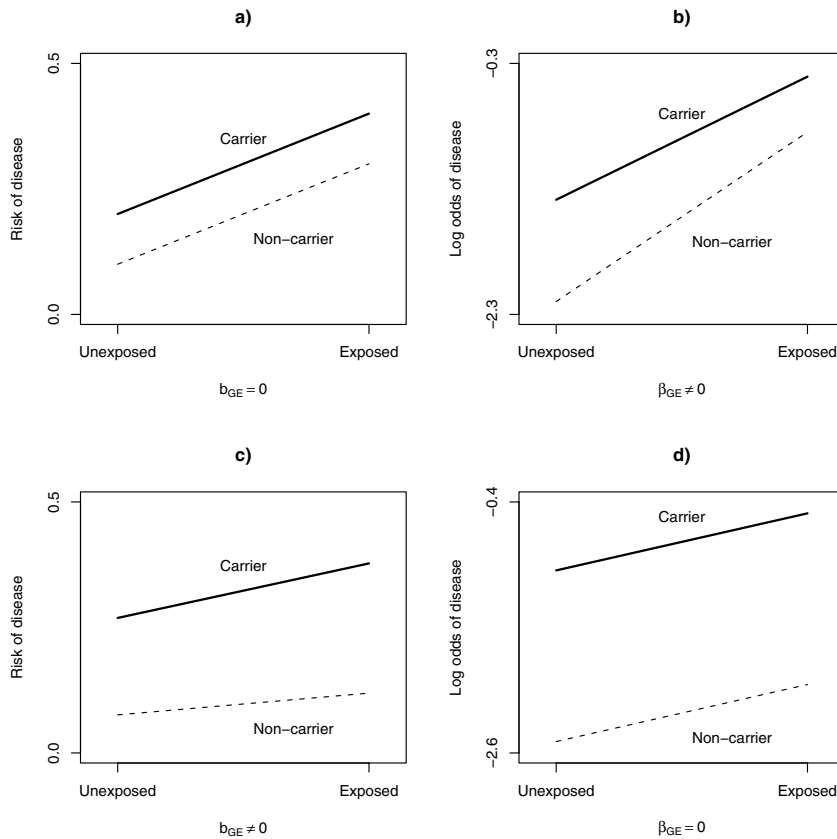


Fig. 11.4 a) No statistical interaction in the absolute risk model. However, there is interaction in the log odds of disease model (b). Similarly, there is no interaction in the log odds of disease model d), but there is interaction in the absolute risk model c). The idea for this figure has been taken from Ref. [375].

scale. However, in contrast to the previous example, there is a synergistic effect of E and G in this example.

Interaction in public health is related to the number of cases of disease occurring in a population and the proportional contribution of each risk factor to this case burden. As given in Ref. [571], the effects of G and E may be considered independent for public health purposes if the number of cases of disease that would occur in the population does not depend on the extent to which G and E occur together in the same individuals. If the number of cases does depend on the extent to which risk factors G and E occur together in the same individuals, then the effects of G and E interact. Thus, for public health purposes interaction between G and E is equivalent to a departure from additivity of attributable risks (AR).

Table 11.17 No biological interaction.

Independence on additive scale			Independence on multiplicative scale		
Relative risk	Not G	G	Relative risk	Not G	G
Not E	Reference = 1	2	Not E	Reference = 1	5
E	2	3	E	2	10

Displayed are relative risks with unexposed non-gene carriers as reference group. Left half of the table: Disease with a prevalence of 1/1000 in unexposed (not E) non-gene carriers (not G). The risk is 2/1000 in exposed subjects or gene carriers. Exposed gene carriers have an absolute risk of 3/1000. The relative risk increases row-wise by 1, i.e., with the same value. It also increases column-wise with the same value, thus represents biological independence on the additive scale. Right half of the table: Disease with a prevalence of 1/1000 in unexposed (not E) non-gene carriers (not G). The risk is 2/1000 in exposed subjects and 5/1000 in gene carriers. Exposed gene carriers have an absolute risk of 10/1000. The relative risk increases row-wise by a factor of 5, i.e., with the same value. It increases column-wise by a factor of 2, thus represents biological independence on the multiplicative scale.

Interaction in individual decision making is similar to that of interaction for public health issues but now based on differences in absolute risks. These absolute risks can be considered on the individual level, while attributable risks are related to the population under consideration. Rothman et al. [571] considered the following example of hypertension and the use of oral contraceptives (Table 11.18). The RR for cerebrovascular complication from oral contraceptives is about the same among normotensive and hypertensive women. A physician advising a woman on whether to use oral contraceptives would reasonably inquire about the presence of hypertension despite this fact. Specifically, the increase in absolute risk is considerably greater for hypertensive women, thus making hypertension interactive with oral contraceptive use for the purpose of evaluating individual risk. On this basis, hypertension becomes a contraindication for oral contraceptive use.

This interaction does not require any biological model. Instead, departures from additivity of RD is the focus whatever the underlying biological mechanisms might be, provided that cost to the individual is taken to be a linear function of risk, as it usually is, especially when risks are small [571].

Before going into detail describing gene-environment and gene-gene interactions, we would like to point out two tools helping in the planning stage of experiments, namely,

Table 11.18 Interaction in individual decision making. Numbers are given per 1000 subjects.

	Normotensive	Hypertensive
No contraceptive	1	2
Contraceptive	2	4

POWER and QUANTO, which are both freely downloadable from the web.

11.5.4 How can gene-environment interaction be statistically tested?

In Section 11.5.3, we have shown that testing for statistical interaction always means that interaction is tested in a specific statistical model. The standard approach for testing statistical interaction in diseases tests for departure from additivity on the log odds ratio scale, i.e., the logistic regression model. An excellent introduction to the logistic regression model has been given, e.g., by Kleinbaum [358]. Although the logistic regression model is used for both testing G-E interaction and testing G-G interaction, we need to distinguish them in the following. Here, we consider G-E interaction, followed by G-G interaction in section 11.5.5.

In the following, we closely follow the presentation of Kraft et al. [376] who considered the simplest case of G-E interaction, i.e., the presence or absence of a variant or an exposure. For the logistic regression model, one may choose any of the standard statistical tests, i.e., the likelihood ratio, the score or the Wald test. Of greater interest is, however, the statistical hypothesis to be tested and the statistical model to be used for this test.

The first hypothesis to be considered is the case-control test for G-E interaction. The statistical model is that of Eq. (11.11)

$$\text{logit}(p) = \ln \frac{p}{1-p} = \ln \frac{P(\text{aff})}{1-P(\text{aff})} = \beta_0^{(1)} + \beta_G^{(1)}G + \beta_E^{(1)}E + \beta_{GE}^{(1)}GE. \quad (11.12)$$

The *GE test* hypothesis for the test of no G-E interaction is

$$H_0: \beta_{GE}^{(1)} = 0 \quad \text{versus} \quad H_1: \beta_{GE}^{(1)} \neq 0.$$

As the second test, the *G-GE test* is also based on model (11.11) but is a joint test that combines information about the marginal gene effect and the G-E interaction effect. The hypothesis thus introduces two restrictions

$$H_0: \beta_G^{(1)} = 0, \beta_{GE}^{(1)} = 0 \quad \text{versus} \quad H_1: \beta_G^{(1)} = 0, \beta_{GE}^{(1)} \neq 0,$$

and subsequently has 2 d.f. Because the same model (11.12) is used as for the first test, parameter estimates are identical.

Alternatively, one may consider the case-control test for the marginal gene effect. This is the test of an overall association between the gene and the disease, and it does not use any environmental data. The statistical model to be considered is

$$\text{logit}(p) = \ln \frac{p}{1-p} = \ln \frac{P(\text{aff})}{1-P(\text{aff})} = \beta_0^{(2)} + \beta_G^{(2)}G,$$

and the third *G Test* hypothesis is

$$H_0: \beta_G^{(2)} = 0 \quad \text{vs.} \quad H_1: \beta_G^{(2)} \neq 0.$$

Here, a different statistical model is used from that in Eq. (11.12) and therefore parameter estimates are generally different. This is indicated by the superscript (2) for this statistical model.

Finally, the *case-only design*, where controls are ignored and not included in the analysis, may be utilized. Here, the data reduce to a simple 2×2 table for the exposure versus the gene. In principle, one could use any standard 2×2 association test in this situation, like the standard χ^2 test from Section 11.2.1. Similarly, association measures as discussed in Section 11.2.3 could be calculated in this situation as well. It is, however, important to note that the RD contrasts the presence or absence of exposure given the case is carrier of the variant in the case-only design.

The statistical model is

$$\text{logit}(\mathbb{E}(E)) = \ln \frac{P(E=1)}{P(E=0)} = \beta_0^{(3)} + \beta_{GE}^{(3)}G,$$

and the fourth GE_{ca} test hypothesis is

$$H_0: \beta_{GE}^{(3)} = 0 \quad \text{versus} \quad H_1: \beta_{GE}^{(3)} \neq 0.$$

In the case-only design, not only the statistical model but also the dependent variable is different. Therefore parameter estimates may be quite different, and the superscript (3) characterizes parameters for this statistical model.

In summary, there are four different approaches for statistical testing:

1. GE case-control for the interaction effect G-E,
2. G case-control for the marginal genetic effect G,
3. $G-GE$ case-control test combining G and G-E,
4. GE_{ca} case-only test for G-E.

Kraft et al. [376] extensively studied the power of the four different approaches using the likelihood ratio test. Table 11.19 summarizes the most important results. The first row displays the model without any effect. Subsequently, all type I error levels match the significance level of 0.01. In the second row, results are given for a model without a marginal mean and without a marginal environmental effect. The test with the greatest power for detecting the G-E interaction clearly is the case-only test, followed by the joint case-control test $G-GE$. In the third row, results are displayed for the case of a major gene effect only. Here, the test for the marginal genetic effect has the greatest statistical power, and the combined $G-GE$ test is adequate as well. GE and GE_{ca} fail because there is no G-E interaction. When there is a major gene effect and G-E interaction (row 4), the combined test has greatest power. The test for the marginal gene effect also has > 99% power in the table.

Results are similar when there is a marginal environmental effect (Table 11.19, rows 5–8). Again, the combined $G-GE$ test has reasonable statistical power in all considered situations, while the other tests fail in at least one situation. Kraft et al. [376] have investigated other parameter constellations as well, and they show that the efficiency of the combined $G-GE$ test relative to the others increases with decreasing significance levels.

When exposure is subject to misclassification [421], all tests can have inflated type I error levels when E and G are correlated. For the GE and $G-GE$ tests this inflation is noticeable only when the G-E dependence is unusually strong. The inflation can be large for the GE_{ca} test even in case of modest correlation.

In summary, the best test for detecting a G-E interaction is the GE_{ca} from the case-only design. The combined $G-GE$ test is generally more powerful than the marginal genetic association test when no marginal gene but a G-E effect is present. It also has greater power when there is no genetic main effect and the measurement error is small to moderate. In addition, it generally outperforms the classical case-control GE test, even when there is exposure misclassification. If the $G-GE$ test is not the most powerful, the loss in power relative to the most powerful test is acceptable. The $G-GE$ test thus is the most appropriate test when the aim is the detection of either a major gene effect or a G-E interaction, even in the presence of exposure misclassification.

Table 11.19 Power for four association tests across a range of environmental and genetic main effects and gene-environment interaction effects.

OR_E	OR_G	OR_{GE}	G	GE	$G-GE$	GE_{ca}
1.0	1.0	1.0	0.01	0.01	0.01	0.01
		2.0	0.29	0.61	0.71	0.96
	1.5	1.0	0.88	0.01	0.81	0.01
		2.0	> 0.99	0.65	> 0.99	0.98
2.0	1.0	1.0	0.01	0.01	0.01	0.01
		2.0	0.7	0.65	0.80	0.98
	1.5	1.0	0.88	0.01	0.80	0.01
		2.0	> 0.99	0.68	> 0.99	0.99

Based on a study with 1000 cases and 1000 controls, significance level of 0.01 with two-sided test. Population prevalence of exposure is 0.25 and disease prevalence is 0.0001. The genetic model is dominant with allele frequency determined from disease prevalence and the genetic effect; G is the marginal genetic test; GE is the standard case-control test for gene-environment interaction; $G-GE$ is the 2 d.f. combined test for a gene and a gene-environment interaction; GE_{ca} is the case-only test for gene-environment interaction.

So far, we have only considered the simple case of a dichotomous exposure and a dichotomous gene effect. If the exposure is dichotomous but an SNP with three genotypes has to be considered instead of the presence or absence of a variant, the simplest approach is the case-only analysis using the exposure as dependent variable together with the trend test for genotypes. However, one cannot generally expect an additive gene acting mechanism, and clear rules of thumb have not been provided. Standard alternatives without assuming a specific genetic model are the two d.f. standard χ^2 test and the case-only logistic regression model

$$\text{logit}(\mathbb{E}(E)) = \ln \frac{P(E=1)}{P(E=0)} = \beta_0 + \beta_{GE,1}1_{G=1} + \beta_{GE,2}1_{G=2},$$

where 1_x is the indicator function, the SNP coded with values 0, 1, and 2, and $\beta_{GE,1}$ and $\beta_{GE,2}$ are the regression coefficients.

A third general approach is the case-control logistic regression model

$$\begin{aligned}\text{logit}(p) = & \beta_0 + \beta_{G,1}1_{G=1} + \beta_{G,2}1_{G=2} \\ & + \beta_{GE,1}1_{G=1} \cdot 1_{E=1} + \beta_{GE,2}1_{G=2} \cdot 1_{E=1}.\end{aligned}$$

Here, the hypothesis to be tested is $H_0: \beta_{GE,1} = \beta_{GE,2} = 0$ leading to a two d.f. test.

The final level of complexity can be reached by moving from a dichotomous exposure variable to an exposure with several unordered or ordered categories or continuous measurements. In any of these cases, a logistic regression model is typically employed and careful modeling required. This, however, exceeds the level of this introductory textbook. We therefore close this section with a simple example on G-E interaction between a variant in the multidrug resistance protein 1 (*MDR1*) gene, pesticide exposure, and Parkinson's disease.

Example 11.6. *In this example, we illustrate the analysis of G-E interaction using data from the two published studies on the c.3435C>T (rs1045642) polymorphism of the MDR1 gene (Table 11.20). Based on the hypothesis that MDR1 variants act as a susceptibility factor for Parkinson's disease in conjunction with several other factors contributing to the development of the disease to different extents, we performed an interaction analysis of pesticide exposure (neurotoxicants: pesticides, herbicides, fungicides, and insecticides) and rs1045642.*

Table 11.20 Genotype frequencies of rs1045642.

	Drożdżik et al. [171]			Zschiedrich et al. [772]		
	C/C	C/T	T/T	C/C	C/T	T/T
Case	26	66	25	69	126	70
Control	24	58	21	21	73	29

	Drożdżik et al. [171]			Zschiedrich et al. [772]		
	C/C	C/T or T/T	Total	C/C	C/T or T/T	Total
Exposed	7	52	59	2	17	19
Non-exposed	19	29	48	24	43	47

The first half of the table reports frequencies in Parkinson patients and controls from the two published studies by Drożdżik et al. [171] and Zschiedrich et al. [772]. The second half of the table displays frequencies in Parkinson patients by pesticide exposure.

Analyses using the Cochran–Armitage trend test revealed p-values of 0.82 and 0.42 in the Drożdżik [171] and the Zschiedrich [772] data, respectively. Even when carriage of the T-allele, i.e., a dominant model, is investigated, p-values are 0.85 ($OR = 0.94$, 95% $CI(OR) = [0.50; 1.77]$) and 0.05 ($OR = 1.71$, 95% $CI(OR) = [0.99; 2.95]$). In summary, there is no or only a weak marginal effect for the c.3435C>T SNP in both data sets.

When we consider the case-only G-E analysis, a different picture is obtained. In Parkinson's disease patients, *c.3435C>T* differs significantly between those exposed to pesticides and those non-exposed to pesticides. This association is seen in both samples with $p = 0.0009$ ($OR = 4.87$, 95% $CI(OR) = [1.83; 12.95]$) and $p = 0.047$ ($OR = 4.74$, 95% $CI(OR) = [1.01; 22.31]$).

11.5.5 How can gene-gene interaction be statistically tested?

Excellent descriptions of the statistical models for testing G-G interactions have been given by Cordell [143, 146], and these form the basis of this section. Given case-control data genotyped at two SNPs from two different loci, the saturated statistical model is given by

$$\begin{aligned} \text{logit}(p) = & \mu + \alpha_1 x^{(A)} + \alpha_2 x^{(AA)} + \beta_1 x^{(B)} + \beta_2 x^{(BB)} \\ & + \iota_{11} x^{(AB)} + \iota_{12} x^{(ABB)} + \iota_{21} x^{(AAB)} + \iota_{22} x^{(AABB)}. \end{aligned} \quad (11.13)$$

This model includes nine different parameters (Table 11.21): a parameter μ that represents the “baseline” log odds for an individual with genotypes a/a and b/b . Parameters α_1 and α_2 represent the effects of replacing one or both alleles at locus A with the modifying allele A. Similarly, β_1 and β_2 represent the effects of replacing one or both alleles at locus B with the modifying allele B. Finally, there are four interaction parameters ι_{11} , ι_{12} , ι_{21} , and ι_{22} . The numbers of the indices represent the number of A and B alleles at the two loci. Finally, the superscripts for the x -variables indicate the number of A and B alleles, respectively. The saturated model of Eq. (11.13) has nine two-locus genotypes that are modeled by nine different parameters. Although the saturated model is the best fitting one, a model with fewer parameters might be preferable, e.g., because of greater stability.

Table 11.21 Saturated gene-gene interaction model. The table shows the 9 parameters of the saturated two-locus interaction model.

Genotypes	bb	bB	BB
aa	μ	$\mu + \beta_1$	$\mu + \beta_2$
aA	$\mu + \alpha_1$	$\mu + \alpha_1 + \beta_1 + \iota_{11}$	$\mu + \alpha_1 + \beta_2 + \iota_{12}$
AA	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_1 + \iota_{21}$	$\mu + \alpha_2 + \beta_2 + \iota_{22}$

To test for interaction, the 4 d.f. test $\iota_{11} = \iota_{12} = \iota_{21} = \iota_{22}$ is used. Because the logistic regression model is used for estimation, the test may be carried out using the likelihood ratio, the score, or the Wald test statistic.

There is a series of genetic models that are simpler than the saturated interaction model. For example, the recessive model for two loci with one interaction term yields the log odds disease of Table 11.22. If instead a dominant model is considered for both loci, the expected log odds are as provided in Table 11.23. In both models, there is only a single interaction parameter ι that may be tested using a 1 d.f. test.

Table 11.22 Recessive gene-gene interaction model. The table shows the four parameters for the recessive–recessive two-locus interaction model.

Genotypes	<i>bb</i>	<i>bB</i>	<i>BB</i>
<i>aa</i>	μ	μ	$\mu + \beta$
<i>aA</i>	μ	μ	$\mu + \beta$
<i>AA</i>	$\mu + \alpha$	$\mu + \alpha$	$\mu + \alpha + \beta + \iota$

Table 11.23 Dominant gene-gene interaction model. The table shows the four parameters for the dominant–dominant two-locus interaction model.

Genotypes	<i>bb</i>	<i>bB</i>	<i>BB</i>
<i>aa</i>	μ	$\mu + \beta$	$\mu + \beta$
<i>aA</i>	$\mu + \alpha$	$\mu + \alpha + \beta + \iota$	$\mu + \alpha + \beta + \iota$
<i>AA</i>	$\mu + \alpha$	$\mu + \alpha + \beta + \iota$	$\mu + \alpha + \beta + \iota$

Both models and several more can be described by the standard two-locus model formulation [689]

$$\text{logit}(p) = \mu + \alpha x_A + \beta x_B + \iota \cdot x_A \cdot x_B, \quad (11.14)$$

where x_A and x_B are appropriate coding schemes for SNPs A and B. Specifically, the recessive–recessive model is obtained by letting $x_A = 1$ if the genotype G_A of a subject at SNP A is A/A , and 0, otherwise. Similarly, $x_B = 1$ if $G_B = B/B$. The dominant–dominant model is obtained by letting $x_A = 1$ if $G_A = A/A$ or a/A , and similarly for x_B . Other G-G models include the additive–additive interaction model, where x_A and x_B take values 0, 1, and 2 depending on the number of A and B alleles, respectively. Combinations of specific coding schemes are possible, resulting, e.g., in an additive–dominant interaction model. All these models include a single interaction parameter ι .

Analogously to studying G-E interactions, an alternative and more powerful approach in case-control studies is to use the case-only approach for identifying G-G interactions [146]. In a case-only analysis, interaction is investigated by testing the null hypothesis that there is no correlation between the two loci in the cases. The test is the simple χ^2 test of independence between genotypes and has 4 d.f. As pointed out by Cordell [146], the main problem with the case-only approach is that the genotype variables have to be uncorrelated in the general population. It is this assumption, rather than the design per se, that provides the increased power compared to case-control analysis. As a consequence, the case-only design is therefore unsuitable for loci that are either in closely chromosomal vicinity or show correlation for another reason. Cordell [146] importantly pointed out that the assumption of independence between SNPs not in linkage disequilibrium is reasonable. This is in contrast to epidemiological studies of environmental factors where confounding

between variables is common. One could use a pre-test approach to first test for correlation between the SNPs in the general population and then use the result of this test to determine whether to perform either the case-only or the case-control interaction test. However, this pre-test approach does not keep its nominal significance level. A preferable approach is to combine the case-only and case-control estimators into a single test. To this end, Zhao et al. [744] proposed considering the difference in inter-locus allelic association between cases and controls.

Irrespective of the statistical model used, detecting G-G interactions is challenging. This is best illustrated using a simple example for a G-G interaction with no main effects (Table 11.24). To illustrate the complexity of G-G interaction, we consider a simple two SNP G-G interaction example for sporadic breast cancer.

Table 11.24 Gene-gene interaction model without main effect.

Genotypes	<i>bb</i> (0.25)	<i>bB</i> (0.5)	<i>BB</i> (0.25)	Marginal effect
<i>aa</i> (0.25)	0	0	1.0	0.25
<i>aA</i> (0.5)	0	0.5	0	0.25
<i>AA</i> (0.25)	1.0	0	0	0.25
Marginal effect	0.25	0.25	0.25	0.25

In this hypothetical example (taken from Ref. [223]), both SNPs interact but do not show main effects. The values in the table reflect the penetrances of each two-locus genotype. The genotype frequencies of each genotype are given in parentheses next to the genotype. It is important to note that the marginal genotype effects for each SNP are equal. If one studied one of these SNPs independent of the other or jointly without taking into account the interaction, no effect would be detectable. This genetic model can be interpreted as follows: The disease arises only when exactly two variant alleles are present, with the double heterozygous genotype not having an effect as pronounced as the two homozygous genotypes because the variant alleles are different.

Example 11.7. For illustration, we consider the data of Ref. [248]. The authors reported an interaction between the $-26G>A$ SNP (rs1799943) of the 5' UTR of the *BRCA2* gene and the Arg72Pro polymorphism (rs1042522) of the *p53* gene in sporadic breast cancer. Absolute frequencies and percentages from this study are displayed in Table 11.25. The biggest frequency difference between patients and controls is seen for the combination Pro Pro, GA, which is observed in only 3% of the patients but 14% of the controls. At the same time, 19% of the patients are homozygous Arg Arg and GG, while only 10% of the controls are homozygous for this genotype combination. The saturated model yielded a *p*-value of 0.0929 for the 4 d.f. interaction test, and this model had the lowest Akaike information criterion (AIC) among all standard interaction models considered in Table 11.26.

In brief, the AIC is a measure of the goodness-of-fit of an estimated logistic regression model. It is often used for model selection and given by $AIC = -2 \log L + 2c$, where L denotes the likelihood function and c the number of parameters of the model. Specifically, given a data set, several competing models are ranked according

Table 11.25 Example for gene-gene interaction.

	Patients			Controls		
	GG	GA	AA	GG	GA	AA
Pro Pro	38 (15.63%)	7 (2.88%)	3 (1.23%)	45 (13.51%)	46 (13.81%)	6 (1.80%)
Pro Arg	59 (24.27%)	38 (15.64%)	12 (4.94%)	78 (23.42%)	68 (20.42%)	14 (4.20%)
Arg Arg	47 (19.34%)	29 (11.93%)	10 (4.11%)	34 (10.21%)	38 (11.41%)	4 (1.20%)

Frequencies (percentages in parentheses) of rs1042522, i.e., the Arg72Pro polymorphism of the *p53* gene codon 72, and rs1799943, i.e., the polymorphism 5'UTR-26G>A of the *BRCA2* gene, in sporadic breast cancer patients and controls from Ref. [248].

to their AIC, with the one having the lowest AIC being the best. For details, see, e.g., Ref. [358].

The standard model with a similar AIC as the saturated model was the Add-Dom model with an additive coding for rs1042522 and a dominant coding for rs1799943. Similar to the saturated model, the interaction effect was marginally significant ($p = 0.0561$). A substantially lower AIC can be obtained by selecting the best model among all models. However, this model is selected from many different interaction models, where first the multiple testing issue is not addressed. Second, the stability of this model has to be scrutinized. Finally, several different models will have similar AIC, and the models probably cannot be distinguished. It is therefore questionable whether the very specific selected model shows adequate goodness-of-fit or even the lowest AIC when applied to a replication, i.e., a new data set.

An alternative to the case-control approach is the case-only analysis, where the standard χ^2 test of independence is used. This test reveals a χ^2 test statistic of 9.936, yielding $p = 0.0415$ at 4 d.f. In contrast to the case-control analysis, the case-only design relies on the assumption that the SNPs are independent in the general population. This assumption is not reasonable for the *BRCA2* and *p53* genes because they form in vivo complexes [445], thus interact in the general population.

Table 11.26 Akaike information criterion of standard gene-gene (G-G) interaction models for the breast cancer data.

Model	AIC	Model	AIC
Saturated	765.425	Rec-Dom	772.264
Recessive	778.956	Dom-Rec	784.025
Dominant	768.805	Add-Dom	766.243
Additive	771.942	Dom-Add	775.401

11.5.6 What is multifactor dimensionality reduction?

In the example at the end of the previous section, we have seen that it may be difficult to derive a reasonable and reliable statistical interaction model even when only two SNPs are considered. Although the possible number of models is low when merely two SNPs are used, the number of possible models increases exponentially with the number of SNPs. Because of computational complexity, not all possible combinations can be considered for high-dimensional data. This does apply not only to data from genome-wide association (GWA) studies but also to smaller scale studies involving just a few hundred SNPs. The high dimensionality has a second side effect: The more variables are involved, the sparser the contingency tables get, and many cells are empty. Many standard statistical algorithms are unable to deal with sparse data and fail. Finally, non-linearities in the data are typically modeled through interaction terms. In Example 11.7, not all standard simple interaction models have been able to pick up the complex structure even in the simple breast cancer data set. Therefore, non-linearities are hard to detect—this is an issue of statistical power—when standard statistical approaches are used. Therefore, several authors have advocated the use of automated searches on the data using *machine learning*, which is the automated computer process of obtaining the best fitting model to a data set. Among the many different machine learning approaches that have been proposed for detecting G-G interactions, including random forests [84] or logic regression [598], the multifactor dimensionality reduction (MDR) have received great interest. Thorough reviews on detecting G-G interactions by machine learning approaches can be found in Refs. [146, 452, 763].

MDR has been reviewed nicely in Refs. [146, 472, 484], and here, we closely follow the presentation by Cordell [146]. The original algorithm of the *multifactor dimensionality reduction* method is described in Algorithm 11.2. An excellent explanation of a more recent algorithm has been given in Ref. [484]. The original MDR approach has been extended in several ways, see Refs. [101, 483, 492] for recent extensions and Refs. [453, 473, 569] for variable selection approaches using MDR. The current version of MDR is able to analyze GWA data. However, an exhaustive search for all combination is impossible on the GWA level, not only with MDR but also with all other approaches. Therefore, MDR is limited to the analysis of a few hundred SNPs for the exhaustive search of all combinations. For analyzing GWA data, only lower-order interactions are considered.

Algorithm 11.2.

1. A 10-fold *cross-validation* is used. This means that the observed data are divided into 10 equal parts, and a model is fit to each 9/10 of the data. These data are termed *training data*. The remaining 1/10, termed *test data*, is used to assess model fit.
2. Within each 9/10 of the data, a set of p SNPs is selected and their possible multifactor classes or cells are considered. For two SNPs, there are nine possible genotype classes or cells.
3. The ratio of the number of cases to the number of controls is estimated in each cell, and the cell is labeled as either high risk if the case-control ratio reaches or exceeds a predetermined threshold, e.g., ≥ 1 , and low risk if it does not reach this threshold. This reduces the original

p -dimensional model to a one-dimensional model. Specifically, there is one variable with two classes: high risk and low risk.

4. The procedure is repeated for each possible p -factor combination. The combination that fits the current 9/10 of the data best, i.e., gives minimum classification error among all p -locus models is selected.
5. The percentage of correct classifications, termed *hit fraction* or *testing accuracy*, of this best p -locus model is estimated using the test data.
6. The whole procedure is repeated for each of the 9/10--1/10 partitions of the data, and the final best p -locus model is the model that maximizes the percentage of correct classification. Specifically, the *cross-validation consistency* is the number of cross-validation partitions in which that same p -locus model was chosen as the best model, i.e., the number of replicates in which it maximized the percentage of correct classification. The average percentage of correct classifications is defined as the average of the fraction of correct classifications over the 10 cross-validation test data sets.

MDR does not test for interactions in a statistical sense. Instead, MDR tries to identify combinations of SNPs that allow distinguishing “high-risk” and “low-risk” groups. In other words, the number of dimensions is reduced in MDR by converting a high-dimensional contingency table to a simple one-dimensional model containing the information “high-risk combination” and “low-risk combination”. For example, applying MDR to the breast cancer data of Example 11.7 results in the classification scheme displayed in Table 11.27. This classification scheme is the natural one when comparing the genotype frequencies of both SNPs in cases and controls (Table 11.25).

This example illustrates the simplicity and elegance of MDR that is one cause for its popularity. An important aspect of the MDR approach is that cross-validation is intrinsic to the procedure. The cross-validation step allows investigators to judge the stability of the model and whether it tends to overfit. *Overfitting* refers to the fact that complex models tend to provide good to excellent fit of the training data used for developing the model. However, when applied to a different test data set, the fit can be substantially worse because the fitted model incorporates some of the random fluctuations of the training data; for a detailed discussion, see Refs. [19, 370].

Table 11.27 Classification scheme from multifactor dimensionality reduction (MDR) for the breast cancer example.

	GG	GA	AA
Pro Pro	High	Low	Low
Pro Arg	High	Low	High
Arg Arg	High	High	High

11.6 PROBLEMS

Problem 11.1 How do two binomial random variables talk to each other?

Problem 11.2 Show that the test statistics χ_c^2 of Eq. (11.1) and χ_A^2 of Eq. (11.3) are equivalent.

Problem 11.3 Derive the Cochran–Armitage trend test statistic.

Problem 11.4 Return to the data of Problem 10.8. Now we are interested in possible differences in the genotypes between asthma patients and healthy blood donors. Assume that there are previous analyses indicating a recessive effect of allele A at SNP 1. Based on this model, test for genetic association. Also, calculate the OR with a 95% confidence interval for being affected for carriers of the specific risk genotype.

Problem 11.5 Derive the X chromosomal trend test statistic.

Problem 11.6 Plan a genetic association study in a case-control design for which you need to calculate the required sample size. Although you are fairly sure that you are investigating SNPs in a gene that is a plausible candidate, you are rather unsure of which values to use for the parameters in the sample size calculation. Hence, you vary some of the parameter values to see how this affects the sizes you need.

Problem 11.6.1. First, you presume the lucky case of genotyping the disease-causing variant itself. You fix a two-sided $\alpha = 0.01$ and set a power $1 - \beta$ of 80%. Based on an OR of 2 associated with the disease allele, how would differing frequencies of this allele in the controls influence the required sample size? Estimate the sample size for allele frequencies in the controls of $p_{A,u} = 0.4, 0.2$, and 0.05 .

Problem 11.6.2. Next, you allow the case of genotyping SNPs that are in LD with the disease-causing variant only. Here, analyze how the OR at the SNP you genotyped changes depending on different values of LD between the SNP and the disease locus. Assuming an OR of the disease variant of 2.0, and the same frequency of 0.2 for the disease and the marker allele, what will be the OR of the marker allele if the $\mathcal{D}$ is maximal, about two-thirds of the maximum, and about half the maximum? On the other hand, assuming you have a maximal $\mathcal{D}$ between marker and disease alleles, how do differing allele frequencies at the marker affect the OR of the marker allele? Consider allele frequencies of the marker allele of 0.2, 0.3, and 0.4.

Problem 11.7 Suppose that a candidate gene case-control study has been undertaken. In addition to the SNP of interest (Table 11.28), 20 null markers have been genotyped. The values of the respective χ^2 statistics are summarized in Table 11.29.

Problem 11.7.1. Check the data group-wise for deviations from HWE.

Table 11.28 Genotype frequencies at a notional single nucleotide polymorphism (SNP) for cases and healthy controls.

	CC	CG	GG	Total
Cases	25	20	15	60
Controls	5	20	45	70
Total	30	40	60	130

Table 11.29 χ^2 statistics at 20 notional single nucleotide polymorphisms (SNPs) for cases and healthy controls.

0.1101	0.1485	1.3233	0.0191	0.1935
0.0158	1.0742	0.0642	0.0039	0.0101
0.7083	0.0268	2.4173	2.2925	6.4653
0.2750	10.8274	4.7093	0.3573	0.5707

Problem 11.7.2. Can population stratification be detected at the null markers?

Problem 11.7.3. Estimate the GC inflation factor assuming an additive model.

Problem 11.7.4. Is the candidate SNP associated with the disease? Calculate the different test statistics available.

Problem 11.7.5. Criticize the approach used in this problem.

Problem 11.8 Consider the data of the replication study given in Table 11.7. Calculate genetic impact measures.

URLs

ADMIXMAP

<http://sourceforge.net/projects/admixmap/>

EECI

<http://www.imbs-luebeck.de/software/>

EIGENSOFT

<http://genepath.med.harvard.edu/~reich/Software.htm>

GMDR

Extension of the MDR program

<http://www.healthsystem.virginia.edu/internet/addiction-genomics/software/>

gmdr_f.cfm

GENOMIC CONTROL

<http://wpicr.wpic.pitt.edu/wpiccompngen/software.htm>

L-POP

<http://statgen.iop.kcl.ac.uk/lpop/>

MDR

Original MDR program

<http://sourceforge.net/projects/mdr>

METAGEN

<http://bioinformatics.biol.uoa.gr/~pbagos/metagen/>

NEWCOMBE'S RESOURCES

<http://www.cardiff.ac.uk/medic/aboutus/departments/primarycareandpublichealth/ourresearch/resources/index.html>

POWER

Estimation of sample size and power for two-locus interactions in both cohort and case-control studies

<http://dceg.cancer.gov/bb/tools/power>

QUANTO

Estimation of sample size and power in matched case-control, case-sibling, case-parent, and case-only designs

<http://hydra.usc.edu/gxe/>

SHELLFISH

<http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php>

STRUCTURE and STRAT

<http://pritch.bsd.uchicago.edu/software.html>

12

Association Analysis in Families

To analyze association of an allele at a genetic locus with a phenotype, different study designs are possible. One classical epidemiological pathway is to conduct a case-control study as described in Chapter 11. However, for a long time, association results from these designs were eyed with suspicion, primarily for the possibility of population stratification (see Section 11.4). As an alternative, analysis of association within families was heralded.

The focus in this chapter is on samples of trios, meaning nuclear families consisting of one affected child and his or her parents, regardless of their affection status. Sections 12.1–12.5 will focus on tests utilizing exactly these family structures. After describing the haplotype relative risk (HRR) method (Section 12.1) and deriving the classical transmission disequilibrium test (TDT) (Section 12.2), risk estimates (Section 12.3) and sample size calculations (Section 12.4) for this design as well as alternative test statistics (Section 12.5) will be delineated. All approaches described in the first sections consider diallelic markers only. Extensions to the use of short tandem repeat (STR) markers with multiple alleles are described in Section 12.6.

For many studies, it might not be practical or even possible to collect trios consisting of an offspring with its both parents. Hence, alternative test statistics have been proposed that use different family structures. These will be the focus of Section 12.7. Whereas all of these procedures are applicable to the situation where a binary phenotype has been assessed, Section 12.8 is devoted to the association analysis of quantitative traits in families. For all analyses considered in this chapter, the genotyping of one single genetic marker at a time is assumed. Association analysis of multiple markers and haplotypes will be the topic of Chapter 13.

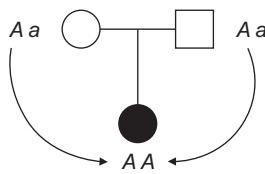


Fig. 12.1 A trio consisting of an affected child and her parents. Because both parents are heterozygous, it can be traced which alleles were transmitted to the offspring.

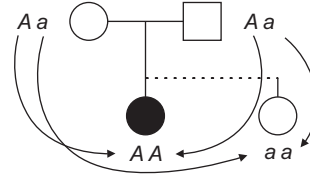


Fig. 12.2 Construction of an internal control. The non-transmitted parental alleles are utilized to comprise a pseudo-control genotype.

The strict distinction between the study of unrelated and related individuals, respectively, may blur in situations where both have been recruited. Indeed, approaches have been proposed recently that deal specifically with this case. We refrain, however, from a detailed description of these methods for two reasons. First, we have already seen in the last chapter that case-control association studies may suffer from population stratification. In contrast, family-based association studies have been advocated because they are not prone to this bias due to their matching for the genetic background within families. Second, it will be discussed in this chapter that parameter estimates from family-based association studies are slightly biased towards the null. In contrast, studies with unrelated individuals can yield unbiased parameter estimates under appropriate sampling. In summary, combining unrelated with related individuals gives up the properties of unbiased estimation and of protection against population stratification. We therefore refrain from a description of these approaches, and the interested reader is referred to a recent overview [245] and the literature for details; see, e.g., Refs. [173, 190, 330, 491, 533, 759].

12.1 WHAT IS THE HAPLOTYPE RELATIVE RISK METHOD?

As a starting point of how association might be analyzed within families, consider a family as in Figure 12.1. Because both parents are heterozygous, it can be traced exactly which alleles were transmitted to the child and which were not. An intuitive idea is to use the alleles that were not transmitted to construct a *pseudo-control* or an *internal control* (Figure 12.2). Please note that in contrast to linkage designs, it does not matter which allele came from the mother or the father. More generally, non-transmitted alleles serve as controls that have the same population genetic background as the affected offspring. Of course, this is valid only if the two non-transmitted parental alleles are a representative random sample of alleles from the population. This requires no inbreeding among parents, no correlation of parental phenotypes, single ascertainment of families, no differential fertility of the disease phenotypes, and no recombination between the disease and the marker locus.

Based on these newly created pseudo-controls, the same 3×2 or 2×2 tables as before can be constructed, and the same estimators and statistical tests are possible

as that shown in Chapter 11. As an example, a classification based on alleles A and a at a diallelic marker is shown in Table 12.1, with the total number of alleles being $4n$ because each parent has two alleles and there are two parents per family. Falk and Rubinstein [199] suggested utilizing this classification for estimating the odds ratio (OR) by

$$\widehat{HRR} = \frac{n_{TA} \cdot n_{Na}}{n_{Ta} \cdot n_{NA}}. \quad (12.1)$$

With the assumption of a low disease prevalence, this OR is similar to the relative risk (RR). This estimator based on the parental haplotypes is thus termed the *haplotype relative risk* (HRR).

Testing for association is similarly straightforward (see Section 11.2). Under the assumption that the two alleles of each parent are independent, the allelic HRR test of independence can be formulated with

$$\chi_{HRR}^2 = \frac{4n(n_{Ta} \cdot n_{NA} - n_{TA} \cdot n_{Na})^2}{n_T \cdot n_N \cdot n_A \cdot n_a},$$

which is asymptotically χ^2 distributed with one degree of freedom (d.f.) under the null hypothesis of independence.

The estimator from Eq. (12.1) is termed haplotype relative risk. But is it truly equivalent to a RR estimated from cases and controls? The properties of the HRR were analyzed by Knapp and colleagues [364] for different modes of inheritance. Assuming positive linkage disequilibrium (LD, see Section 10.2) between the marker and the disease, they proved that the HRR is not larger than an increased RR . More generally, they showed $|HRR - 1| \leq |RR - 1|$, regardless of the recombination fraction. Concerning the estimators for HRR and RR , they showed that the estimator of the RR is stochastically larger than the estimator of the HRR:

$$P(\widehat{RR} \geq x) \geq P(\widehat{HRR} \geq x) \text{ if } \mathcal{D} > 0, \quad P(\widehat{RR} \leq x) \geq P(\widehat{HRR} \leq x) \text{ if } \mathcal{D} < 0.$$

This means that irrespective of the recombination frequency, the HRR will not overestimate the RR .

Although the HRR approach is easy and intuitive, it is based on the crucial assumption that the parental alleles are independent. Also, the allele frequency at the investigated marker has to be similar in all trios. To overcome these drawbacks, a matched analysis as described below was developed that has become the standard.

Table 12.1 Observed allele frequencies at a diallelic genetic marker for cases and pseudo-controls, the latter consisting of the non-transmitted parental genotypes.

	A	a	Total
Transmitted alleles	n_{TA}	n_{Ta}	n_T
Non-transmitted alleles	n_{NA}	n_{Na}	n_N
Total	n_A	n_a	$4n$

12.2 WHAT IS THE TRANSMISSION DISEQUILIBRIUM TEST?

The more famous variant of the trio design is the TDT, which is usually ascribed to Spielman and colleagues [627]. Its importance for genetic epidemiology was recently emphasized by *Human Heredity* in its 15th anniversary special issue [256]. This test does not rely on the construction of pseudo-controls. Instead, the core is the comparison of transmitted and non-transmitted alleles in a statistical test for matched samples. More specifically, it evaluates whether the proportion of transmitted alleles from heterozygous parents to the affected offspring deviates from the 50% that would be expected under Mendelian frequency assuming no linkage.

To fully acknowledge the features of the TDT, we need to derive the test statistic and discuss its properties in some detail. For this, let us consider a diallelic trait locus with alleles D and N . As in the previous chapters, D is the disease and N the normal allele. The allele frequencies are given by $P(D) = p$ and $P(N) = 1 - p = \bar{p}$. We are investigating a diallelic marker locus with alleles A and a , having frequencies $P(A) = q$ and $P(a) = 1 - q = \bar{q}$, respectively. For the following, we assume that the alleles D and A are in positive LD, hence,

$$\begin{aligned} P(DA) &= pq + \mathcal{D}, & P(NA) &= \bar{p}q - \mathcal{D}, \\ P(Da) &= p\bar{q} - \mathcal{D}, & P(Na) &= \bar{p}\bar{q} + \mathcal{D}. \end{aligned}$$

In the first step, let us investigate which parental haplotypes are possible in general. These are listed in Table 12.2 (left-hand side column). Given a specific parental haplotype, only a limited number of constellations for the transmission and non-transmission of alleles to the child are possible; these are shown in Table 12.2.

Using these and the probabilities derived before, we can calculate the probabilities for specific constellations of transmissions and non-transmissions. For example, a disease allele D is transmitted together with allele A at the marker locus, and at the same time, another allele A is not transmitted to the child in all constellations where both parental haplotypes are $\begin{smallmatrix} D \\ A \end{smallmatrix}$. The probability that both parental haplotypes are $\begin{smallmatrix} D \\ A \end{smallmatrix}$ is $(pq + \mathcal{D})^2$. Similarly, with no preferential transmission, haplotype $\begin{smallmatrix} D \\ A \end{smallmatrix}$ is transmitted in 50% of the parents with one haplotype being $\begin{smallmatrix} D \\ A \end{smallmatrix}$ and the other haplotype being $\begin{smallmatrix} N \\ A \end{smallmatrix}$. The probability that a parent has these two haplotypes is $2(\bar{p}q - \mathcal{D})(pq + \mathcal{D})$ and thus the probability for a transmitted $\begin{smallmatrix} D \\ A \end{smallmatrix}$ haplotype and a non-transmitted A allele is

$$P(T = \begin{smallmatrix} D \\ A \end{smallmatrix}, NT = A) = (pq + \mathcal{D})^2 + \frac{1}{2} \cdot 2(\bar{p}q - \mathcal{D})(pq + \mathcal{D}) = pq^2 + \mathcal{D}q,$$

where T denotes transmitted and NT non-transmitted. The other probabilities can be derived analogously and are shown in Table 12.3. One only has to note that the recombination fraction θ is involved if a parent is heterozygous at both the trait and the marker locus. As a consequence of these calculations, the row sum of Table 12.3 gives the marginal probabilities for the different haplotypes in the offspring.

So far, we have ignored the underlying study design: all families consist of trios with a single affected offspring (SAO). If we focus on a recessive model, only transmitted haplotypes including a D allele are of interest. As a consequence, we

Table 12.2 Possible parental haplotypes and probabilities for transmitted and non-transmitted haplotype.

Parental haplotypes	Transmitted haplotype Non-transmitted haplotype							
	$\begin{smallmatrix} D \\ A \end{smallmatrix} \begin{smallmatrix} \cdot \\ A \end{smallmatrix}$	$\begin{smallmatrix} D \\ A \end{smallmatrix} \begin{smallmatrix} \cdot \\ a \end{smallmatrix}$	$\begin{smallmatrix} D \\ a \end{smallmatrix} \begin{smallmatrix} \cdot \\ A \end{smallmatrix}$	$\begin{smallmatrix} D \\ a \end{smallmatrix} \begin{smallmatrix} \cdot \\ a \end{smallmatrix}$	$\begin{smallmatrix} N \\ A \end{smallmatrix} \begin{smallmatrix} \cdot \\ a \end{smallmatrix}$	$\begin{smallmatrix} N \\ A \end{smallmatrix} \begin{smallmatrix} \cdot \\ a \end{smallmatrix}$	$\begin{smallmatrix} N \\ a \end{smallmatrix} \begin{smallmatrix} \cdot \\ A \end{smallmatrix}$	$\begin{smallmatrix} N \\ a \end{smallmatrix} \begin{smallmatrix} \cdot \\ a \end{smallmatrix}$
$\begin{smallmatrix} D \\ A \end{smallmatrix} \begin{smallmatrix} D \\ A \end{smallmatrix}$	1							
$\begin{smallmatrix} D \\ A \end{smallmatrix} \begin{smallmatrix} D \\ a \end{smallmatrix}$		$\frac{1}{2}$	$\frac{1}{2}$					
$\begin{smallmatrix} D \\ a \end{smallmatrix} \begin{smallmatrix} D \\ A \end{smallmatrix}$		$\frac{1}{2}$	$\frac{1}{2}$					
$\begin{smallmatrix} D \\ a \end{smallmatrix} \begin{smallmatrix} D \\ a \end{smallmatrix}$				1				
$\begin{smallmatrix} N \\ A \end{smallmatrix} \begin{smallmatrix} D \\ A \end{smallmatrix}$	$\frac{1}{2}$				$\frac{1}{2}$			
$\begin{smallmatrix} D \\ A \end{smallmatrix} \begin{smallmatrix} N \\ A \end{smallmatrix}$	$\frac{1}{2}$				$\frac{1}{2}$			
$\begin{smallmatrix} N \\ A \end{smallmatrix} \begin{smallmatrix} D \\ a \end{smallmatrix}$		$\frac{\theta}{2}$	$\frac{1-\theta}{2}$			$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	
$\begin{smallmatrix} D \\ a \end{smallmatrix} \begin{smallmatrix} N \\ A \end{smallmatrix}$		$\frac{\theta}{2}$	$\frac{1-\theta}{2}$			$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	
$\begin{smallmatrix} N \\ a \end{smallmatrix} \begin{smallmatrix} D \\ A \end{smallmatrix}$		$\frac{1-\theta}{2}$	$\frac{\theta}{2}$			$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	
$\begin{smallmatrix} D \\ A \end{smallmatrix} \begin{smallmatrix} N \\ a \end{smallmatrix}$		$\frac{1-\theta}{2}$	$\frac{\theta}{2}$			$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	
$\begin{smallmatrix} N \\ a \end{smallmatrix} \begin{smallmatrix} D \\ a \end{smallmatrix}$				$\frac{1}{2}$				$\frac{1}{2}$
$\begin{smallmatrix} D \\ a \end{smallmatrix} \begin{smallmatrix} N \\ a \end{smallmatrix}$				$\frac{1}{2}$				$\frac{1}{2}$
$\begin{smallmatrix} N \\ A \end{smallmatrix} \begin{smallmatrix} N \\ A \end{smallmatrix}$					1			
$\begin{smallmatrix} N \\ A \end{smallmatrix} \begin{smallmatrix} N \\ a \end{smallmatrix}$						$\frac{1}{2}$	$\frac{1}{2}$	
$\begin{smallmatrix} N \\ a \end{smallmatrix} \begin{smallmatrix} N \\ A \end{smallmatrix}$						$\frac{1}{2}$	$\frac{1}{2}$	
$\begin{smallmatrix} N \\ a \end{smallmatrix} \begin{smallmatrix} N \\ a \end{smallmatrix}$								1

condition on a D allele being present in the transmitted haplotypes given in Table 12.3. Thus, the two bottom lines are neglected, and the probabilities are divided by the probability for the D allele, which is p . Table 12.4 gives the resulting conditional probabilities and differentiates the transmitted alleles with the non-transmitted alleles at the marker locus. As abbreviation for these probabilities, the first index letter denotes the transmitted allele, while the second denotes the non-transmitted allele. For example, $p_{A,a} = P(T=A, NT=a|SAO)$ is the probability that the A allele has been transmitted from a parent to the single affected offspring, while the a allele has not been transmitted.

To test for differential transmission, we consider heterozygous parents only, hence the off-diagonal elements in Table 12.4, and investigate whether A has been trans-

Table 12.3 Probabilities of transmitted haplotypes and non-transmitted alleles in children.

Transmitted haplotype	Non-transmitted allele		Total
	<i>A</i>	<i>a</i>	
$\begin{smallmatrix} D \\ A \end{smallmatrix}$	$pq^2 + \mathcal{D}q$	$pq\bar{q} + \mathcal{D}(\bar{q} - \theta)$	$pq + (1 - \theta)\mathcal{D}$
$\begin{smallmatrix} D \\ a \end{smallmatrix}$	$pq\bar{q} - \mathcal{D}(q - \theta)$	$p\bar{q}^2 - \mathcal{D}\bar{q}$	$p\bar{q} - (1 - \theta)\mathcal{D}$
$\begin{smallmatrix} N \\ A \end{smallmatrix}$	$\bar{p}q^2 - \mathcal{D}q$	$\bar{p}q\bar{q} - \mathcal{D}(\bar{q} - \theta)$	$\bar{p}q - (1 - \theta)\mathcal{D}$
$\begin{smallmatrix} N \\ a \end{smallmatrix}$	$\bar{p}q\bar{q} + \mathcal{D}(q - \theta)$	$\bar{p}\bar{q}^2 + \mathcal{D}\bar{q}$	$\bar{p}\bar{q} + (1 - \theta)\mathcal{D}$
Total	<i>q</i>	$\bar{q}$	1

mitted with a different frequency than the *a* allele. More formally, we test the hypotheses

$$H_0: \frac{p_{A,a}}{p_{A,a} + p_{a,A}} = \frac{1}{2} \quad \text{versus} \quad H_1: \frac{p_{A,a}}{p_{A,a} + p_{a,A}} \neq \frac{1}{2}.$$

For this test problem, tests for matched samples, i.e., tests of the McNemar type, are adequate. These are based on the observed frequencies of transmitted and non-transmitted alleles from one parent (see Table 12.5), where the marginals are the entries of Table 12.1. The standard McNemar test statistic is given by

$$T_{\text{TDT}} = \frac{(n_{A,a} - n_{a,A})^2}{n_{A,a} + n_{a,A}}. \quad (12.2)$$

Under the null hypothesis of equal transmission from heterozygous parents, T_{TDT} is asymptotically χ_1^2 distributed. For small sample sizes, the exact distribution can be computed based on the binomial distribution with sample size $(n_{A,a} + n_{a,A})$ and probability $\frac{1}{2}$. For testing a one-sided alternative, one usually employs $\sqrt{T_{\text{TDT}}}$, which is asymptotically standard normally distributed under H_0 . Critical regions then take into account the sign of $n_{A,a} - n_{a,A}$.

Table 12.4 Probabilities of transmitted and non-transmitted alleles for a recessive genetic model in families consisting of a single affected offspring and two parents.

Transmitted allele	Non-transmitted allele		Total
	<i>A</i>	<i>a</i>	
<i>A</i>	$p_{A,A} = (q + \frac{\mathcal{D}}{p})q$	$p_{A,a} = (q + \frac{\mathcal{D}}{p})\bar{q} - \theta\frac{\mathcal{D}}{p}$	$q + (1 - \theta)\frac{\mathcal{D}}{p}$
<i>a</i>	$p_{a,A} = (\bar{q} - \frac{\mathcal{D}}{p})q + \theta\frac{\mathcal{D}}{p}$	$p_{a,a} = (\bar{q} - \frac{\mathcal{D}}{p})\bar{q}$	$\bar{q} - (1 - \theta)\frac{\mathcal{D}}{p}$
Total	$q + \theta\mathcal{D}/p$	$\bar{q} - \theta\mathcal{D}/p$	1

The question is, however, what does the TDT statistic of Eq. (12.2) test? This can be easily seen by contrasting the theoretical probabilities for the off-diagonal elements in Table 12.4, yielding

$$p_{A,a} - p_{a,A} = \frac{D}{p}(1-2\theta).$$

Under H_0 , this difference is 0, which occurs either if $D = 0$ or if $\theta = \frac{1}{2}$. This means either

that linkage equilibrium between the trait and the marker allele must be present, case $D = 0$, or that there is no linkage between the trait and the marker locus, case $\theta = \frac{1}{2}$. Turned around, the null hypothesis can be rejected only if there are both association and linkage. Through this, we have formulated the core feature of the TDT, which is that the TDT is a simultaneous test for linkage and association.

Compared to the HRR, where alleles are separately considered and the total sample size is $4n$, the total sample size is halved for the TDT because paired data are used (see Table 12.5). Furthermore, only the off-diagonal elements of Table 12.5 are used in the TDT; hence, only heterozygous parents are informative.

The calculation of the TDT is illustrated in Example 12.1 and can be practiced in Problem 12.1. For practical purposes, the TDT is implemented in several software packages, including, for example, S.A.G.E. Also, a stand-alone implementation TDT is available.

Table 12.5 Observed frequencies for transmitted and non-transmitted alleles for the transmission disequilibrium test using data from families with a single affected offspring and both parents.

Transmitted allele	Non-transmitted allele		Total
	<i>A</i>	<i>a</i>	
<i>A</i>	$n_{A,A}$	$n_{A,a}$	n_{TA}
<i>a</i>	$n_{a,A}$	$n_{a,a}$	n_{Ta}
Total	n_{NA}	n_{Na}	$2n$

12.3 HOW CAN THE RISK BE ESTIMATED FROM FAMILY DATA?

In Section 12.1, the HRR statistic was introduced as the first design for the classical trio situation. In this line, the HRR had been proposed as an estimator of the *OR*. However, in contrast to genotype relative risks (GRRs) as described in Chapter 2, the HRR is based on transmitted alleles, not on genotypes. Hence, it is a weighted average of the *OR*s for affection being homozygous with the disease allele *D* compared to being homozygous for the normal allele *N*, and for affection being *D/N* compared to *N/N*. To disentangle this problem, the corresponding GRRs might be estimated, which will be the focus of this section.

Several approaches for estimating GRRs have been proposed in the past 15 years, starting with the well-recognized papers by Self and colleagues [604] and Schaid and Sommer [585], followed by several extensions, see, e.g., Refs. [145, 147, 367]. Here, we restrict our attention to a simple method for estimating the GRRs and corresponding confidence intervals for trio data consisting of both parents and an affected offspring. This method is based on the work of Schaid and Sommer [585], and the corresponding asymptotic and exact confidence intervals have been proposed by Scherag and colleagues [589].

In the following, we consider a diallelic trait locus that is identical to the marker locus and, for simplicity, is assumed to be in Hardy–Weinberg equilibrium (HWE). The methods can be extended easily to the case of a marker not being in HWE, and the reader may refer to the literature [589]. The GRRs have already been defined in Section 2.3.1. Using the Bayes formula, they can alternatively be written in the form (see Problem 12.2)

$$\gamma_1 = \frac{P(DN|\text{affected})P(NN)}{P(NN|\text{affected})P(DN)} \quad \text{and} \quad \gamma_2 = \frac{P(DD|\text{affected})P(NN)}{P(NN|\text{affected})P(DD)}. \quad (12.3)$$

In addition, the *population attributable risk (PAR)*, which is the proportion of affected individuals in the population that can be ascribed to the disease genotype, can be defined by [585]

$$\begin{aligned} \text{PAR} &= \frac{P(\text{affected}) - P(\text{affected}|\text{no disease genotype})}{P(\text{affected})} \\ &= \frac{p^2 f_2 + 2pqf_1 + q^2 f_0 - f_0}{p^2 f_2 + 2pqf_1 + q^2 f_0} = \frac{p^2(\gamma_2 - 1) + 2pq(\gamma_1 - 1)}{p^2\gamma_2 + 2pq\gamma_1 + q^2}. \end{aligned}$$

For estimating these parameters, let n be the total number of trios in the sample; n_0 , n_1 , and n_2 the number of families in which zero, one, and two D alleles, respectively, were transmitted to the offspring; and K the number of D alleles among the $2n$ alleles that were not transmitted to the children.

The maximum likelihood (ML) estimator for the frequency of the disease allele is $\hat{p} = K/2n$. $P(DN|\text{affected})$ and $P(NN|\text{affected})$ can be estimated by the number of trios in which one and zero D alleles were transmitted to the affected child, respectively. The probability of carrying no and one disease allele is estimated by $(1 - \hat{p})^2$ and $2\hat{p}(1 - \hat{p})$, respectively. Hence, using Eq. (12.3) and the derivations by Scherag and colleagues [589], ML estimators of γ_1 and γ_2 are given by

$$\hat{\gamma}_1 = \frac{n_1(1 - \hat{p})}{2n_0\hat{p}} \quad \text{and} \quad \hat{\gamma}_2 = \frac{n_2(1 - \hat{p})^2}{n_0\hat{p}^2}.$$

Finally, inserting the obtained estimators into the formula for the PAR shown above yields

$$\widehat{\text{PAR}} = 1 - \frac{4n n_0}{(2n - K)^2}.$$

Both asymptotic and exact confidence intervals for the parameters have been derived by Scherag and colleagues [589]. Referring to their work for details, we present only their results for the approach that assumes HWE. We abbreviate $a = 1 + \hat{p} \cdot (2\hat{\gamma}_1 - 1)$ and $b = 1 + \hat{p} + \hat{p}^2(\hat{\gamma}_1 - 1)$ to denote the approximate confidence intervals for a coverage probability of $1 - \alpha$. Then, the asymptotic confidence intervals for p , γ_1 , γ_2 , and PAR are given by

$$\text{expit} \left(\text{logit}(\hat{p}) \pm z_{1-\frac{\alpha}{2}} (2n\hat{p}(1 - \hat{p}))^{-1/2} \right),$$

$$\begin{aligned} & \hat{\gamma}_1 \cdot \exp \left(\pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{a\hat{\gamma}_2\hat{p}^2 + (a^2 + \hat{\gamma}_1)(1-\hat{p})}{2n\hat{p}(1-\hat{p})^2\hat{\gamma}_1}} \right), \\ & \hat{\gamma}_2 \cdot \exp \left(\pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\gamma}_2^2\hat{p}^4 + 2b\hat{\gamma}_2\hat{p}(1-\hat{p}) + a(1-\hat{p})^3}{n\hat{p}^2(1-\hat{p})^2\hat{\gamma}_2}} \right), \end{aligned}$$

and

$$\widehat{\text{PAR}} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}}{n(1-\hat{p})^2} \cdot \frac{2(1-\hat{p})(1+\hat{\gamma}_1) + \hat{\gamma}_2\hat{p}}{(\hat{\gamma}_2\hat{p}^2 + a(1-\hat{p}))^2}},$$

respectively, where logit denotes the logistic function $\text{logit}(x) = \ln \frac{x}{1-x}$ and expit its inverse. For its use, see Problem 12.1.

We again want to stress that these estimates are valid only if the disease locus itself is genotyped, HWE holds at the disease locus, and asymptotic results can be applied. In most applications, however, it is uncertain whether the analyzed marker is the truly disease-causing variant. Using these derived estimates for the underlying genetic effect might therefore be biased. Investigating the extent and direction of this bias, we found that it crucially depends on the LD between disease and marker locus, with greater LD leading to greater bias [221]. In addition, the bias is more pronounced if the allele frequencies of the genetic marker and the disease locus differ more. In most instances, the bias is toward the null hypothesis, so that the risk estimates can well be used if the marker is not identical to the disease locus. However, we did not investigate the situation of deviations from HWE at the marker and/or the trait locus. Furthermore, we considered only cases where the approximation through the asymptotic distribution was valid.

12.4 HOW CAN SAMPLE SIZE AND POWER BE CALCULATED FOR TRIO DATA?

Apart from quantifying the contribution of the mutation to the disease, estimation of GRRs is important both for planning further studies and sample size and for power calculation. These computations can be done easily for the HRR approach with the methods discussed in the previous chapter (see Section 11.3), where required sample sizes for the OR have been calculated.

For the TDT, we distinguish the following two situations. In the first, results from a previous study are available. In this very simple situation, the available data can be utilized for sample size and power calculation. We repeat the general sample size formula from Section 8.5.2 for the two-sided alternative hypothesis

$$n \geq (z_{1-\beta} + z_{1-\frac{\alpha}{2}})^2 \cdot \frac{1}{\Delta^2},$$

where $\Delta = \frac{\mu_1 - \mu_0}{\sigma}$ for the one-sample Gauss test. The power $1 - \beta$ analogously is given by $1 - \Phi(z_{1-\frac{\alpha}{2}} + \Delta\sqrt{n})$, with Φ , as before, denoting the cumulative

distribution function of the standard normal distribution. If the alternative hypothesis is one-sided, $z_{1-\frac{\alpha}{2}}$ is replaced by $z_{1-\alpha}$ in the respective formulae.

In the TDT, we test for preferential transmission. Thus, we investigate whether the transmission probability for the A allele deviates from 50%. We can therefore replace μ_1 by p_T , μ_0 by $\frac{1}{2}$, and σ by $p_T(1 - p_T)$, with p_T denoting the transmission probability under H_1 . Note that with this approach, we calculate the required number of heterozygous parents. Therefore, in the second step, we have to compute the total sample size by taking into account the estimated probability for a parent to be heterozygous.

Table 12.6 Exemplary data for estimation of the sample size in the transmission disequilibrium test (TDT) based on previous results.

Transmitted allele	Non-transmitted allele	
	A	a
A	55	30
a	10	5

Example 12.1. Consider the notional data from 50 trio families with an SAO displayed in Table 12.6. The TDT statistic is $T_{TDT} = (30 - 10)^2 / (30 + 10) = 10$, giving a p -value of 0.0016 from the χ^2 distribution with 1 d.f. The transmission probability for the A allele is estimated to be $\hat{p}_T = 30 / (30 + 10) = 0.75$. The aim is to validate the preferential transmission of the A allele in a second trio sample at the 5% test level with 90% power. In the validation study, the alternative hypothesis is considered to be one-sided. Therefore, $z_{1-\alpha} = z_{0.95} \approx 1.6449$ and $z_{1-\beta} = z_{0.9} \approx 1.2816$. Δ is approximated by $(\hat{p}_T - \frac{1}{2}) / \sqrt{\hat{p}_T(1 - \hat{p}_T)} \approx 0.5774$ so that the required number of heterozygous parents for the validation study is given by $n \geq (z_{1-\beta} + z_{1-\alpha})^2 \cdot \frac{1}{\Delta^2} \approx (1.6449 + 1.2816)^2 \cdot \frac{1}{0.5774^2} \approx 25.6915$. The proportion of heterozygous parents in the sample is $(30 + 10) / 100 = 0.4$; thus, at least $\frac{26}{0.4} = 65$ parents, corresponding to 33 nuclear families with an SAO and both parents, are required in the validation study.

In the second situation for calculating sample sizes for the TDT, however, the detailed information required for the former approach is not given. Also, one might not want to rely on the new study population and the result being similar to a previous one. If one is not willing to simply adopt the previous values, more elaborate sample size calculations have to be performed. In a seminal paper, Risch and Merikangas [565] derived sample size formulae using the GRR assuming a multiplicative genetic model. For arbitrary modes of inheritance, formulae have been given by Knapp [360]. It is important to note that both approaches assume complete LD between the disease and the investigated marker locus. If, however, LD is not maximal or if the disease and the marker locus have different allele frequencies, a loss in power usually occurs, as demonstrated in the context of association studies with unrelated probands in Section 11.3.

12.5 WHAT ARE ALTERNATIVE STATISTICS TO THE TRANSMISSION DISEQUILIBRIUM TEST?

The original TDT as described above is based on the transmission of a specific allele. An alternative representation of the transmission pattern in trios has been described by Khoury [351] and Schaid and Sommer [586]. Considering again a marker with alleles A and a , the number of A alleles that are transmitted or not transmitted are counted as shown in Table 12.7.

From this, the relative risks RR_1 and RR_2 , specifying the risk for carriers of one or two A alleles over carriers of no A allele, respectively, as well as the risk RR_3 for carriers of two A alleles over carriers of one A allele can be estimated to be

Table 12.7 Transmitted and non-transmitted number of A alleles.

Transmitted	Non-transmitted		
	0	1	2
0	$n_{0,0}$	$n_{0,1}$	$n_{0,2}$
1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$
2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$

$$\widehat{RR}_1 = n_{1,0}/n_{0,1}, \quad \widehat{RR}_2 = n_{2,0}/n_{0,2}, \quad \widehat{RR}_3 = n_{2,1}/n_{1,2}.$$

Summarizing these ratios, a matched summary statistic is obtained with

$$\chi_{MS}^2 = \frac{(n_{1,0} - n_{0,1} + n_{2,0} - n_{0,2} + n_{2,1} - n_{1,2})^2}{n_{1,0} + n_{0,1} + n_{2,0} + n_{0,2} + n_{2,1} + n_{1,2}},$$

which is asymptotically χ_1^2 distributed under the null hypothesis $H_0: RR_1 = RR_2 = RR_3 = 1$.

In addition to these statistics, specific genetic models can be tested via a likelihood approach using a score statistic [586]. For example, the restriction $RR_1 = RR_2$ yields a dominant model, and $RR_1 = 1$ a recessive model, leading to the following test statistics for the dominant and the recessive genetic model:

$$\chi_{dom}^2 = \frac{\left((n_{2,0} + n_{1,1} - \frac{3n_2}{4}) + (n_{1,0} - \frac{n_1}{2})\right)^2}{\frac{3n_2}{16} + \frac{n_1}{4}}, \quad \chi_{rec}^2 = \frac{\left((n_{2,1} - \frac{n_3}{2}) + (n_{2,0} - \frac{n_2}{4})\right)^2}{\frac{n_3}{4} + \frac{3n_2}{16}}.$$

Different from previous sections, here n_1 , n_2 , and n_3 denote the number of parental matings with one, two, and three A alleles, respectively. Using our above notations, these are calculated from $n_1 = n_{0,1} + n_{1,0}$, $n_2 = n_{0,2} + n_{1,1} + n_{2,0}$, and $n_3 = n_{1,2} + n_{2,1}$.

Having described different methods that are all based on the same kind of sample, i.e., two parents with an SAO, the question is which statistic one should use in a given situation. The answer to this mostly depends on the genetic model assumed to underlie the trait under study. Simulation studies have confirmed that in the case of a dominant, a recessive, and an additive mode of inheritance, χ_{dom}^2 , χ_{rec}^2 , and the classical TDT, respectively, perform best [586]. If the mode of inheritance is unknown, the TDT is preferable because it performs well under a variety of models.

12.6 HOW CAN MORE THAN TWO ALLELES AT THE MARKER BE ANALYZED IN THE TDT?

So far, we have considered a very simple situation for the TDT. Specifically, we assumed that trios with one affected child and both parents are available and that one diallelic marker is genotyped. Family structures that deviate from the classical TDT design will be considered in Section 12.7, whereas in this section we will focus on extensions to multiallelic markers. If a marker with m alleles $A_1, A_2, \dots, A_m$ has been typed, Table 12.5 needs to be extended to Table 12.8.

12.6.1 Test of single alleles

The easiest approach to deal with multiple alleles is to simply compute several TDTs, namely, one for each allele against all others. If the marker has m alleles and hence m tests are performed, the result needs to be corrected for multiple testing.

To reduce the number of tests and thereby the need for correction, it might be possible to specify a smaller number of alleles a priori that are of specific interest. This selection might be based on previous studies or on presumed functional relevance. Alternatively, if no such specification is possible, it has been suggested to simply test the most frequent allele. The drawback of this approach is that even if a meaningful reduction can be done, the TDT might lose its specific property of testing for linkage and association and be a test for linkage only. However, if only a single allele is of interest, this approach is recommended.

An alternative strategy is to dichotomize the alleles of an STR marker according to their length, resulting in long alleles and short alleles. This becomes meaningful if the evolution and mutation pattern of STR markers is taken into account [100]: mutations are primarily single-step mutations [733]. Because repeating tri-, tetra-, and pentanucleotide sequences are also associated with a number of hereditary diseases [713], this approach is definitely promising for some disorders.

Table 12.8 Observed frequencies and theoretical probabilities in parenthesis for transmitted and non-transmitted alleles for the transmission disequilibrium test for a multiallelic marker using data from affected children and their parents.

Transmitted allele	Non-transmitted allele				Total
	A_1	A_2	$\dots$	A_m	
A_1	$n_{1,1} (p_{1,1})$	$n_{1,2} (p_{1,2})$	$\dots$	$n_{1,m} (p_{1,m})$	$n_{1,\cdot}$
A_2	$n_{2,1} (p_{2,1})$	$n_{2,2} (p_{2,2})$	$\dots$	$n_{2,m} (p_{2,m})$	$n_{2,\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
A_m	$n_{m,1} (p_{m,1})$	$n_{m,2} (p_{m,2})$	$\dots$	$n_{m,m} (p_{m,m})$	$n_{m,\cdot}$
Total	$n_{\cdot,1}$	$n_{\cdot,2}$	$\dots$	$n_{\cdot,m}$	$n_{\cdot,\cdot} = n$

12.6.2 Global test statistics

Instead of testing single alleles, one of the several global test statistics may be applied. For instance, for an m -allelic marker, the Bowker test of symmetry can be applied [58]. This is the multivariate extension of the McNemar test and tests the null hypothesis $H_0 : p_{i,j} = p_{j,i} \forall i \neq j$. The test statistic is given by

$$T_c = \sum_{i < j}^m \frac{(n_{i,j} - n_{j,i})^2}{n_{i,j} + n_{j,i}}$$

and under H_0 of no asymmetry is approximately $\chi_{m(m-1)/2}^2$ distributed. The disadvantage of this procedure is that the number of d.f. is high for markers with many alleles. Hence, the test will rapidly lose power if the marker is highly polymorphic. Also, each genotype is separately tested, thus disregarding the fact that one allele is present in more than one genotype. If one allele is transmitted preferentially, this should be the case in each of the genotypes where it is present.

An alternative global test is the *marginal homogeneity test*. For this, the test statistic

$$T_m = \sum_{i=1}^m \frac{(n_{i,\cdot} - n_{\cdot,i})^2}{n_{i,\cdot} + n_{\cdot,i}}$$

is used, which is approximately χ_{m-1}^2 distributed under H_0 of marginal homogeneity. It hence has inherently higher power than T_c . However, it shows the disadvantage of testing only for linkage, not for association.

To recapture this property, the T_m has been modified to include only heterozygous parents. This leads to the statistic

$$T_{mhet} = \frac{m-1}{m} \sum_{i=1}^m \frac{(n_{i,\cdot} - n_{\cdot,i})^2}{n_{i,\cdot} + n_{\cdot,i} - 2n_{i,i}},$$

which again is approximately χ_{m-1}^2 distributed under H_0 [625]. In addition to having fewer d.f., the T_{mhet} tests for both linkage and association.

For all these statistics, one should take into account that many STR markers have alleles with low frequencies. This can result in a breakdown of both power and asymptotic properties. In this case, grouping of the low-frequency alleles is advisable. They are grouped either according to length or as a separate category termed “low-frequency alleles.” To obtain exact p -values for sparse contingency tables, a Monte-Carlo permutation approach can be employed.

For this, the easiest approach is to generate transmissions at random according to the null hypothesis of equal transmission, conditional on the observed parental genotypes [133, 345]. More specifically, under H_0 of no preferential transmission, alleles A_j and $A_{j'}$ will be transmitted to the SAO from a parent with genotype $A_j A_{j'}$ with equal frequency (see Algorithm 12.1).

Algorithm 12.1.

1. Compute the TDT statistic using the original n trio families
2. Set counter $C = 0$
3. For a specified number N of permutations,
 - (a) Permute the off-diagonal elements $n_{i,j} + n_{j,i}$ for $i = 1, \dots, m$ and $j > i$
 - (b) Compute the TDT statistic using the permuted data
 - (c) If the TDT statistic from the permuted data is at least as large as the TDT statistic using the original data, add 1 to C
4. The Monte--Carlo p -value is C/N

This permutation approach leads to another global TDT, termed $TDT_{\max}$. Here, the standard TDT is calculated for each allele A_i , and the maximum of the m TDT statistics is taken. With the permutation approach from Algorithm 12.1, a global p -value can be determined [476].

Table 12.9 Contingency table for the multi-allelic transmission. Observations of transmitted and non-transmitted alleles at a short tandem repeat with four alleles.

Transmitted allele	Non-transmitted allele				Total
	A_1	A_2	A_3	A_4	
A_1	4	3	5	7	19
A_2	3	3	3	4	13
A_3	1	4	3	3	11
A_4	1	1	4	1	7
Total	9	11	15	15	50

An interesting alternative approach to the contingency table methods presented in this section is the *extended TDT* or *ETDT* utilizing the logistic regression framework [607]. We conclude this section by illustrating the calculation of the multiallelic TDTs with notional data.

Example 12.2. Consider a study in which 25 trios have been recruited, and all children and their parents were successfully genotyped at an STR with the four alleles A_1 , A_2 , A_3 , and A_4 . The observed pattern of transmitted and non-transmitted alleles is shown in Table 12.9.

If a specific allele had been singled out beforehand, for instance, allele A_1 because of the results of a previous study, the classical TDT would be calculated using data only from parents who were heterozygous for the allele A_1 . This leads to $T_{TDT} = \frac{(15-5)^2}{15+5} = 5$. Utilizing the approximate χ_1^2 distribution, the associated p -value is approximately 0.0253.

If, instead, no previous result was known, a global test might be preferred. We stress that the table is sparse so that the approximation of the distribution of the test statistics may not be appropriate. The test of symmetry uses the diagonal elements within the table so that

$$T_c = \frac{(3-3)^2}{3+3} + \frac{(5-1)^2}{5+1} + \frac{(7-1)^2}{7+1} + \frac{(3-4)^2}{3+4} + \frac{(4-1)^2}{4+1} + \frac{(3-4)^2}{3+4} \approx 9.2524.$$

As the numbers of d.f. are given by $\frac{4(4-1)}{2} = 6$, the p -value for T_c from the χ^2_6 distribution is 0.1599. In contrast to the T_c statistic, the marginal homogeneity test utilizes only the row and column sums. The resulting statistic is

$$T_m = \frac{(19-9)^2}{19+9} + \frac{(13-11)^2}{13+11} + \frac{(11-15)^2}{11+15} + \frac{(7-15)^2}{7+15} \approx 7.2626$$

and the p -value utilizing the $\chi^2_{m-1} = \chi^2_3$ distribution is given by 0.0640. Finally, the T_{mhet} is similar to T_m but leaves out the instances of homozygous parents to retrieve the feature of simultaneously testing for linkage and association. The respective statistic for the example is given by

$$T_{mhet} = \frac{(19-9)^2}{19+9-24} + \frac{(13-11)^2}{13+11-23} + \frac{(11-15)^2}{11+15-23} + \frac{(7-15)^2}{7+15-21} \approx 9.2222.$$

This is the same value that would be obtained by summing all single-test statistics. T_{mhet} has an approximate χ^2_3 distribution, yielding $p \approx 0.0265$.

12.7 HOW CAN DIFFERENT FAMILY STRUCTURES BE ANALYZED?

In the previous sections, we discussed the analysis of a genetic marker with two or more alleles in the classical TDT situation. Specifically, this means that trios have to be available consisting of one affected child and both parents. However, in practice, it might be difficult to obtain data from both or even one parent. Especially for late-onset diseases, parents are often unavailable, so there is a clear need for further development of the original test statistic to analyze these partial data. However, the methods dealing with missing parental data usually require stronger assumptions and thus do not keep the full robust property of the original TDT. This will be discussed in Section 12.7.1 in some detail.

In other situations, available data might deviate from the classical trio structure. For instance, there might be a sibling of the affected child who may or may not be affected (see Section 12.7.2). Or, more generally, there could be more complex family structures with affected and unaffected members. Therefore, the obvious question arises how one might incorporate this additional information in the analysis.

As a simple solution, if, instead of trios, larger pedigrees with affected members are available, one obvious choice would be to split the family into several trios with an SAO. Then, the TDT could be calculated in every trio. As the χ^2 statistic assumes independent observations that would then be violated, the resulting test is only a test for linkage, not for association. One alternative is to select one trio from each pedigree for analysis, so that it is again tested both for linkage and for association. However, valuable information might be lost by discarding the other family members. A valid test for linkage and association for families with multiple affected and unaffected family members is described in Section 12.7.3.

12.7.1 How can families with missing parental genotypes be analyzed?

If the sample consists of not only complete trios with both parents and an SAO but also families where the data from one parent is missing, the original TDT statistic cannot be computed from these families. The easiest solution to this problem is simply to exclude families with missing parental data. If, conditional on the parental genotypes and the offspring's phenotype, the missing of parental data is independent of the offspring's genotype for each sub-population in the data, the TDT using only complete trios is still valid [119]. However, it certainly leads to a severe loss of information.

To recapture some of the lost information, it seems straightforward to at least include those trios in which it can be deduced unambiguously that alleles were transmitted and non-transmitted from the available parent. For this, consider the three cases where the available parent is heterozygous at a diallelic marker and the affected offspring is heterozygous or homozygous with one or the other allele as shown in Figure 12.3. In case (a), it is clear that the mother transmitted allele A , whereas allele a was not transmitted. In contrast, in case (b), a was transmitted and A was not transmitted from the mother. The only unsolvable and hence uninformative setting is case (c), where allele A or a might have been inherited from the mother. However, to discard only these uninformative cases and to use the others might lead to biased information [151]. To understand this, imagine the situation that allele a is extremely rare in the general population. This means that most of the missing parents will have the genotype AA , and if the available parent is heterozygous and therefore potentially informative, the affected offspring might be either homozygous AA or heterozygous Aa . However, the available transmission is informative only if the child is homozygous AA and hence the allele A was transmitted. Consequently, the estimated transmission frequencies will be distorted in favor of the A allele being preferably transmitted to the affected offspring.

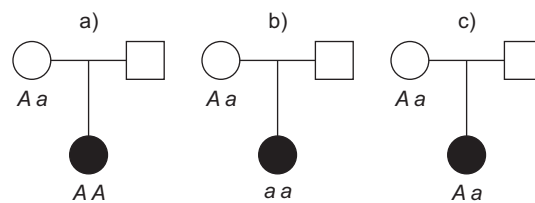


Fig. 12.3 Three cases with one missing and one heterozygous parent at a diallelic marker. Families (a) and (b) are informative, where the mother transmitted allele A and a , respectively. Family c) is uninformative, as the mother might have transmitted allele A or allele a .

To avoid this bias, only the data from the genotyped parent may be utilized, yielding the 1-TDT [650]. For this, the counts of the observed genotypes of the present parent and the offspring are tabulated as shown in Table 12.10. From these observed counts, the informative instances where the missing parent must have contributed an A allele ($n_{21} + n_{10}$) are contrasted with the informative cases where an a allele was transmitted from the missing parent ($n_{12} + n_{01}$). These numbers are utilized to calculate the

1-TDT that is asymptotically χ_1^2 distributed under H_0 of no linkage or no association with

$$T_1 = \frac{((n_{21} + n_{10}) - (n_{12} + n_{01}))^2}{(n_{21} + n_{10}) + (n_{12} + n_{01})}.$$

However, this simple version is valid only under the assumption that males and females with the same genotype have the same mating preference or that the father and mother in each trio are missing with the same probability. If both issues are violated, a more complicated version of the test statistic needs to be employed; for details see Ref. [650].

In applications, the sample will often comprise trios with both complete and missing parental data. In this case, the

Table 12.10 Observed genotype counts of the available parent and the affected offspring for the 1-transmission disequilibrium test (1-TDT).

Child's genotype	Parental genotype		
	<i>AA</i>	<i>Aa</i>	<i>aa</i>
<i>AA</i>	n_{22}	n_{21}	$n_{20} = 0$
<i>Aa</i>	n_{12}	n_{11}	n_{10}
<i>aa</i>	$n_{02} = 0$	n_{01}	n_{00}

original TDT from Eq. (12.2) can be combined with the 1-TDT, resulting in a score-type test. For this, the combined transmission count is calculated with $y_1 = n_{A,a} + (n_{21} + n_{10})$ having mean and variance under H_0 of no linkage or no association

$$\begin{aligned} \mathbb{E}_{H_0}(y_1) &= \frac{1}{2}(n_{A,a} + n_{a,A}) + \frac{1}{2}((n_{21} + n_{10}) + (n_{12} + n_{01})) \quad (12.4) \\ \text{and } \text{Var}_{H_0}(y_1) &= \frac{1}{4}(n_{A,a} + n_{a,A}) + \frac{1}{4}((n_{21} + n_{10}) + (n_{12} + n_{01})). \end{aligned}$$

The combined test statistic is then given by

$$T = \frac{y_1 - \mathbb{E}_{H_0}(y_1)}{\sqrt{\text{Var}_{H_0}(y_1)}},$$

which is asymptotically standard normally distributed under H_0 . The calculation of the 1-TDT alone and in combination with the TDT can be practiced in Problem 12.1.

For the 1-TDT, we considered SNP markers, and the missing parent was neglected in the analysis. An alternative approach for SNP markers is to reconstruct the missing parental alleles or the missing parental genotype by an expectation maximization (EM) approach [110, 126, 702]. However, these methods are valid only when the genotype distribution, conditional on the available data, among available parents is the same as that among the missing parents [11], and the approach of Clayton [126] also requires the assumption of no population stratification. In contrast, the 1-TDT is valid under population stratification, but it relies on the assumption that the genotype distribution among the missing parents does not differ from that among the observed [11]. If, however, the parental genotype influences the availability of the parent for genotyping, this assumption does not hold. An approach independent of this assumption is the *robust TDT (rTDT)* [600], which returns minimum and maximum values of TDT that are consistent with all the possible completions of the missing data. Another approach reconstructing missing parental data without estimating allele or genotype frequencies is the *reconstruction combined TDT (RC-TDT)* [361], which will be described in the next section in the context of tests using sibling data.

12.7.2 How can data from sibships be analyzed for transmission disequilibrium?

In the previous section, we dealt with only partially available parental data. A different approach if parental data are missing is to completely dispense with parental genotypes. Instead, siblings of the affected child can be used as controls, which will be the focus in this section.

An important question here about the selection of the sibling controls. In complex diseases, penetrance will usually be incomplete and will depend on age. Hence, it is usually recommended to recruit older unaffected siblings as controls to avoid the possibility that a younger sibling is unaffected simply because he or she is younger than the affected proband. However, for traits that are influenced by secular trends, this might lead to a confounding by time. Another issue is related to possible over-matching by using family controls, and the reader may refer to Witte and colleagues [726] for an in-depth discussion of this topic.

If families with at least one affected and one unaffected offspring have been recruited, the *sib TDT* (*S-TDT*) can be employed [626]. It should not be confused with the *sibship disequilibrium test* (*SDT*) [311], which will be discussed below. A requirement for the S-TDT is that the children in the same sibship must not all have the same genotype. Basically, the S-TDT compares the frequency of a specific marker allele A between affected and unaffected offspring. It is a valid test for linkage in the presence of association if the data from only one discordant sib-pair (DSP) per family are used; for multiple sibships, it is a test for linkage. Per family i , the number of affected and unaffected children is denoted by $n_{\text{aff},i}$ and $n_{\text{unaff},i}$, respectively, rendering the number of children in the family to be $n_i = n_{\text{aff},i} + n_{\text{unaff},i}$. For a diallelic marker testing for allele A , there are $n_{AA,i}$ siblings with the genotype AA and $n_{Aa,i}$ siblings with the genotype Aa . The test is based on the number of A alleles among the affected offspring, denoted by y . It can be shown (see Problem 12.3) that under H_0 of no preferential transmission, with i now being the family index,

$$\mathbb{E}_{H_0}(y) = \sum_{i=1}^n (2n_{AA,i} + n_{Aa,i}) \frac{n_{\text{aff},i}}{n_i} \quad \text{and} \quad (12.5)$$

$$\mathbb{V}\text{ar}_{H_0}(y) = \sum_{i=1}^n n_{\text{aff},i} n_{\text{unaff},i} \frac{4n_{AA,i}(n_i - n_{AA,i} - n_{Aa,i}) + n_{Aa,i}(n_i - n_{Aa,i})}{n_i^2(n_i - 1)}. \quad (12.6)$$

The score test statistic $T_{S-TDT} = (y - \mathbb{E}_{H_0}(y)) / \sqrt{\mathbb{V}\text{ar}_{H_0}(y)}$ is then asymptotically standard normally distributed and its use will be illustrated in Example 12.3.

If the sample also comprises classical trios without siblings, they can be added to the same test statistic. In addition to the number y of A alleles observed in the affected children, we use the statistic y_1 from the previous section with mean and variance under H_0 as given in Eq. (12.4). Then, under the null hypothesis, the sum $x = y + y_1$ has mean $\mathbb{E}_{H_0}(x) = \mathbb{E}_{H_0}(y) + \mathbb{E}_{H_0}(y_1)$ and variance $\mathbb{V}\text{ar}_{H_0}(x) = \mathbb{V}\text{ar}_{H_0}(y) + \mathbb{V}\text{ar}_{H_0}(y_1)$. Finally, we obtain the *combined TDT* statistic $T_{cTDT} = (x - \mathbb{E}_{H_0}(x)) / \sqrt{\mathbb{V}\text{ar}_{H_0}(x)}$ that is, under the null hypothesis, approximately standard normally distributed. If the sample includes families where both parents and siblings

have been genotyped, it is suggested to neglect the siblings and compute the original TDT in these families.

The S-TDT was proposed as a test for a specific allele against all others. This will be most useful in the case of a diallelic marker. An extension of the S-TDT explicitly allowing for the analysis of multiallelic markers has been suggested by Schaid and Rowland [583]. Implementations of the S-TDT are available as given in the URLs section.

For the same situation of families with at least one affected and one unaffected child, an alternative is the SDT [311], which is a valid test for both linkage and association regardless of the size of the sibship. It is an application of the standard sign test and can be used as follows. Considering a diallelic marker and using the same notation as before, the numbers of A alleles in the affected and the unaffected offspring are counted per family to be $n_{\text{aff},A,i}$ and $n_{\text{unaff},A,i}$, respectively. From this, the mean number of A alleles $m_{\text{aff},i} = n_{\text{aff},A,i}/n_{\text{aff},i}$ and $m_{\text{unaff},i} = n_{\text{unaff},A,i}/n_{\text{unaff},i}$ are computed. Under the null hypothesis, these should be equal, and the SDT is a sign test on their difference $d_i = m_{\text{aff},i} - m_{\text{unaff},i}$. Discarding all families where $d_i = 0$, b gives the number of sibships with $d_i > 0$ and c is the count of families with $d_i < 0$. The test statistic then takes the form of the sign test $T_{SDT} = (b - c)/\sqrt{b + c}$, which is asymptotically standard normally distributed.

An example of the practical computation is given in Example 12.3. For more details and extensions, e.g., to multiallelic markers or a combination with standard TDT data, the reader may refer to Horvath and Laird [311]. The SDT is implemented, for instance, in the very general software FBAT (see also Section 12.7.3).

Example 12.3. To demonstrate the computation of the S-TDT and the SDT, consider the three families with five affected and nine unaffected children shown in Figure 12.4. From this, we can deduce the data given in Table 12.11.

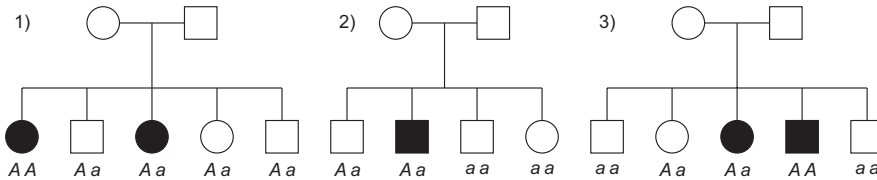


Fig. 12.4 Illustrative data for the computation of the sib transmission disequilibrium test (S-TDT) and the sibship disequilibrium test (SDT). Three families have been genotyped at one diallelic marker with alleles A and a .

For the S-TDT, we observe $y = 3 + 1 + 3 = 7$ A alleles in the affected offspring. Furthermore,

$$\begin{aligned} \mathbb{E}_{H_0}(y) &= \sum_{i=1}^n (2n_{AA,i} + n_{Aa,i}) \frac{n_{\text{aff},i}}{n_i} \\ &= (2 \cdot 1 + 4) \frac{2}{5} + (2 \cdot 0 + 2) \frac{1}{4} + (2 \cdot 1 + 2) \frac{2}{5} = 4.5, \end{aligned}$$

Table 12.11 Illustrative data to calculate the sib transmission disequilibrium test (S-TDT) and the sibship disequilibrium test (SDT).

Family	$n_{\text{aff},i}$	$n_{\text{unaff},i}$	n_i	$n_{AA,i}$	$n_{Aa,i}$	$n_{\text{aff},A,i}$	$n_{\text{unaff},A,i}$	$m_{\text{aff},i}$	$m_{\text{unaff},i}$
1	2	3	5	1	4	3	3	$\frac{3}{2}$	$\frac{3}{3}$
2	1	3	4	0	2	1	1	$\frac{1}{1}$	$\frac{1}{3}$
3	2	3	5	1	2	3	1	$\frac{3}{2}$	$\frac{1}{3}$

$$\begin{aligned}
\mathbb{V}ar_{H_0}(y) &= \sum_{i=1}^n n_{\text{aff},i} n_{\text{unaff},i} \frac{4n_{AA,i}(n_i - n_{AA,i} - n_{Aa,i}) + n_{Aa,i}(n_i - n_{Aa,i})}{n_i^2(n_i - 1)} \\
&= 2 \cdot 3 \cdot \frac{4 \cdot 1(5-1-4) + 4(5-4)}{5^2(5-1)} + 1 \cdot 3 \cdot \frac{4 \cdot 0(4-0-2) + 2(4-2)}{4^2(4-1)} \\
&\quad + 2 \cdot 3 \cdot \frac{4 \cdot 1(5-1-2) + 2(5-2)}{5^2(5-1)} \approx 1.3333,
\end{aligned}$$

rendering $T_{S-TDT} \approx \frac{7-4.5}{\sqrt{1.3333}} \approx 2.1561$, with a two-sided p -value of 0.0304.

For the SDT, we compute the differences in mean numbers of A alleles in affected and in unaffected offspring to be $d_1 = 3/2 - 3/3 = 0.5$, $d_2 = 1/1 - 1/3 \approx 0.6667$, and $d_3 = 3/2 - 1/3 \approx 1.1667$, yielding $b = 3$, $c = 0$, and $T_{SDT} = (3 - 0)/\sqrt{3} = \sqrt{3} \approx 1.7321$, and the two-sided asymptotic p -value is approximately 0.0833.

Other tests based on sibling data have been proposed in the literature. For instance, the S-TDT is equivalent to *SIBASSOC* [150] and a statistic by Boehnke and Langefeld [69] if a diallelic marker and a single DSP per family are considered [470]. For multiallelic markers, *SIBASSOC* shows lower power than the others. Implementation of these tests is available as given in the URLs section at the end of the chapter.

It has also been suggested to combine the approaches of reconstructing parental genotypes and utilizing sibling data for a better exhaustion of the available data. In line with this suggestion, Knapp [361] proposed to first reconstruct parental data. If this cannot be achieved unambiguously, siblings are used to compute the S-TDT, and both sets of information are combined into one test statistic of the RC-TDT. More specifically, Knapp distinguishes five different types of families:

1. Families with both parents being genotyped; at least one parent is heterozygous.
2. Families with one parent being genotyped; the other can be reconstructed. For a diallelic marker, this means that the reconstructed parent has to be heterozygous.
3. Families where both parental genotypes are missing but can be reconstructed. Both parents therefore have to be heterozygous.
4. Families in which only one parent has been genotyped and the other cannot be reconstructed unambiguously. However, the conditions for performing the S-TDT are fulfilled.
5. All other families; these are discarded from the analysis.

To base the statistic on the number of A alleles in the affected children, y , for families of types 1 through 4, $\mathbb{E}_{H_0}(y)$ and $\mathbb{V}\text{ar}_{H_0}(y)$ have to be determined. For families of type 4, these were given in Eqns. (12.5) and (12.6); the derivation of the others is described in detail by Knapp [361]. The results of the derivations are presented in Table 12.12 for a diallelic marker.

Table 12.12 Mean values and variances under H_0 for the number of A alleles in the affected children in family i , y_i , depending on the family type. A diallelic marker with alleles A and a is considered.

	Family type	$\mathbb{E}(y_i)$	$\mathbb{V}\text{ar}(y_i)$
1	1 parent Aa	$\frac{1}{2}n_{\text{aff},i}$	$\frac{1}{4}n_{\text{aff},i}$
	2 parents Aa	$n_{\text{aff},i}$	$\frac{1}{2}n_{\text{aff},i}$
2	Typed parent AA	$\frac{3}{2}n_{\text{aff},i}$	$\frac{n_{\text{aff},i}}{4} \cdot \frac{2^{n_i-1} - n_{\text{aff},i}}{2^{n_i-1} - 1}$
	Typed parent Aa	$n_{\text{aff},i}$	$\frac{n_{\text{aff},i}}{2} + n_{\text{aff},i} \left(\frac{(5 - 2n_{\text{aff},i})3^{n_i-2} - 2^{n_i-1}}{4^{n_i} - 2 \cdot 3^{n_i} + 2^{n_i}} \right)$
	Typed parent aa	$\frac{1}{2}n_{\text{aff},i}$	$\frac{n_{\text{aff},i}}{4} \cdot \frac{2^{n_i-1} - n_{\text{aff},i}}{2^{n_i-1} - 1}$
3		$n_{\text{aff},i}$	$\frac{n_{\text{aff},i}}{2} + n_{\text{aff},i} \left(\frac{(5 - 2n_{\text{aff},i})3^{n_i-2} - 2^{n_i-1}}{4^{n_i} - 2 \cdot 3^{n_i} + 2^{n_i}} \right)$
4		See Eq. (12.5)	See Eq. (12.6)

Using these values, the test statistic for the RC-TDT can be calculated as the sum over all families i with

$$T_{\text{RC-TDT}} = \frac{\sum_{i=1}^n (y_i - \mathbb{E}(y_i))}{\sqrt{\sum_{i=1}^n \mathbb{V}\text{ar}(y_i)}}.$$

Under the null hypothesis of no linkage or no association, $T_{\text{RC-TDT}}$ is approximately standard normally distributed [313]. An example for computing the RC-TDT follows.

Example 12.4. For illustrative purposes, we assume that we have conducted a small study in which eight families were included that had at least 1 affected offspring. Members of these families were genotyped at one candidate SNP with alleles A and a , and we wish to test for linkage and association in these families (see Figure 12.5). Note that in the first two families, both the parents and the child were genotyped; hence these represent families of type 1. The following two families are of type 2 as in both cases, the father is missing. Families 5 and 6 are instances of type 3 because the missing father in the one case and both missing parents in the other can be reconstructed unambiguously from the offspring's genotype. Finally, the last two families belong to type 4. Here, missing genotypes cannot be reconstructed without doubt, but the S-TDT can be carried out. Table 12.13 shows the values of the number of A alleles in the affected, y , as well as $\mathbb{E}(y)$ and $\mathbb{V}\text{ar}(y)$ for each family.

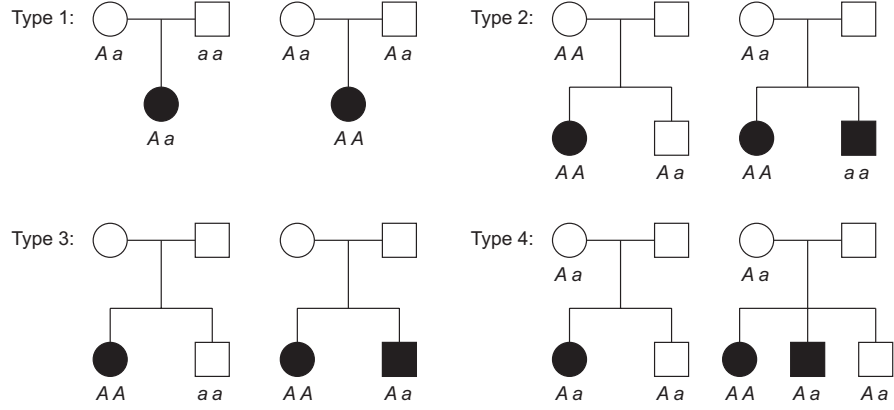


Fig. 12.5 Illustrative data for the computation of the reconstruction combined transmission disequilibrium test (RC-TDT). Eight families of four different types as described in Table 12.12 have been genotyped at one diallelic marker with alleles A and a .

Table 12.13 Number of A alleles Y in the affecteds, $\mathbb{E}_{H_0}(y)$, and $\mathbb{V}\text{ar}_{H_0}(y)$ for each family from Figure 12.5.

Family	y	$\mathbb{E}_{H_0}(y)$	$\mathbb{V}\text{ar}_{H_0}(y)$
1	1	$\frac{1}{2} \cdot 1 = 0.5$	$\frac{1}{4} \cdot 1 = 0.25$
2	2	1	$\frac{1}{2} \cdot 1 = 0.5$
3	2	$\frac{3}{2} \cdot 1 = 1.5$	$\frac{1}{4} \cdot \frac{2^{2-1}-1}{2^{2-1}-1} = 0.25$
4	2	2	$\frac{2}{2} + 2 \cdot \left(\frac{(5-2 \cdot 2) \cdot 3^{2-2} - 2^{2-1}}{4^2 - 2 \cdot 3^2 + 2^2} \right) = 0$
5	2	1	$\frac{1}{2} + 1 \cdot \left(\frac{(5-2 \cdot 1) \cdot 3^{2-2} - 2^{2-1}}{4^2 - 2 \cdot 3^2 + 2^2} \right) = 1$
6	2	2	$\frac{2}{2} + 2 \cdot \left(\frac{(5-2 \cdot 2) \cdot 3^{2-2} - 2^{2-1}}{4^2 - 2 \cdot 3^2 + 2^2} \right) = 0$
7	1	$(2 \cdot 0 + 2) \frac{1}{2} = 1$	$1 \cdot 1 \cdot \frac{4 \cdot 0 \cdot (2-0-2) + 2 \cdot (2-2)}{2^2 \cdot (2-1)} = 0$
8	3	$(2 \cdot 1 + 2) \frac{2}{3} \approx 2.6667$	$2 \cdot 1 \cdot \frac{4 \cdot 1 \cdot (3-1-2) + 2 \cdot (3-2)}{3^2(3-1)} \approx 0.2222$

From this, the test statistic is calculated to be

$$T_{RC-TDT} \approx \frac{(1 - 0.5) + (2 - 1) + \dots + (3 - 2.6667)}{\sqrt{0.25 + 0.5 + \dots + 0.2222}} \approx 2.2361,$$

with an associated two-sided p -value of approximately 0.02535.

In contrast to other statistics based on reconstructing parental genotypes, the RC-TDT does not rely on the assumption that the missing of parental data is independent

of the genotype through the combination of reconstruction and use of siblings. The only nuclear families that are neglected are those with an SAO and one parent. These could, however, be included by combining the RC-TDT and the 1-TDT. Although we have described the RC-TDT for a diallelic marker, it is directly applicable to multiallelic [361] and X-linked markers [312], and the software for the RC-TDT is available on the Web.

For the situation of nuclear families with at least one affected offspring, where parental genotypes may or may not be available, an alternative approach was recently suggested by Guo and colleagues [266]. In this *informative-transmission disequilibrium test (i-TDT)*, transmission information from heterozygous parents to all, affected and unaffected, offspring is utilized. It is a valid test for linkage and association, and the TDT and the SDT are special cases if only nuclear families with one affected child or only affected and unaffected offspring without parental genotypes are available, respectively.

12.7.3 How can extended pedigrees be analyzed?

An even further extension is to allow not only for sibships but also for arbitrary extended pedigrees including both affected and unaffected offspring. For this situation, Martin and colleagues [447, 448] developed the *pedigree disequilibrium test (PDT)*. This is a test utilizing data from related nuclear families and is a valid test for linkage and association even with population substructure. The test can be applied to informative pedigrees that include at least one informative trio with an SAO or one DSP consisting of one affected and one unaffected sibling.

Let us consider a diallelic marker with alleles A and a . The first step is to construct the following two random variables, one for each single trio with an SAO t within family i and one for each DSP t' within family i :

$$\begin{aligned} x_{\text{Trio},it} &= (\#A \text{ transmitted}) - (\#A \text{ non-transmitted}) \\ x_{\text{DSP},it'} &= (\#A \text{ in affected}) - (\#A \text{ in unaffected}) \end{aligned}$$

For non-informative trios or DSPs, these random variables are 0. In the second step, x_{Trio} and x_{DSP} are summed within each single pedigree:

$$D_i = \frac{1}{n_{\text{Trio},i} + n_{\text{DSP},i}} \left(\sum_{t=1}^{n_{\text{Trio},i}} x_{\text{Trio},it} + \sum_{t'=1}^{n_{\text{DSP},i}} x_{\text{DSP},it'} \right),$$

with $n_{\text{Trio},i}$ and $n_{\text{DSP},i}$ being the number of trios with an SAO and DSPs within pedigree i , respectively. The statistic is based on the sum over all pedigrees $D = \sum_{i=1}^n D_i$ having mean 0 under H_0 . The variance can be estimated by $\sum_{i=1}^n (D_i - \bar{D})^2$ with $\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i$, so that the Wald-type PDT finally is

$$T_{\text{PDT}} = \frac{\sum_{i=1}^n D_i}{\sqrt{\sum_{i=1}^n (D_i - \bar{D})^2}}, \quad (12.7)$$

which is asymptotically standard normally distributed under H_0 of no linkage or no association. The PDT of Eq. (12.7) has been termed *PDT-average* by Martin and colleagues [447] because it averages all families regardless of their size. Contributions are normed D_i through $n_{\text{Trio},i} + n_{\text{DSP},i}$. A slightly more powerful version of the PDT can be constructed by weighting families according to their size [447]. Before we illustrate the calculation of the PDT, we want to point out that the originally proposed PDT did not include the subtraction of $\bar{D}$ in the denominator [448] and thus was not a true Wald-type statistic.

Example 12.5. We first consider family 1 with members from three generations having been included in a study (see Figure 12.6). A disease of interest occurs in the third generation only, and all members have been genotyped at a diallelic marker with alleles A and a . To test for linkage and association using the PDT, the first step is to identify all possible trios with an SAO and all DSPs in this pedigree. There are three affected children in the last generation (children 1, 2, and 4), and each of these together with his or her parents forms one trio with an SAO, thus $n_{\text{Trio},1} = 3$. Basing the statistic on the allele A , for the first, the second, and the fourth child, values for $x_{\text{Trio},1t}$ are $x_{\text{Trio},11} = 1 - 1 = 0$, $x_{\text{Trio},12} = 2 - 0 = 2$, and $x_{\text{Trio},13} = 1 - 0 = 1$, respectively. In addition to these trios, there is one DSP in the family consisting of children 3 and 4 with contribution $x_{\text{DSP},11} = 1 - 0 = 1$. Next, we summarize these values within the pedigree to $D_1 = \frac{1}{3+1}(0 + 2 + 1 + 1) = 1$.

Now, we consider the second, smaller family. This family consists of one DSP with $x_{\text{DSP},21} = 2 - 0 = 2$ and $D_2 = \frac{1}{1} \cdot 2 = 2$. Then, $\bar{D} = \frac{1}{2} \cdot (1 + 2) = \frac{3}{2}$ for both families. From this, the statistic $T_{\text{PDT}} = \frac{1+2}{\sqrt{(1-\frac{3}{2})^2 + (2-\frac{3}{2})^2}} \approx 4.2426$ is calculated, rendering a two-sided p -value of < 0.0001 from the standard normal distribution. Again, we want to stress that the asymptotic normal distribution is used only for illustrating the calculation.

An alternative to the PDT is a very general testing procedure for extended families that is summarized under the term *family-based association tests* or *FBAT* for short. It was first proposed by Rabinowitz and Laird [552] and can be employed for testing a variety of different phenotypes, including binary, survival [395], time-to-onset [395], and quantitative traits (see Section 12.8), even in a multivariate fashion [396].

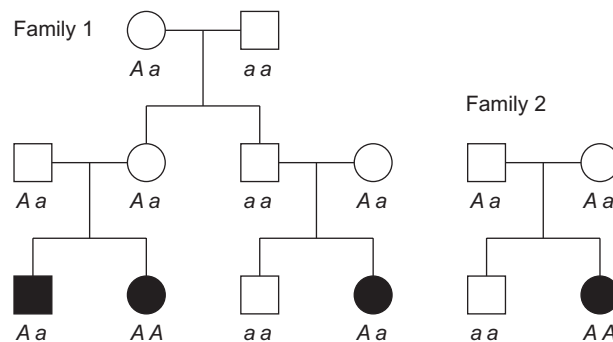


Fig. 12.6 Illustrative data for the computation of the pedigree disequilibrium test (PDT). One family with three generations (left) and one with two generations (right) have been genotyped at one diallelic marker with alleles A and a .

The general approach comprises two stages. In the first, a test statistic for the specific situation is described that tests for association and linkage between the genetic marker and the trait. Consider that n families are included in total, with each including n_i affected and/or unaffected children. The marker genotype and the phenotype are denoted by X_{it} and Y_{it} for the t th child within family i . Then, the general test statistic is of the Mantel statistic-type space–time clustering form (for an overview, see Ref. [50]):

$$S = \sum_{i=1}^n \sum_{t=1}^{n_i} X_{it} Y_{it} \quad (12.8)$$

X_{it} can be chosen in different ways, reflecting the assumed genetic model and the number of alleles at the genetic marker. One standard coding for a diallelic marker with alleles A and a is to let X_{it} be the number of A alleles. Alternatively, a dominant or a recessive model can be utilized by setting X_{it} to 1 if the genotype is Aa or AA in the dominant and AA in the recessive case. Codings for the multiallelic setting can be found in Ref. [313]. For Y_{it} , different functions of the phenotype might be used. For the binary phenotype affection versus non-affection that we consider here, $Y_{it} = 1$ for an affected child and 0 otherwise.

In the second stage of the FBAT approach, the null distribution of the marker data is calculated by treating the offspring genotype data as random and conditioning on the observed phenotype and parental genotype data. If parental data are complete, the distribution under H_0 may be obtained by conditioning on the observed phenotypes in all family members and the parental genotypes. In this case, the conditional distribution of the marker alleles in the children is derived from Mendelian transmission probabilities. If parental data are incomplete, we need to condition on the genotype configuration in the offspring as well. It has been shown that even partially observed parental genotypes and the genotype constellation in children are sufficient statistics for the missing parental genotypes. Hence, by conditioning on these for obtaining the null distribution, biases resulting from population stratification are avoided.

To derive the null distribution if parental data are missing, Rabinowitz and Laird presented algorithms based on the observed genotypes. For example, if both parental genotypes are missing, and the children's genotypes are AA and Aa , parental genotypes need to be either Aa, Aa or AA, Aa . Which of these alternatives is present then determines the conditional distribution of the offspring's genotypes. Therefore, the algorithm fixes the number of AA and Aa genotypes in the children and randomly permutes the genotypes among the offspring. In another case, the genotypes in the children might be AA and aa . Then, the genotypes of the parents are Aa, Aa , and the algorithm randomly assigns the sibs the genotypes AA, Aa , and aa with probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ while keeping only sibships where at least one AA and one aa genotype are present [387]. A detailed description can be found in Ref. [552].

Using the conditional distribution of the offspring genotypes under H_0 , the distribution of the test statistic may be obtained by Monte–Carlo simulations. Alternatively, the conditional distribution can be utilized to compute $\mathbb{E}_{H_0}(S)$ and $\mathbb{Var}_{H_0}(S)$. The derivation of these is described in a technical report accompanying the program FBAT. Using these, the resulting score statistic $T_{\text{FBAT}} = (S - \mathbb{E}_{H_0}(S)) / \sqrt{\mathbb{Var}_{H_0}(S)}$

is approximately standard normally distributed; a further discussion is found in Refs. [313, 388].

12.8 HOW CAN ASSOCIATION WITH QUANTITATIVE TRAITS IN FAMILIES BE ANALYZED?

In the previous sections, we assumed that a binary trait such as affection versus no affection was the focus of the study. However, many applications in complex traits concentrate instead on associations with quantitative traits. The advantages of studying these over dichotomous traits were considered in Chapter 9 in the context of linkage studies, and a recent discussion in the context of association studies is found in Ref. [203]. In this section, we will describe how association between a quantitative trait and a genetic marker can be analyzed in a family setting.

Several modifications to the original TDT for quantitative traits have been suggested. Some of these, for example, the elegant variance components (VC; see Chapter 9) approach by Fulker and colleagues [232], extended by Abecasis and others [1], assume a normally distributed trait. An alternative test was suggested by Rabinowitz [551] that makes no assumptions on the distribution of the quantitative trait. In essence, it is similar to an approach by Allison [12] but does not rely on a normal distribution of the trait.

The development of the test by Rabinowitz includes two steps. In the first, a statistic is defined for association between a specific marker allele and the trait. The second step modifies the statistic in order to remove possible spurious association by including parental information. If a diallelic marker is investigated and the focus is on allele A , indicator variables for the transmission of allele A are defined by

$$X_{it,m} = \begin{cases} 1, & \text{if } A \text{ is transmitted maternally to child } t \text{ in family } i; \\ 0, & \text{otherwise} \end{cases}$$

and

$$X_{it,p} = \begin{cases} 1, & \text{if } A \text{ is transmitted paternally to child } t \text{ in family } i; \\ 0, & \text{otherwise.} \end{cases}$$

With Y_{it} denoting the quantitative trait of child t in family i and $\bar{Y}$ being the average of the trait over all children in all families, the statistic for association is given by

$$\sum_{i=1}^n \sum_{t=1}^{n_i} (Y_{it} - \bar{Y})(X_{it,m} + X_{it,p}).$$

If there is no association between A and the trait, the indicator variables will be independent of the trait; hence, the expected value of the statistic will be 0 under H_0 , and association will be reflected by any deviation from 0.

In the second step, parental information is included in the statistic. In detail, those indicators will be removed where the parent is not heterozygous, and then $\frac{1}{2}$ is

subtracted from each remaining indicator. Letting $X_{i,m}^*$ and $X_{i,p}^*$ be 1 if parent i is heterozygous and 0 otherwise, the modified statistic is of the form

$$S_{\text{Rab}} = \sum_{i=1}^n \sum_{t=1}^{n_i} (Y_{it} - \bar{Y}) \left(X_{i,m}^* (X_{it,m} - \frac{1}{2}) + X_{i,p}^* (X_{it,p} - \frac{1}{2}) \right),$$

which reflects association only in the presence of linkage. Under the H_0 of no linkage or no association, $\mathbb{E}_{H_0}(S_{\text{Rab}}) = 0$, and the variance is given by

$$\text{Var}_{H_0}(S_{\text{Rab}}) = \frac{1}{4} \sum_{i=1}^n \sum_{t=1}^{n_i} (Y_{it} - \bar{Y})^2 (X_{i,m}^* + X_{i,p}^*).$$

$T_{\text{Rab}} = S_{\text{Rab}} / \sqrt{\text{Var}_{H_0}(S_{\text{Rab}})}$ subsequently is the Rabinowitz score test statistic, which is approximately standard normally distributed. An illustration of the calculation will be given below.

With the approach of selecting the maximal statistic from all possible alleles in a multiallelic situation, Rabinowitz shows how the procedure can be extended to multiallelic markers. In addition, possible environmental covariates can be included by first regressing the trait on the covariates and then utilizing the residuals for the T_{Rab} . The test is implemented as an option in the software package QTDT, which is available on the Web.

The same statistic as the one by Rabinowitz [551] can be obtained by using the general framework of FBAT. Equivalence can be obtained by defining the phenotype Y_{it} as $Y_{it} = (Y_{it} - \bar{Y})$ (see Ref. [387]). A concise comparison of these two approaches was recently given by Ewens and colleagues [196].

The above-mentioned methods are prospective approaches in nature, meaning that they model the outcome phenotype conditional on the genotypes [236]. Alternatives are retrospective approaches as proposed, e.g., in Refs. [357, 425]. Here, the genotypes of the study subjects are modeled conditional on their phenotypes and the parental genotypes. Detailed descriptions and comparisons of these approaches can be found in Ref. [714].

Example 12.6. As an illustration of the calculation of the Rabinowitz test, consider a study in which 20 trios were ascertained and genotyped at a diallelic marker with alleles A and a . For simplicity, we assume that five different genotype combinations as shown in Figure 12.7 were observed with four families each. In addition, a

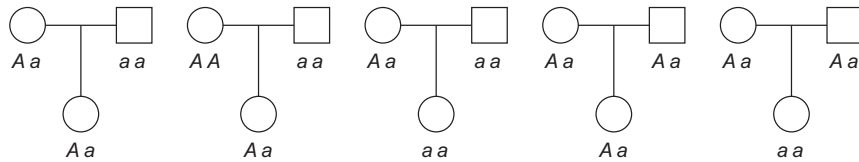


Fig. 12.7 Trio types ascertained for Example 12.6 illustrating the calculation of the Rabinowitz test. Twenty families, four of each type, were genotyped at a diallelic marker.

Table 12.14 Illustrative data to calculate the Rabinowitz test.

Family	Y	$X_{it,m}$	$X_{it,p}$	$X_{i,m}^*$	$X_{i,p}^*$	Family	Y	$X_{it,m}$	$X_{it,p}$	$X_{i,m}^*$	$X_{i,p}^*$
1	26	1	0	1	0	11	61	0	0	1	0
2	22	1	0	1	0	12	17	0	0	1	0
3	21	1	0	1	0	13	84	0	1	1	1
4	29	1	0	1	0	14	82	0	1	1	1
5	18	1	0	0	0	15	45	0	1	1	1
6	72	1	0	0	0	16	91	0	1	1	1
7	21	1	0	0	0	17	99	0	0	1	1
8	72	1	0	0	0	18	51	0	0	1	1
9	18	0	0	1	0	19	1	0	0	1	1
10	80	0	0	1	0	20	38	0	0	1	1

quantitative trait has been recorded in the offspring, and the values of Y are shown in Table 12.14.

The average trait value is calculated to be $\bar{Y} = 47.4$. Also, the indicator variables denoting the transmission of the A allele paternally $X_{it,p}$ and maternally $X_{it,m}$ are shown, as are $X_{i,m}^*$ and $X_{i,p}^*$ denoting whether the parent is heterozygous. For instance, in the first four families, A is transmitted maternally but not paternally, so that $X_{1,m} = 1$ and $X_{1,p} = 0$. Further, the mother is heterozygous and the father is homozygous, yielding $X_{1,m}^* = 1$ and $X_{1,p}^* = 0$. From these values, it is straightforward to compute $S_{Rab} = -38.4$, $\text{Var}_{H_0}(S_{Rab}) = 5973.86$, and $T_{Rab} \approx -0.4968$, rendering a two-sided p -value of approximately 0.6193.

12.9 PROBLEMS

Problem 12.1 As an example of the analysis of trio data, consider the data shown in Figure 12.8. Here, a sample is available consisting of 10 trios with both parents and an SAO.

Problem 12.1.1. Test for association and linkage using the TDT. In addition, estimate the GRRs γ_1 and γ_2 as well as the PAR and corresponding asymptotic 95% confidence intervals under the assumption of HWE.

Problem 12.1.2. Imagine that all paternal genotypes are missing in Figure 12.8. Test for linkage and association using an appropriate test.

Problem 12.2 Show the equality of Eq. (2.2) and Eq. (12.3) by using the Bayes' theorem.

Problem 12.3 Confirm Eq. (12.5) and Eq. (12.6) by assuming a hypergeometric distribution for Y_1 and Y_2 .

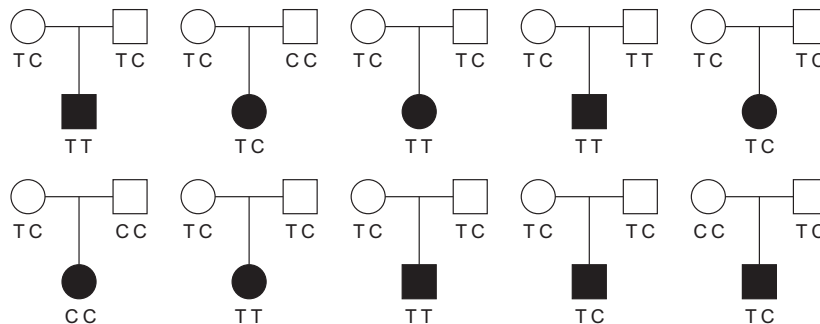


Fig. 12.8 Ten trios with both parents and a single affected offspring genotyped at a diallelic marker for Problem 12.1.

URLs

ETDT

<http://www.smd.qmul.ac.uk/statgen/dcurtis/software.html>

FBAT INCLUDING SDT

<http://www.biostat.harvard.edu/fbat/default.html>

FBAT, TECHNICAL REPORT

<http://www.biostat.harvard.edu/fbat/fbattechreport.ps>

RC-TDT

<http://www.uni-bonn.de/umt70e/soft.htm>

rTDT

<http://www.ugr.es/mabad/rTDT/rTDT.html>

S.A.G.E.

<http://darwin.cwru.edu>

SIBASSOC

<http://www.mds.qmul.ac.uk/statgen/dcurtis/software.html>

TDT AND S-TDT

<http://genomics.med.upenn.edu/spielman/TDT.htm>

TRANSMIT

<http://www-gene.cimr.cam.ac.uk/clayton/software/>

13

Haplotypes in Association Analyses

In conducting a genetic epidemiological study, one will very rarely restrict the genotyping to a single genetic marker, even if there is a strong candidate gene. In the context of linkage studies, it was already observed that it might be advantageous to utilize the information from several markers simultaneously (see Section 8.6). Consequently, the obvious question is how can information from multiple genetic markers be used reasonably in association analyses? In this chapter, the focus will be on *haploid genotypes* or *haplotypes*, meaning a specific set of alleles at linked genetic loci. As will be seen, a combination of alleles at a number of loci that are linked or in linkage disequilibrium (LD) can have a different impact on the phenotype under study depending on whether the alleles are on the same chromosome, hence in *cis* position, or on different ones, thus in *trans* position.

In Section 13.1, we describe the reasons for studying haplotypes. Here and in other sections, only single nucleotide polymorphisms (SNPs) will be considered as markers. Because in most instances the haplotypes are not directly observed but only the genotypes, we will depict methods to reconstruct haplotypes or estimate haplotype frequencies (Section 13.2). Because in practice haplotypes are usually analyzed in data from unrelated individuals, Section 13.3 is devoted to these testing approaches. In contrast, we will not discuss haplotype association analysis in families in this book, and the interested reader may refer to the literature (see, e.g., Refs. [129, 362, 739, 743]).

In addition to testing for association, an indirect use of haplotypes lies in their utilization for selecting markers for a study (Section 13.4). We want to stress that the role of haplotypes in association is a rapidly evolving topic. Hence, many approaches will only be sketched to give the basic idea.

13.1 WHY ARE HAPLOTYPES INTERESTING?

The analysis of haplotypes in addition to single markers has gained immense interest over the past decade. As an indication of this, the journal *Genetic Epidemiology* dedicated a special issue to the topic *genetic epidemiology and haplotypes* in 2004 [581]. But where does this stem from? Several reasons can be identified, most of them having been nicely summarized by Clark [124].

1. Haplotypes are biologically relevant. There is strong evidence that if there are several mutations within the same gene in *cis* position, these might interact to cause changes in the amino acid chain, thus yielding different protein products. Characteristics of the protein such as folding structure and stability might depend on exactly these interactions. Consequently, we can think of a number of alleles in *cis* position creating a kind of superallele that in turn has a more pronounced effect on the phenotype. Because in recent years it has become more feasible to genotype markers in close vicinity to each other, thus more likely to interact, these possibilities are acknowledged to a greater extent. To give an example, Fitze and colleagues investigated influences of and interactions between mutations in the *RET* proto-oncogene and polymorphisms on the development of Hirschsprung's disease [216]. In patients with a more severe and a milder form, they found that the c135G/A polymorphism, or sequence variations in LD with this marker, modulated the phenotypic influences of *RET* germline mutations. Because the c135A variant was associated with the milder form when it was in *cis* with the mutation, it seems that the c135A allele has a protective effect, and the *RET* haplotype seems to influence disease severity. However, it needs to be pointed out that the observation of a haplotype with a specific phenotype does not necessarily substantiate a true biological function of the investigated haplotype. Consider, for instance, a haplotype consisting of alleles *A* and *B* that may be more strongly associated with disease than either allele *A* or *B* alone. This might be due to the fact that the true disease-causing mutation is neither *A* nor *B* but some mutation *C* that is located between the other alleles. Hence, if a person has haplotype *AB*, he may also be more likely to carry *C* and thus the disease mutation [128].
2. The variation in populations is inherently structured into haplotypes. At some point in time, the disease-causing mutation was introduced into the population, and while the haplotype structure is a mirror of the population history of genetic drift, recombination, mutation, selection, and migration, individuals carrying the mutation tend to share more ancestry than those without the mutation.
3. Both simulation and analytical studies have shown that haplotype analyses have a greater power than those based on single markers [7], especially when the causative variant is not investigated directly or when there are multiple disease-causing alleles or even interactions. However, if there is just one single or a small number of alleles that cause the disease, testing single markers may be more successful [39, 477].

4. In Chapter 7, we saw that the informativity of phased haplotypes may be greater than that of unphased single markers. Similarly, the inherently lower informativity of a diallelic SNP (Section 3.2) can be increased by combining several SNPs and thus yielding multiallelic markers.
5. Knowledge of haplotypes might guide the sensible selection of markers for large-scale association studies. This will be discussed in Section 13.4.

13.2 HOW CAN WE INFER HAPLOTYPES?

When genetic markers are genotyped, the resulting information usually is unphased, meaning that it is not known which pairs of alleles are lying on the same chromosome. As we saw in Chapter 10, a haplotype is unambiguous only if there is no more than one heterozygous site. If there is not complete LD among the markers, the probability for a haplotype to be ambiguous increases with the number of investigated loci and as the allele frequencies approach 0.5. In principle, it is possible to obtain phase information, thus haplotypes, in the genotyping process by specific molecular methods such as allele-specific polymerase chain reaction (PCR) or cloning. However, these molecular haplotyping approaches are still costly and time-consuming and thus rarely applied in practice. If only the unphased genotype information is given, it may be possible in families to unambiguously assign phases over generations. However, in smaller pedigrees and especially in unrelated individuals, this is usually not possible. Hence, algorithms are required that infer haplotypes from genotype data. For family data, this was described in Chapter 7, and here we will focus on the derivation of haplotypes from unrelated individuals.

Haplotype information can be obtained by two fundamentally different approaches. One group of algorithms focuses on the individual and assigns pairs of haplotypes or haplotype probabilities (see Section 13.2.1); the other looks at the population and estimates haplotype frequencies (see Section 13.2.2). After assigning haplotype probabilities for each individual or estimating haplotype frequencies in the population, the seemingly most straightforward way for further analyses would be simply to use the most probable haplotype combination for each individual and treat it as if it had been observed. However, any ambiguity associated with that reconstruction is then neglected, which falsely decreases the variance and is a potential source of bias. Further, if there were other haplotype combinations just slightly less likely for an individual, this potentially important information is discarded. Hence, it is instead preferable to utilize haplotype probabilities in the individual or estimated haplotype frequencies in the population to test for group differences (see Section 13.3).

13.2.1 Which algorithms assign haplotypes or haplotype probabilities to individuals?

One of the earliest algorithms for haplotype reconstruction was suggested by Clark [123] and explicitly aims at assigning each individual a pair of haplotypes. Considering a sample of unrelated individuals whose genotypes have been determined, it works as follows.

Algorithm 13.1.

1. Identify all unambiguous haplotypes, that is, all homozygotes and single-site heterozygotes. Consider the haplotypes of these individuals *resolved*.
2. Determine whether any of the resolved haplotypes could be one of the haplotypes in a still unresolved individual. If not, STOP. Otherwise, continue.
3. Each time a resolved haplotype is identified as one of the possible alleles in an unresolved individual, identify the complement and consider this newly identified homologue resolved. Go back to step 2.
4. Continue until all haplotypes have been resolved or until no more new haplotypes can be found.

An example for the algorithm is given in Problem 13.1, and the Fortran implementation HAPINFERX of Clark's algorithm is available.

The immediate advantages of this algorithm are that it is relatively straightforward and works well for a large number of loci with a limited diversity. However, the disadvantages are that the algorithm does not start if there are no unambiguous haplotypes in the sample and that there might be unresolved haplotypes when the algorithm is stopped. Furthermore, the solution may not be unique because a different order of the individuals may yield different results. Hence, Clark recommended repeating the algorithm with different orderings of the data and found that the solution that resolves most of the haplotypes with the smallest number of haplotypes usually is valid, which is why the algorithm is based on the principle of maximum parsimony. Another problem with Clark's algorithm may be that its performance is relatively sensitive to deviations from the Hardy–Weinberg equilibrium (HWE), although this assumption does not enter the algorithm explicitly.

A number of Bayesian algorithms that aim at phasing the haplotypes, thus assigning each individual a pair of haplotypes, have been proposed in the literature. All these algorithms treat the unknown haplotypes as random quantities and utilize both prior information and a term of the likelihood to determine the posterior probability of the unknown haplotypes given the known genotypes. By selecting the most likely haplotype for each individual, haplotype assignment is carried out. One of the Bayesian algorithms that has gained currency is the approach by Stephens and colleagues [634]. Here, assumptions based on the approximate coalescent are made for the prior information. Specifically, it is surmised that an unresolved haplotype will differ only slightly from the progenitor haplotype, often by only a single base. This coalescence assumption distinguishes this approach from all other procedures.

To describe the Stephens and colleagues' algorithm rather intuitively, start with an initial guess for haplotypes. Then, for each iteration, one individual is randomly chosen, and new haplotypes for this individual are assigned under the assumption that all other haplotypes are correctly specified. A central issue is that for this step, the conditional distribution of a haplotype based on the genotype and other haplotype distribution has to be specified. As an approximation for this distribution, the above-mentioned prior information based on coalescence comes into play. In-depth information on this algorithm can be found in the work by Stephens and colleagues [631, 632, 634]. This general idea was first formulated in 2001 [634]. In the following years, modifications and extensions adopting a similar structure were continuously suggested [503, 633].

Further algorithms have been suggested in the past few years. Some of them explicitly utilize the structure of LD or the block-like structure of haplotypes in the human genome as described in Section 10.2.3. Because this is a fast emerging research field, we refrain from describing the numerous approaches in detail; instead, the reader may refer to reviews [502, 633]. A comprehensive overview over the currently available methods including software references was given by Liu and colleagues [424]. Because a number of different approaches to infer haplotypes are in use, the question of which approach to utilize remains open. Results from the Genetic Analysis Workshop 14 [50] indicate that the Bayesian coalescence-based method is preferable to other approaches. The quality of all procedures strongly depends on the specifics of the data under study, especially the LD between the genetic markers and the haplotype diversity.

13.2.2 Which algorithms estimate haplotype probabilities in the population?

The easiest approach for estimating haplotype probabilities can be employed if haplotypes have been determined on the molecular level. Then, they can be estimated by counting. This method can also be used if haplotypes have been assigned to or haplotype probabilities have been estimated for all individuals in the population.

An alternative is to directly estimate the haplotype probabilities of the population from the unphased data. Probably the most common approach is to utilize an EM algorithm. Several variants have been suggested [197, 288, 428], with the algorithm by Excoffier and Slatkin [197] being applied most frequently.

The basic idea is that if the phases were known, the frequencies could be directly estimated from them, and if the frequencies were known, the phase could be deduced. Hence, in the E step of the algorithm, the expected phase is calculated, and in the M step, the maximum likelihood estimates of the frequencies are determined.

Assume that a sample of n unrelated individuals has been genotyped and that r different multilocus genotypes $g_1, \dots, g_r$ have been observed having frequencies $n_{g_1}, \dots, n_{g_r}$. A multilocus genotype of an individual is the unphased joint information from all marker loci of interest. These r multilocus genotypes are compatible with s haplotypes. One variant of the EM algorithm is given by Algorithm 13.2.

Algorithm 13.2.

1. Set convergence criterion $\varepsilon < 10^{-v}$ for precision of haplotype frequency estimates. Typically, $4 \leq v \leq 8$.
2. Determine frequency of multilocus genotypes $n_{g_1}, \dots, n_{g_r}$.
3. Set maximum number of iterations N . Typically, $25 \leq N \leq 100$.
4. Set starting log likelihood for convergence check, e.g., $L^{(-1)} = 1$.
5. Initialize iteration $k = 0$.
6. Set starting values for the haplotype frequencies $f_{h_i}^{(0)}$, $i = 1, \dots, s$. Typically, $f_{h_i}^{(0)} = 1/s$.
7. While $k \leq N$,
 - (a) Assume HWE and calculate probabilities for all possible haplotype combinations by $\hat{P}^{(k)}(h_i, h_i) = f_{h_i}^{(k)} f_{h_i}^{(k)}$, for $i = j$ $\hat{P}^{(k)}(h_i, h_j) = 2 f_{h_i}^{(k)} f_{h_j}^{(k)}$, for $i \neq j$.
 - (b) Estimate multilocus genotype probabilities $\hat{P}^{(k)}(g_l)$ for $l = 1, \dots, r$ from $\hat{P}^{(k)}(h_i, h_j)$, where $i = 1, \dots, s$ and $j \geq i$ is the sum of all haplotype combinations (h_i, h_j) that are compatible with the multilocus genotype g_l .
 - (c) Calculate the log likelihood $\ln L^{(k)} = \sum_{l=1}^r n_{g_l} \ln \hat{P}^{(k)}(g_l)$.
 - (d) If $|L^{(k-1)} - L^{(k)}| < \varepsilon$, then JUMP to 9.
 - (e) Estimate the expected number of haplotype i for $i = 1, \dots, s$ by $\hat{E}^{(k)}(n_{h_i}) = \sum_{l=1}^r n_{g_l} \hat{P}^{(k)}(h_i, h_j | g_l, g_l \text{ compatible with } (h_i, h_j))$.
 Here $\hat{P}^{(k)}(h_i, h_j | g_l, g_l \text{ compatible with } (h_i, h_j)) = \frac{2 f_{h_i}^{(k)} f_{h_j}^{(k)}}{\hat{P}^{(k)}(g_l)}$
 for both $h_i \neq h_j$ and $h_i = h_j$.
 - (f) $k = k + 1$.
 - (g) Update haplotype frequencies: $f_{h_i}^{(k+1)} = \hat{E}^{(k)}(n_{h_i}) / (2n)$.
8. STOP with no convergence.
9. STOP with convergence. Haplotype frequency estimates are $f_{h_i}^{(k)}$.

Because the description of the algorithm is quite technical, we will give an example below. The advantages of this algorithm are that it is fairly robust against deviations from HWE and it is based on solid statistical theory. However, it is sensitive to initial values and might yield local maxima only. It is therefore advisable to rerun the algorithm using different start values. Furthermore, the EM algorithm cannot handle a large number of loci. An alternative is the stochastic-EM algorithm proposed by Trégouët et al. [667], which avoids the convergence to local maxima.

For diallelic markers such as SNPs, simulations have shown that the absolute error in frequency estimates is very low and that estimates are better if there are some very common haplotypes and many very rare haplotypes in the population [200]. With less LD between the markers, the estimation quality deteriorates [50]. A problem might be that very rare haplotypes from the original population are not detected in the sample, although this might be due to sampling error rather than haplotype inference. Further simulation studies have shown that the EM algorithm, if modified to explicitly

handle missing data, is surprisingly accurate even if a high proportion of genotype information is missing [350]. However, genotyping errors strongly decrease the accuracy of frequency estimates [356].

There are a number of implementations of the algorithm available in more comprehensive packages. For example, the R package HAPLO.STAT includes the program HAPLO.EM for EM estimation of haplotype frequencies. Another implementation specifically designed for estimating frequencies of large haplotypes within SNPs is SNPHAP.

Example 13.1. Consider three SNPs genotyped in a sample of 10 unrelated individuals. The first SNP has alleles G and C , the second, A and C , and the third, A and G . In the sample, three different multilocus genotypes have been observed. Three individuals carried genotype $g_1 = (GG, CC, AA)$, three genotype $g_2 = (GG, AC, AA)$, and the last four genotype $g_3 = (GG, AC, AG)$. In principle, from three SNPs there are eight possible haplotypes, but in this sample only four haplotypes are compatible with the data:

$$h_1 = \begin{pmatrix} G \\ C \\ A \end{pmatrix}, \quad h_2 = \begin{pmatrix} G \\ A \\ A \end{pmatrix}, \quad h_3 = \begin{pmatrix} G \\ A \\ G \end{pmatrix}, \quad \text{and} \quad h_4 = \begin{pmatrix} G \\ C \\ G \end{pmatrix}$$

For the first two genotypes, there is only one possible haplotype combination. Specifically, g_1 has haplotype combination (h_1, h_1) and g_2 has (h_1, h_2) . However, the third genotype g_3 is compatible with the two haplotype combinations (h_1, h_3) and (h_2, h_4) .

In the initialization step, we assign every compatible haplotype the same probability, so that $f_{h_1}^{(0)} = f_{h_2}^{(0)} = f_{h_3}^{(0)} = f_{h_4}^{(0)} = 0.25$. From this, we compute for each genotype the probability assuming HWE. Thus,

$$\begin{aligned} \hat{P}^{(0)}(g_1) &= \hat{P}^{(0)}(h_1, h_1) = f_{h_1}^{(0)} f_{h_1}^{(0)} = 0.25^2 = 0.0625, \\ \hat{P}^{(0)}(g_2) &= \hat{P}^{(0)}(h_1, h_2) = 2f_{h_1}^{(0)} f_{h_2}^{(0)} = 2 \cdot 0.25 \cdot 0.25 = 0.125, \text{ and} \\ \hat{P}^{(0)}(g_3) &= 2f_{h_1}^{(0)} f_{h_3}^{(0)} + 2f_{h_2}^{(0)} f_{h_4}^{(0)} = 2 \cdot 0.25 \cdot 0.25 + 2 \cdot 0.25 \cdot 0.25 = 0.25. \end{aligned}$$

These probabilities are then utilized in the second step to calculate the expected numbers of each haplotype. For this, remember that the first haplotype may occur in all three multilocus genotypes and the second haplotype in the second or third multilocus genotype, but haplotypes 3 and 4 are compatible only with the third multilocus genotype. This gives

$$\begin{aligned} \hat{E}^{(0)}(n_{h_1}) &= 3 \cdot \frac{2 \cdot 0.25 \cdot 0.25}{0.0625} + 3 \cdot \frac{2 \cdot 0.25 \cdot 0.25}{0.125} + 4 \cdot \frac{2 \cdot 0.25 \cdot 0.25}{0.25} = 11, \\ \hat{E}^{(0)}(n_{h_2}) &= 3 \cdot \frac{2 \cdot 0.25 \cdot 0.25}{0.125} + 4 \cdot \frac{2 \cdot 0.25 \cdot 0.25}{0.25} = 5, \\ \hat{E}^{(0)}(n_{h_3}) &= 4 \cdot \frac{2 \cdot 0.25 \cdot 0.25}{0.25} = 2, \text{ and } \hat{E}^{(0)}(n_{h_4}) = 4 \cdot \frac{2 \cdot 0.25 \cdot 0.25}{0.25} = 2. \end{aligned}$$

Table 13.1 Results of the first five iterations in the expectation maximization algorithm in Example 13.1.

Iteration k	$f_{h1}^{(k)}$	$f_{h2}^{(k)}$	$f_{h3}^{(k)}$	$f_{h4}^{(k)}$	$\ln L^{(k)}$
1	0.5500	0.2500	0.1000	0.1000	-14.7903
2	0.5875	0.2125	0.1375	0.0625	-14.0365
3	0.6218	0.1782	0.1718	0.0282	-13.3618
4	0.6410	0.1590	0.1910	0.0090	-13.0214
5	0.6477	0.1523	0.1977	0.0023	-12.9132

In the third step, we utilize the expected haplotype counts to update the haplotype frequencies to $f_{h1}^{(1)} = \frac{11}{2 \cdot 10} = 0.55$, $f_{h2}^{(1)} = \frac{5}{2 \cdot 10} = 0.25$, and $f_{h3}^{(1)} = f_{h4}^{(1)} = \frac{2}{2 \cdot 10} = 0.1$. The genotype probabilities are then updated as

$$\begin{aligned}\hat{P}^{(1)}(g_1) &= 0.55^2 = 0.3025, & \hat{P}^{(1)}(g_2) &= 2 \cdot 0.55 \cdot 0.25 = 0.275, \\ \hat{P}^{(1)}(g_3) &= 2 \cdot 0.55 \cdot 0.1 + 2 \cdot 0.1 \cdot 0.25 = 0.16.\end{aligned}$$

The log likelihood based on these updated probabilities is given by $\ln L = 3 \cdot \ln(0.3025) + 3 \cdot \ln(0.275) + 4 \cdot \ln(0.16) \approx -14.7903$. These steps are repeated until convergence. For illustration, Table 13.1 shows the results for the first five iterations.

13.3 HOW CAN WE TEST FOR ASSOCIATION USING HAPLOTYPES?

The available approaches to test for differences in haplotype frequencies have been nicely reviewed in depth by Schaid [580] and by Liu et al. [424]. One early procedure was to utilize an EM algorithm, principally as described above. As before, the haplotypes are treated as missing data within the algorithm, and the frequencies are estimated so that the log likelihood is maximized. As a modification, the algorithm is run in the affected cases, in the unaffected controls, and in the pooled sample. Using the different likelihoods from the three runs, a likelihood ratio test statistic can be constructed by

$$LRT = 2 \cdot (\ln L_{\text{cases}} + \ln L_{\text{controls}} - \ln L_{\text{pool}}),$$

which is, under H_0 of equal haplotype frequencies in cases and controls, asymptotically χ^2 distributed with the degrees of freedom (d.f.) given by the number of haplotypes compatible with the data minus 1. A limitation of this approach, however, is that the approximation might not be adequate if there are many haplotypes. Also, adjustments for other covariates such as environmental factors are not possible. It is restricted to a categorical phenotype, and a specific haplotype cannot be tested against the others. As the EM algorithm itself, it is based on the assumption of HWE.

A more general alternative is to model the effect of haplotypes on the phenotype in a generalized linear model (GLM) [580, 584]. The technical level of this method

is beyond the scope of this book, and we therefore sketch only the fundamental idea. With this very flexible approach, different types of phenotypes can be analyzed, including binary affection status, quantitative traits, or even survival phenotypes. It also allows the inclusion of other variables such as environmental covariates and interactions between environmental effects and haplotypes [389].

Because haplotype phases have to be estimated in practice for a sample of unrelated individuals, the GLM needs to be modified in order to include the ambiguity of the haplotype assignments. Hence, the probabilities of the individual haplotype combinations need to be included in the model. Very generally, the contribution of an individual to the likelihood is given by $L = \sum_{h \in g} P(y|x_e, x_g(h), \beta)P(h)$, where y denotes the phenotype, g the observed genotypes, h the haplotypes, x_e environmental covariates, $x_g(h)$ genetic covariates, β the vector of regression coefficients, and $P(h)$ the prior probability for the haplotype. The conditional probability of the phenotype given the covariates $P(y|x_e, x_g(h), \beta)$ can then be modeled using a GLM.

Various methods have been described that are based on this general approach, and these are summarized by Schaid [580] in Table II. For application, the regression model is fitted in the first step. Here, it is necessary to maximize the log likelihood over the regression parameters, for which unphased haplotypes can be treated as missing data within the EM framework. In a second step, the GLM framework is used to construct score statistics for hypothesis testing. Advantages of these statistics are that they are simple to compute and that it is feasible to calculate p -values by Monte-Carlo permutations. Both the regression and the score statistics have been implemented in the R package HAPLO.STAT, and an example of the application is given below.

One major disadvantage of this approach is that if the haplotype phase is unknown, it depends on the estimation of haplotype frequencies. For this, the EM algorithm is applied within the pooled sample. In case-control studies, this leads to bias because cases are over-represented compared to their population prevalence [639]. The extent of the bias depends on how well the haplotypes are predicted from the genotypes and is hence usually less severe with higher LD. As a possible solution to this problem, introduction of sampling weights that mirror the actual disease prevalence has been suggested [639]. However, this might not always be known. A different solution has been proposed by Epstein and Satten [189, 578]. They model the probability of the observed genotype given the phenotype by summing over all compatible pairs of haplotype combinations. A major disadvantage of this is that it is still not possible to include covariates in the analysis. Furthermore, it yields biased results if there is deviation from HWE in the controls.

One problematic issue in all testing approaches is that for haplotypes comprising more loci, a large number of haplotypes will be possible, and many of them will be rare. This problem is further aggravated if interactions between haplotypes and covariates are tested. Consequently, both frequency estimates and parameter estimates will have large variances, leading to unstable models. How should one deal with rare haplotypes? One possibility is to include them in the baseline category, another to pool them together into one group. An often suggested frequency threshold for this is to define haplotypes as rare if their estimated frequency is below 1–2%. However,

both ways yield results that may be difficult to interpret. Also, it excludes the possibility that a rare haplotype might be of specific interest. A more appealing approach, although it requires more computational time, is to shrink the effects of each single rare haplotype toward the common mean or toward the effect of a similar haplotype. This is described in more detail by Schaid [580].

For applications, a number of software solutions have been made available. A very comprehensive collection has been published by Salem and colleagues [575], and we agree with the authors that no program is ideal for all situations. Instead, the researcher should carefully scrutinize the main features, which are given in the review and are also available in the web.

The approaches discussed so far somewhat represent the opposite of the association analyses that were the focus of Chapter 11. There only single markers were investigated, possibly in interaction, whereas here only haplotypes are investigated, possibly in interaction with other covariates. Another way is the simultaneous analysis of single markers and haplotypes [116, 141].

Example 13.2. *In a study on the effect of variants in the human gene coding surfactant protein D (SFTPD) on serum collectin surfactant protein D (SP-D), 6 SNPs were genotyped in a sample of 290 healthy blood donors, and the quantitative phenotype SP-D level was determined [292]. Applying the EM algorithm as implemented in HAPLO.STAT, nine haplotypes with frequencies greater than 1% were identified, as shown in Table 13.2.*

Table 13.2 Estimation of haplotype frequencies in the gene coding surfactant protein D and association with collectin surfactant protein D.

Haplo ^a	1 ^b	2	3	4	5	6	Freq ^c	Score	<i>p</i> ^d
1	A	G	T	A	G	G	0.1352	-7.1568	< 0.0001
2	G	G	T	G	G	G	0.0478	-1.5547	0.1254
3	A	A	T	A	G	G	0.0686	-1.4096	0.1630
4	A	G	T	G	G	G	0.0150	-0.0033	0.9974
5	G	A	C	A	A	G	0.0251	0.3022	0.7651
6	A	A	C	A	A	G	0.3593	0.5739	0.5685
7	A	G	T	G	G	A	0.0471	1.4250	0.1639
8	G	G	T	G	G	A	0.1809	1.9542	0.0530
9	A	A	T	G	G	G	0.0694	2.3559	0.0213

^a Haplotype. ^b 1 = 759A>G, 2 = 92A>G, 3 = 4694C>T, 4 = 11208A>G, 5 = 11384A>G, 6 = 11466G>A. ^c Estimated haplotype frequency. ^d Monte-Carlo *p*-value from 10⁶ replications.

*Modeling SP-D as an ordinal trait, the global score statistic was $U \approx 71.9397$ with 9 d.f. The scores associated with each haplotype are shown in Table 13.2 with their respective *p*-values. Specifically, haplotype 1 is associated with a lower SP-D and haplotype 9 with a higher SP-D.*

13.4 HOW CAN HAPLOTYPES AND LINKAGE DISEQUILIBRIUM STRUCTURE BE UTILIZED TO SELECT MARKERS FOR STUDY?

In Section 10.2.3, we explained that the concept of indirect association largely hinges upon existing LD between the investigated marker and the truly disease-causing variant. If the actual extent of LD is largely unknown, however, it means that in practice, a dense coverage of the region to investigate is required. On the other hand, it was observed that for many genomic regions, LD showed a blocky pattern with some markers being in tight LD, interspersed with recombinational hot spots. This gave rise to the idea that the structure of LD and haplotypes might be utilized to reduce the number of markers to be typed in a specific study.

An example of this is shown in Figure 13.1, where 9 SNPs with alleles *G* and *A* were genotyped. We assume that for these SNPs, it has been shown that they are structured in two distinct haplotype blocks as described in Section 10.2.3. Focusing on the first block of three SNPs, it can be seen that the first two SNPs are sufficient to distinguish all three haplotypes, and the last one is redundant. For the second block, the first four SNPs are required for a unique constellation of alleles identifying each of the five haplotypes, rendering the last two SNPs superfluous.

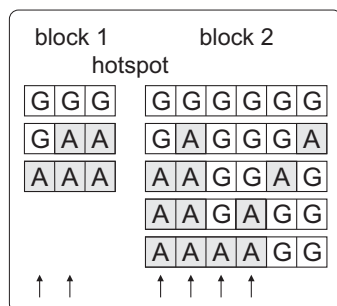


Fig. 13.1 Nine fictional single nucleotide polymorphisms (SNPs) with alleles *G* and *A* each structured in two distinct haplotype blocks. In the first block, the first two SNPs are sufficient to distinguish all three haplotypes. In the second block, the first four SNPs are required for a unique constellation of alleles identifying each of the 5 haplotypes. Arrows indicate the relevant SNPs. This figure has been adapted from Ref. [673].

Abstracting from this example, a rationale might be sought to select relevant markers. In principle, this selection could be based on the aim of finding direct associations with causative variants as described in Section 10.1.1. For this, the selection would focus on the sequence information and preferably identify functional SNPs in coding regions, splice junctions, and promoter regions. Especially, SNPs leading to non-conservative amino acid changes are more likely to be functional. The main advantage of such an approach is that in all, fewer SNPs are required, with an estimated 50,000 to 100,000 for a genome scan. The major problem, however, is that functional SNPs not identified might be missed.

Another approach for selecting SNPs is to aim for indirect associations and utilize LD and haplotype information [277]. The assumed scenario for this is to first choose a region of interest for the analysis. Then, the study population is defined, and a preliminary analysis of all possible SNPs in the region is performed on a representative sample of individuals from that population. Based on this, a subset of SNPs is identified. For the underlying causal model of indirect association, this requires

that the SNPs be selected such that the probability is maximized for substantial LD between the unknown causative mutation and at least one of the genotyped markers.

The important difference from the direct association approach is that the location and the type of the SNP itself are irrelevant and that functional SNPs not identified in the previous approach might be found. However, it can be expected that more SNPs have to be genotyped, probably on the order of 500,000 to 1,000,000, and the LD structure across the genome has to be known for the relevant population. To facilitate SNP selection without having to type all possible SNPs for every study, the international *HapMap project* was performed, which genotyped millions of SNPs in four major populations to describe the pattern of LD (see the web site of the project for more details).

13.4.1 What are methods to select markers based on haplotypes or linkage disequilibrium?

Basically, the selection of SNPs for indirect association can be based on one of three strategies (for overviews, see Refs. [277, 695]):

1. Selection of *haplotype-tagging SNPs* or *htSNPs* that retain haplotype diversity: Because this requires that haplotypes have been reconstructed or that haplotype frequencies have been estimated, this is only rarely applied in practice and therefore not described in depth.
2. Selection of *tagging SNPs* or *tagSNPs*: This avoids the explicit phasing and utilizes information of pairwise LD between all SNPs instead (see below for details).
3. Construction of maps of LD units as described in Section 5.3 [438].

In the following, we will focus on the family of methods for marker selection which calls on the structure of the pairwise LD between all SNPs in question. The underlying rationale here is that if the SNPs that are not selected are in high LD with those that are eventually genotyped, the underlying disease-causing mutation will be as well. The aim therefore is to identify a subset of SNPs that best predicts any other SNP within the considered block. Hence, the approach of indirect association might be successful.

One popular procedure was suggested by Carlson and colleagues [106]. Using the pairwise squared correlation Δ^2 (see Section 10.2), SNPs are sorted into bins so that the pairwise LD within a bin is sufficiently high. In the following, we explain the algorithm in detail.

Algorithm 13.3.

1. Set a binning criterion $\Delta^2 = c$. Common choices in the literature are $0.5 \leq \Delta^2 \leq 0.8$, with $\Delta^2 = 0.8$ being the most popular choice.
2. Compute all pairwise Δ^2 between all SNPs.
3. For each SNP, count the number l of other SNPs to which $\Delta^2 \geq c$.
4. If the maximal $l = 0$, STOP. Otherwise, identify the SNP with the maximal l . Group this SNP with all associated SNPs into one bin.

5. Select any SNP within this bin as a tagSNP if for all other SNPs within the bin $\Delta^2 \geq c$.
6. For SNPs as yet unbinned, repeat steps 3 to 5.

In applications, SNPs with low minor allele frequency are filtered and thus not considered as tagSNPs. An advantage in using Δ^2 over other measures of LD is the direct connection with power. If $\Delta^2 = 1$, clearly no information is lost, and with lower values, the loss in power might be compensated with a greater sample size. In fact, it has been shown that the sample size needs to be increased by a factor of $1/\Delta^2$. Hence, with a threshold of 0.8, for instance, the sample size needs to be increased by 25% in order to retain the power. The binning algorithm is implemented in the software LDSELECT, and an illustration is given with Example 13.3.

Although using Δ^2 for SNP selection is conceptually related to power, another metric associated with case-control studies is the selection of SNPs directly based on power. Hu and colleagues [317], for instance, suggested two algorithms, one that would maximize the average power of the tagSNPs to detect all SNPs in the original set and the other that would maximize the minimum power of the tagSNPs to detect each individual SNP in the original set. One problem with this procedure, though, is that a specific genetic model needs to be defined a priori. Specifically, assumptions have to be made about the location and the allele frequency of the underlying disease mutation, which will prove difficult in practice.

In a different vein of approaches, the LD structure is analyzed using principal components analysis (PCA) and the selection of SNPs is based on the resulting factor structure [310, 461, 508]. In the most elaborate of these works [310], a two-stage procedure has been proposed that aims at maximizing the proportion of genetic variation captured using a minimal number of SNPs. In the first stage, a PCA is performed on the correlation between the SNPs to determine LD groups by extracting factors with eigenvalues of at least 0.7. These factors then undergo an oblique rotation to allow some degree of interrelatedness between the LD groups. The number of extracted factors, thus the number of relevant LD groups, is determined using the cumulative explained variance, with the aim of accounting for at least 90% of the variance. All SNPs with a factor loading of at least 0.4 are assigned to a factor.

After obtaining thus the relevant LD groups, the second step selects group tagging SNPs. For this, separate PCAs are performed for every group. Here, the number of extracted factors indicates the number of required SNPs, and the relevant SNPs are those with the highest factor loading. Again, an eigenvalue of > 0.7 and a proportion of explained variance of 90% are the criteria for factor and SNP selection, respectively. The SNPs with the highest factor loadings are selected as tagSNPs. Computations can be carried out with standard statistical software. An advantage of this approach is that the LD groups need not be on contiguous chromosomal fragments. Also, as opposed to most other procedures, it gives a rationale for the optimal number of SNPs. In the end, this section gives an illustration for this approach.

A host of other algorithms for SNP selection have been proposed over the last few years [124, 277, 637], and both the ideas and the quality of the resulting selection are similar.

Example 13.3. In a study on the role of the interleukin-21 receptor (IL21R) and the interleukin-4 receptor α genes in inflammation, 12 SNPs in the IL21R and the IL4R gene were genotyped in 300 healthy blood donors [290]. For the binning algorithm by Carlson and colleagues [106], the pairwise LD was estimated using Δ^2 . Applying a threshold of $\Delta^2 \geq 0.5$, it was found that only three pairs, SNP 4 with SNP 5, SNP 7 with SNP 8, and SNP 11 with SNP 12, exceeded the threshold. Consequently, these were grouped into one bin each, with the remaining six SNPs staying unbinned. From the bins, one of the two SNPs may be selected as a tagSNP.

For the PCA approach [310], we subsequently performed an oblique factor rotation. Although only five factors showed eigenvalues ≥ 0.7 , six factors were extracted in order to explain more than 90% of the variance. After rotation, factor loadings of the SNPs were obtained as shown in Table 13.3.

Table 13.3 Loadings of 6 extracted factors resulting from a principal components analysis and oblique factor rotation for the 12 single nucleotide polymorphisms (SNPs) from Hecker and colleagues [290].

SNP	Factor					
	1	2	3	4	5	6
1	0.8864	0.1123	-0.0529	0.0660	0.4346	0.1820
2	0.9186	0.1667	-0.0149	0.0970	0.4432	0.1750
3	0.9600	0.0886	0.0075	0.0579	-0.0861	0.2169
4	0.9600	0.0886	0.0075	0.0579	-0.0861	0.2169
5	0.0989	0.1238	-0.1243	0.0014	0.9738	-0.0492
6	0.1656	0.8834	0.0981	0.2398	0.2052	-0.0059
7	0.2514	0.8513	-0.0115	0.0795	0.1792	-0.0690
8	-0.1123	0.8488	-0.0243	0.0913	-0.1038	-0.2147
9	0.0606	0.1468	0.0006	0.9980	0.0030	0.2995
10	0.1764	-0.1177	0.0309	0.2989	-0.0303	0.9980
11	-0.0102	0.0169	0.9949	-0.0024	-0.0877	0.0329
12	-0.0222	0.0207	0.9951	0.0083	-0.0918	0.0290

From these values, SNPs with factor loadings ≥ 0.4 were grouped together, so that factor 1 contained SNPs 1, 2, 3, and 4; factor 2 SNPs 6, 7, and 8; factor 3 SNPs 11 and 12; factor 4 SNP 9; factor 5 SNPs 1, 2, and 5; and factor 6 SNP 10.

In the second step, separate PCAs were performed for each factor to determine the tagSNPs. Again extracting a number of factors so that the proportion of explained variance was $\geq 90\%$, the number of factors in the groups, thus the number of tagSNPs per group, were two, two, one, two, and one, respectively. The SNPs with the highest factor loadings in a group were then selected as tagSNPs, yielding a set comprising SNPs 1 and 3 for the first group, SNPs 6 and 8 for the second, SNP 11 for the third, SNP 9 for the fourth, SNPs 2 and 5 for the fifth, and SNP 10 for the last group.

13.4.2 How good is marker selection based on haplotypes or linkage disequilibrium?

Although the idea has gained popularity, there are important critical points to be noted concerning the procedures for selecting SNPs based on LD or haplotypes. Some theoretical caveats have nicely been summarized by Halldorsson et al. [277]. Recall that SNP selection is part of a process in which, first, a preliminary set of SNPs is typed in a representative sample of the intended population. In this way, the tagSNPs or htSNPs can never be better suited for association studies than the preliminary set, and there will always be a loss in power after SNP selection. Hence, the density that adequately mirrors the variation in the population of the preliminary set is crucial.

Second, it may be questioned whether the sample in which SNPs are selected is truly representative of the latter population in practical situation. If every single study includes a step for SNP selection based on the intended population, this may be achievable. Using tagging SNPs selected in other contexts, however, may be problematic. Comparisons have shown that there may be substantial differences even between European populations, rendering the transferability questionable [485].

The third limitation lies in the specific composition of the preliminary SNPs. Here, it has been shown that there might be a considerable ascertainment bias in the initial SNP set, especially if the SNPs are not selected in a balanced way between the analyzed populations [8]. Most SNPs that are available in databases, thus amenable for selection, have been validated only in small samples. Also, it should be added that there is a bias in databases toward commoner SNPs, which will probably not be representative of the mutations that may cause a disease. This latter point is all the more problematic because most tagging approaches are restricted to the selection of commoner SNPs. Hence, it is very unlikely that these procedures will be able to detect rare disease-causing variants [124]. Termed differently, the success of tagging SNPs hinges upon the validity of the common disease—common variant hypothesis (Section 2.3.5). However, because this debate is not finished yet, the impact of this limitation is still unknown.

In addition to these theoretical concerns, empirical evaluations have been carried out in both simulated and real data sets. These have yielded surprisingly consistent results [50, 219, 328, 349, 737, 741]:

- Even though the different approaches select different SNPs, a comparable proportion of information is captured.
- The proportion of hidden SNPs that are detected is low for rare variants. Also, the power to detect association with a rare allele decreases faster with selecting tagging SNPs than with a random selection of SNPs.
- The number of selected SNPs depends on the number of SNPs investigated in the original set.
- The number of selected SNPs with a high minor allele frequency is almost independent of the sample size. A sample size of 50 individuals is sufficient for efficient selection of tagging SNPs with common alleles. However, if SNPs

with small minor allele frequencies are to be selected, results depend on the sample size.

- Even though the tagging SNPs might capture most of the variation of the SNPs in the preliminary set, this need not be the case for hidden SNPs. Specifically, if the density of SNPs in the original set is not sufficient, it is difficult to capture the variance of hidden SNPs. For example, htSNPs based on a marker set with a distance of 10 kilobases (Kb) captures only 84% of the variation in the hidden SNPs that was indicated by the htSNP analysis.
- The proportion of captured information and the accuracy of constructed haplotypes is generally higher if there is a higher LD in the region.

Although the quality of SNP selection is a topic of dispute, tagging SNPs form the basis of recent study approaches, which are large-scale, genome-wide association studies. This will be described in the following chapter.

13.5 PROBLEMS

Problem 13.1 Consider that a sample of unrelated probands has been genotyped at three SNPs with alleles G and A, C and A, and A and G, respectively. The overall size of the sample is of no concern, but 6 different genotype combinations, as shown in Figure 13.2, have been observed. Reconstruct the haplotypes using Clark's algorithm.

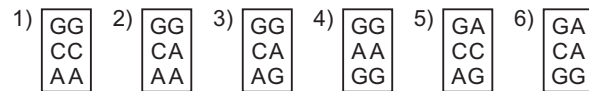


Fig. 13.2 Exemplary genotype data for Clark's algorithm. Three single nucleotide polymorphisms have been genotyped in 6 unrelated individuals.

Problem 13.2 In the study by Berghöfer and colleagues [56], 5 common SNPs in the toll-like receptor 9 (*TLR9*) gene were genotyped in 527 unrelated healthy blood donors to investigate possible associations of this candidate gene with atopic reactions. LD between these SNPs A1923C, T1486C, T1237C, G1174A, and G2848A was estimated, and the pairwise correlation Δ is shown in Table 13.4.

Problem 13.2.1. Based on the binning algorithm by Carlson and colleagues [106], which SNPs should be selected for further study at a threshold of $\Delta^2 = 0.5$?

Problem 13.2.2. Compared to typing all SNPs, by how many orders of magnitude would the sample size need to be increased for the same power?

Table 13.4 Pairwise Δ between five single nucleotide polymorphisms (SNPs) in the toll-like receptor 9 gene from Berghöfer and colleagues [56].

	T1486C	T1237C	G1174A	G2848A
A1923C	0.2931	0.1005	0.2177	0.2192
T1486C		0.3446	0.7419	0.7480
T1237C			0.3665	0.3533
G1174A				0.9929

URLs

HAPINFERX

<http://www.nslj-genetics.org/soft/hapinferx.f>

HAPLO.STAT FOR R

http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm

THE HAPMAP PROJECT

<http://hapmap.ncbi.nlm.nih.gov/>

LDSELECT

<http://droog.gs.washington.edu/LdSelect.html>

PHASE AND FASTPHASE

<http://www.stat.washington.edu/stephens/software.html>

SNPHAP

<http://www-gene.cimr.cam.ac.uk/clayton/software/>

REVIEW ON HAPLOTYPING SOFTWARE

<http://polymorphism.scripps.edu/HapSoftwareReview/>

14

Genome-wide Association Studies

Devoting an entire chapter to one specific approach of study inevitably raises the question: What makes these studies so special? According to the National Institutes of Health (NIH), a *genome-wide association (GWA) study* is a study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits. Furthermore, the density of genetic markers and the extent of linkage disequilibrium (LD) should be sufficient to capture large proportion of common variation, and the number of samples should provide sufficient power.

Therefore, the first specific issue is pure numbers: Before GWA studies were conducted, a few single nucleotide polymorphisms (SNPs) were selected for association studies, about 500 to 1000 microsatellite markers were used for genome-wide linkage studies, and in the order of maybe 50,000 transcripts in expression studies. At most, hundreds of probands were included. For GWA studies, on the other hand, more than 1,000,000 SNPs are possibly genotyped and compared in usually thousands of probands.

Second, in the last three chapters, we dealt with association analyses in a rather special setting. Specifically, it was mostly assumed that candidate genes or regions were known in which association was then tested. This, of course, presupposes that there are candidates that are plausible because of biological function or previous study results, with some of them having strong genetic effects. In contrast to that, GWA studies systematically investigate SNPs in the entire human genome without an a priori specified hypothesis on the location. Thus, GWA studies rely on the concept of indirect association, which was described in Section 10.1. Consequently, we do not necessarily suppose that any of the SNPs is causal but rather that they may be in the vicinity of functional DNA variants and therefore associated with the disease

because of LD. The functional variant might not even be an SNP, but an inversion, an insertion, a deletion or a copy number variation (CNV) [767].

Finally, what kind of genetic effects are detected? In linkage studies, we were able to investigate rare diseases in families caused by rare variants. In contrast, a GWA study can be successful only if the common disease – common variant assumption holds (see Section 2.3) that the common diseases we are investigating are mostly caused by a small to moderate number of disease loci having common variants. This is a direct consequence of the technology that makes the typing of rare variants difficult.

Even though the conductance of GWA studies was proposed more than 10 years ago [565], it became feasible only a few years ago. For this, the important advancements were that microarray-based genotyping systems have become available at an affordable cost (Chapter 3), that information from the HapMap project (see Section 13.4) can be utilized to select SNPs adequately capturing differing levels of variation within individual genes, and finally, that large enough case-control or family-based samples or cohorts have been assembled for an adequate power of this approach.

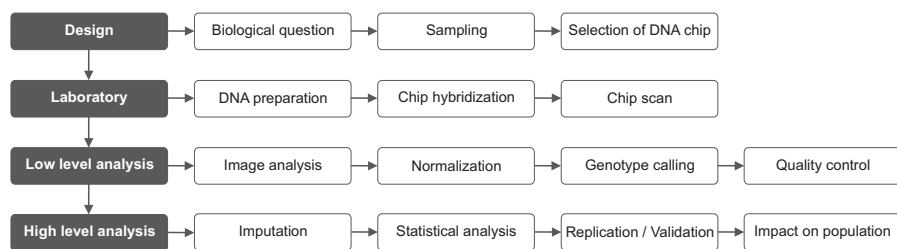


Fig. 14.1 Typical workflow of a genome-wide association (GWA) study. Reproduced from Ref. [760] with kind permission from Wiley-Blackwell.

Although we will not describe every single step in great detail here, the typical workflow of a GWA study is depicted in Figure 14.1. As in every epidemiological study, possibilities for the design need to be considered in the first step, and specific options for GWA studies are described in the first section. Steps that take place in the laboratory have been the topic of Chapter 3. As part of the general quality control procedures, most of the issues for evaluating a GWA study's quality were included in Chapter 4. The specific challenge of dealing with population stratification was part of Section 11.4. The high level analysis then may or may not begin with the imputation of missing genotype data (Section 14.2). We then describe a number of points on the statistical analysis itself (Section 14.3). There will be a focus on how to handle the multiple testing problem (Section 14.4) and how, starting with GWA results, evidence from different data sources may be accumulated (Section 14.5). This will include multistage designs, replication of GWA results, and meta-analyses. Furthermore, we broach the issue of the impact of results on the population (Section 14.6) before finally giving an outlook to the near future (Section 14.7).

14.1 WHAT DESIGN OPTIONS ARE THERE?

As in other genetic association studies, the first decision that has to be made in a GWA study pertains to the subjects who are included in the study. Here, the same options exist that have already been described in Section 10.1. Because most of the published studies have used a classical case-control design, our further focus will be on that.

A design issue that is specific to GWA studies is the selection of the DNA chip. Generally, a commercial fixed chip has two major advantages over a self-designed chip. First, it is more cost-efficient and, second, it has greater potential for a combination across studies (Section 14.5.3). Specific strategies can be used to create a chip suitable for GWA studies [45]:

- Tag strategy: Using information on genome-wide LD, a set of tagSNPs is chosen so as to maximize the amount of variation captured per SNP (see Section 13.4). An example is the Human660W-Quad chip by Illumina.
- Random strategy: A set of SNPs is chosen that is distributed approximately randomly across the genome. The Genome-Wide Human SNP Array 5.0 by Affymetrix is constructed this way.
- Combined strategy: Random SNPs are used as a basis, and the data are augmented by a fill-in set.
- Functional strategy: Using all known non-synonymous SNPs, a panel is constructed, possibly enriched by SNPs in genes in general.

The second criterion to be considered is the number of SNPs on a chip or the coverage of the genome it represents. This is usually evaluated in comparison with the data from the HapMap project. Classically, the genomic coverage is estimated by the percentage of common SNPs being in high LD, say $r^2 \geq 0.8$ (Section 10.2), with at least one SNP on the platform. For Caucasian populations, the Human1M (Illumina) has the greatest coverage of 93%, followed by the HumanHap550 (Illumina) and the Genome-Wide Human SNP Array 6.0 (Affymetrix) with 87% and 83%, respectively [408]. Since the selection of the SNPs is usually based on an older version of the HapMap than the one used for evaluating the chip, the coverage often is lower than would be theoretically expected. Furthermore, rare variants are only badly covered.

Coverage is, however, not the only measure to compare chips. Primarily, we are interested in the power to detect a functionally relevant variant that is not always highest with greatest coverage [624]. Furthermore, high quality of raw genotypes is important because high-quality chips require fewer efforts in data cleaning. For example, both the duplicate error fraction and the Mendelian error fraction were lower for the Illumina 1M BeadChip than for the Affymetrix Genome-Wide SNP Array 6.0 in the study of Ref. [293].

14.2 HOW CAN WE IMPUTE MISSING DATA?

We have seen above that the coverage of current genotyping platforms is far from perfect. To deduce the genotype information at untyped SNPs, one might make use of the available SNPs and the LD structure within the genome. Indeed, the LD between SNPs allows to capture much of the variation, even if these were not genotyped. This *imputation* of missing genotype data has a number of important applications in GWA studies [93]:

1. For any SNP, a number of probands will not have valid genotypes after quality control. These could be filled by using the information from the surrounding SNPs and the individuals who were successfully genotyped at the respective SNP.
2. Although being represented on the platform, a number of SNPs fail the quality control altogether. In this case, these can be imputed.
3. When comparing results from studies that have used different platforms, one might impute genotypes at SNPs that were not on the respective platform. This is especially useful if a meta-analysis is to be conducted.

It is important to note that for the second and third application, external information describing the LD pattern across the genome is utilized. In most applications, data from the HapMap project is used as a reference, although more complete data from the *1000 Genomes Project* will soon be available (for details on this project, see web site). Thus, in order for this to be a valid approach [279], we need to assume that the reference samples within HapMap, the cases and the controls are all sampled from the same population. Through this, the three populations share the same LD structure for every set of SNPs. This is a critical assumption because it ignores the hypothesis that controls are more related than cases close to a disease locus, which is the underlying rationale behind haplotype sharing statistics (see, e.g., Ref. [655]). If this assumption is violated, statistical tests are still valid. The estimation of effect sizes is, however, biased. An additional assumption for the first application is that the genotypes are missing completely at random meaning that there is no dependency of the missingness mechanism and the phenotype of interest.

14.2.1 What are common algorithms for imputation?

For genotype imputation, most methods that have been suggested for haplotype phase inference (see Section 13.2) could also be used. A first alternative for application 1 is to use standard statistical techniques for imputation, such as linear regression with variable selection or regression trees. The commonly used alternative is to utilize specialized algorithms that are able to deal with any of the three situations mentioned above.

One commonly used imputation algorithm is IMPUTE [442] that uses the following two fundamental ideas to estimate SNP-specific genotype probabilities. First, each individual's L locus genotype g_i is estimated independent of the genotypes of other individuals. Second, given a set of known reference haplotypes h , e.g., the hap-

lotypes provided in the HapMap project, each individual's genotype vector $P(g_i|h)$ is a Hidden Markov Model in which the hidden states are a sequence of pairs of the K known haplotypes in the set h . A subset of haplotypes is selected as reference set, and each reference haplotype represents a hidden state. The true but unobserved haplotypes are assumed to be imperfect mosaics of the reference haplotypes. Recombinations can be modeled as changes between reference haplotypes, and mutations are included as differences between observed and underlying true alleles. To model the transition probabilities between SNPs, a fine-scale recombination map is utilized [490].

Theoretically, information from all genotyped SNPs on the same chromosome can be utilized in this algorithm. However, since the influence of a specific SNP decreases with increasing distance, some window is defined within which the genotyped SNPs are used. Based on this model, the probability of each possible genotype for every missing genotype in the study can be obtained. A major difference between IMPUTE and most comparable methods is that both the reference panel and the sample data are used, so that more information is available. IMPUTE can also utilize information from different reference panels [316], including unphased panels in addition to phased panels.

An alternative algorithm MACH [410] is based on the same model and operates similarly. A major difference lies in the fact that recombination and mutation rates are estimated within the algorithm. They are not specified by the user. Another difference is that, in its current form, MACH cannot handle more difficult data situations, including a number of different reference panels.

In addition to these two algorithms, a large variety of different methods have been proposed and implemented for imputation, and the interested reader may refer to the excellent reviews in Refs. [93, 279, 524].

14.2.2 How good are imputations of missing genotypes?

The imputation of missing genotypes explicitly or implicitly relies on the accuracy of the haplotype estimates. If the estimated imputation quality is low, the imputed data should be discarded altogether. As a measure for imputation quality, the ratio of the empirically observed variance of allele dosage to the theoretically expected variance is used. If the imputation has led to adequate information in predicting the unobserved genotypes from the observed haplotypic background, the ratio should be close to 1. As the observed variance diminishes, so does the ratio. This metric is given, e.g., as `rsqr_hat` from MACH or as `info` from PLINK.

For the remaining imputed SNPs, the uncertainty of imputation should be taken into account. To this end, likelihood-based tests can be used for imputed and ungenotyped variants that integrate over uncertain haplotype phase and missing data values [419]. Furthermore, most imputation methods provide posterior probabilities for imputed genotypes, which allows to take into account uncertainty without taking the full likelihood approach. In any case, when imputed SNPs show signs of association, the genotypes need to be confirmed by separate genotyping of the respective SNPs in the imputed sample.

Given the wealth of methods available, the obvious question arises which algorithm to use in practice. Comparisons including IMPUTE [442], MACH [410], BEAGLE [94], PLINK [548], and fastPHASE [587] found that although the concordance between observed and imputed genotypes was comparably high, the proportion of successfully imputed SNPs differed strongly, with PLINK performing substantially worse than MACH and IMPUTE [506, 524].

A number of factors have been shown to affect the quality of imputation. The first is the minor allele frequency of the SNP to be imputed, with rarer SNPs being easier to impute. Second, stronger LD leads to better imputations, although it seems that only a moderate LD of $\mathcal{D}' \approx 0.7$ is required [748]. Third, the more SNPs are missing, the lower the imputation accuracy is [322].

The final factor is the adequacy of the reference panel for the investigated sample, which usually stems from the HapMap project. In the above comparisons, the samples were drawn from a population similar to one in the HapMap project. In contrast, the central limitation of the HapMap is that only a very limited number of ethnicities are included, and these are represented by only a very small sample size. Specifically, there are 30 trios with northern and western European ancestry from the USA (CEU panel), 30 trios of Yoruban individuals from Ibadan, Nigeria, 45 unrelated individuals from Tokyo, Japan, and 45 unrelated Han Chinese individuals from Beijing. Thus, low frequency alleles will be extremely difficult to cover [93]. Also, untyped SNPs in individuals from other ethnicities are more difficult to impute [748]. A useful recommendation has been that when trying to impute genotype information on other populations, a mixed reference population with different proportions of the HapMap data sets should be used [322].

14.3 HOW DO WE EVALUATE THE DATA STATISTICALLY?

Once the cleaned genotype and phenotype data are available, the typical first-round statistical association analysis of a GWA study is rather straightforward. The standard approach here is to test every single SNP separately for association using one or a few of the tests described in Section 11.2. The main differences in comparison to the analysis of a candidate region are that first, the computational burden is higher, and second, it is not possible to manually scrutinize every single SNP result.

Instead, to derive an impression of the results, these are often displayed in the so-called confetti or Manhattan plot (see Figure 14.2 for an example from the study by Samani and colleagues [576]). For this, p -values from any association test could be utilized in principle, although the Cochran–Armitage trend test is the most common. In the example plot, it can easily be seen that whereas the majority of p -values are lower than 10^{-3} , there is a major peak on chromosome 9 in the Wellcome Trust Case Control Consortium (WTCCC) study followed by a couple of smaller ones with p -values less than 10^{-4} , which is mirrored in the results from the German Myocardial Infarction Family Study.

From this wealth of results, the obvious question is how to identify the SNPs or regions that merit further attention. It is clear that simply applying a classical

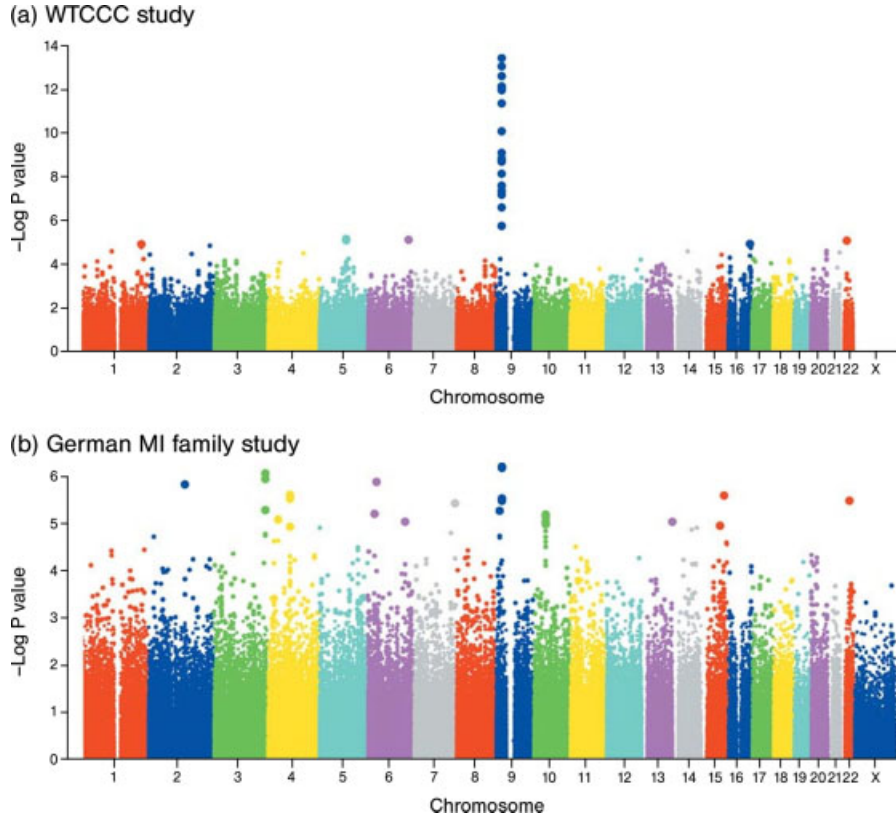


Fig. 14.2 Confetti or Manhattan plot to display results from the Wellcome Trust Case Control Consortium (WTCCC) study (a) and the German Myocardial Infarction (MI) family study (b), two genome-wide association studies. Shown are the $-\log p$ -values of all single nucleotide polymorphisms (SNPs) sorted by position. The fairly long flat bars between p -values of 10^{-3} and 10^{-2} represent the skyscrapers, and the low p -values represent the antennas. Figure is taken from Ref. [576]. Copyright 2007 Massachusetts Medical Society.

threshold for statistical significance of 0.05 is not helpful but would yield many false positive results. Thus, adjustment for multiple testing is an important issue and will be the topic of the following Section 14.4.

Independent of the designation of statistical significance, a usual practice is that SNPs are ranked in the first stage of a GWA study and that some cutoff is used to select SNPs that are then taken forward for confirmation in the next stage. Alternatively, a fixed number of best SNPs is selected. The ranking may be done based on different criteria such as p -values, test statistics, power for replication, or the false positive report probability (FPRP, see Section 14.4). For the situation of ranking by test statistics or p -values, Zaykin and Zhivotovsky [735] investigated the probability that, within a pre-specified number of most significant SNPs, the true positive effects

would be included. For example, if the power to detect an individual true effect is 85% and when 100,000 SNPs are tested in total, they found that the probability to have three true effects among the 500 most significant ones is only 15%. Adjusting for multiple testing improved the situation only marginally and increased the probability to 20%. More robust results may be obtained when ranking the SNPs not according to a single p -value but to the minimum p -value across the different test statistics for the dominant, recessive, and additive genetic model [751].

14.4 HOW CAN WE DEAL WITH THE PROBLEM OF MULTIPLE TESTING?

As we have pointed out, GWA studies differ from other studies regarding the pure numbers involved, concerning, e.g., the number of tests that are eventually performed. Thus, the question of how to adjust for multiple testing is most pertinent. In a very practical approach, the WTCCC [708] decided to use a point-wise significance threshold of $5 \cdot 10^{-7}$ simply because they estimated that this would give them a power of 80% to detect effects for common SNPs with a relative risk of at least 1.5. More formally, the easiest approach would certainly be to assume that every SNP is tested independent of all others. Then, classical procedures for multiple test adjustments to control the *family-wise error fraction*, which erroneously is termed *family-wise error rate (FWER)* can be used. It is the probability of committing at least one error. Hence, with α_p as the point-wise error frequency, M the number of tests performed and α_g the global error fraction that one aims to control, the *Bonferroni correction* yields $\alpha_l = \alpha_g/M$ as local significance level. Applying the Bonferroni correction to the analysis of 1,000,000 SNPs then leads to an appropriate local significance level of $5 \cdot 10^{-8}$ for a global significance level of 5%.

However, assuming independence between SNPs is inappropriate, so extreme test statistics are less likely, making the procedure overly conservative. It therefore becomes interesting to take into account the actual LD between the SNPs. In this line, Dudbridge and Gusnanto [174] used a permutation approach based on the data from the WTCCC. Extrapolating to genotyping at infinite density, they estimated the genome-wide significance threshold to be at $7.2 \cdot 10^{-8}$. Although feasible, permutation is computationally too intensive to be applied for every GWA study separately, so that a number of approximations have been suggested. Instead of comparing the p -value of every SNP with the local significance level, one may use approaches that

Table 14.1 Proposed local significance levels α_l for GWA studies in European populations to obtain a global level of 5%.

Approach	α_l
Bonferroni	$5 \cdot 10^{-8}$
Dudbridge and Gusnanto [174]	$7.2 \cdot 10^{-8}$
Pe'er et al. [523]	10^{-7}
Gao et al. [235]	$1.3 \cdot 10^{-7}$
Gao et al. [235]	$2.4 \cdot 10^{-7}$
Duggal et al. [175]	$3 \cdot 10^{-7}$
WTCCC [708]	$5 \cdot 10^{-7}$

estimate p -values adjusted for multiple testing for a specific study. In Section 14.4.1, we discuss a permutation-based approach that can be used for region-wide adjustments only because of its computational complexity. In Section 14.4.2, we sketch a neat approach to simulate p -values on the genome-wide level. In contrast to the above approach, other procedures specifically tackle the multiple testing problem in GWA studies without any simulation or permutation, and these are described in Section 14.4.3. An overview of different suggested local levels is given in Table 14.1.

With all of these different suggestions advocating strict significance levels, inevitably the question of power arises. For example, with typical sample sizes in GWA studies such as approximately 1000 cases and 1000 controls, using a point-wise significance level of 10^{-6} and assuming a risk allele frequency of 0.2, the power is approximately 4% to detect an odds ratio (OR) of 1.3. Therefore, it has been suggested to use other error concepts instead of the FWER. Some authors have suggested that the *false discovery rate* (FDR) might be more appropriate in the context of GWA studies (for a definition of the FDR and an excellent overview on error concepts, see Ref. [87]). However, the FDR is virtually identical to the FWER in GWA studies because a large proportion of tested null hypotheses is true. We therefore refrain from describing FDR based approaches here in further detail.

Yet another way of interpreting the statistical relevance of the data is to use a Bayesian approach as suggested by Wacholder et al. [680]. This method tries to answer the question: Given that the null hypothesis of no association is rejected, what is the probability that this is not correct? This leads to the definition of the *FPRP*, for which the grounds of classical statistical testing are abandoned. For details, the interested reader may refer to Refs. [680, 683] and applications in Refs. [576, 708].

Finally, based on the suggestions by Lander and Kruglyak [393], Duggal et al. [175] proposed to use different levels to claim suggestive, significant, and highly significant evidence for association, which, for the same reasons as given in Chapter 7, we view with caution.

14.4.1 How can region-wide p -values be simulated?

The approach discussed in this section is not well suited to adjust for multiple testing on the genome-wide level. However, in replication stages of GWA studies, regions, not only single SNPs, merit further attention, and to adjust for multiple testing in a specific region, this approach by Ge et al. [239] may be used. It has been described in the context of haplotype analysis in Ref. [49].

Given that we are testing SNPs in a confined region, we can simplify this procedure to the testing of single SNPs where the distribution of the test statistic is known (Algorithm 14.1):

Algorithm 14.1.

1. Compute test statistic T^m and p -values p^m for every SNP using the original n data
2. Derive the minimal p -value over all SNPs $p^{min} = \min\{p^m\}$

3. Set counter $C = 0$
4. For a specified number N of permutations
 - (a) Permute the case-control status and arrange the data into n new subjects
 - (b) Compute test statistics $T_{(j)}^m$ and p -values $p_{(j)}^m$ for every SNP using the j th permuted data
 - (c) Derive the minimal p -value across all SNPs $p_{(j)}^{min} = \min\{p_{(j)}^m\}$
 - (d) If $p_{(j)}^{min}$ is smaller than p^{min} , add 1 to C
5. Compute the global p -value as $p_g = C/N$

Example 14.1. To illustrate the approach, let us consider a region containing merely three SNPs as shown in Table 14.2. The first row shows the test statistics in the real data along with the corresponding p -values. Here, the minimal p -value is given by 0.0102. Although, for applications, the number of permutations clearly needs to be greater (see above), we display in the bottom lines the results from 10 permutations of the case-control status and the resulting minimal p -values in the right-most columns. For replicates 3 and 10, the minimal p -value is less than the observed value from the real data, so that the global p -value is $2/10 = 0.2$.

Table 14.2 Example data for region-wide multiple testing based on Ref. [49].

Data	Results for SNP m						
	T^1	p^1	T^2	p^2	T^3	p^3	p^{min}
Real data	2.57	0.0102	2.16	0.0308	1.96	0.0500	0.0102
Permutation samples							
1	2.56	0.0105	2.11	0.0349	1.35	0.1770	0.0105
2	1.68	0.0930	0.98	0.3271	0.15	0.8808	0.0930
3	0.81	0.4179	2.66	0.0078	2.76	0.0058	0.0078
4	0.22	0.8259	2.05	0.0404	0.01	0.9920	0.0404
5	2.22	0.0264	1.65	0.0989	0.08	0.9362	0.0264
6	1.43	0.1527	0.58	0.5619	0.89	0.3735	0.1527
7	1.00	0.3173	2.10	0.0357	1.87	0.0615	0.0357
8	0.12	0.9045	1.62	0.1052	2.46	0.0139	0.0139
9	2.37	0.0178	1.34	0.1802	1.06	0.2891	0.0178
10	2.43	0.0151	1.77	0.0767	2.60	0.0093	0.0093

14.4.2 How can p -values be simulated on the genome-wide level?

To overcome the necessity of permuting genome-wide data as described above, Lin [418] suggested a simulation approach that can be applied on the genome-wide level. It is based on the idea that all commonly used test statistics can be approximated by score statistics. Specifically, $T_k = s_k' \Sigma_k^{-1} s_k$ can be used for the hypotheses $k =$

$1, \dots, K$, where $s_k = \sum_{i=1}^n s_{ki}$, n is the sample size, and $\Sigma_k = \sum_{i=1}^n s_{ki}s'_{ki}$. Under the null hypothesis H_k of no association, s_k is approximately normally distributed with mean 0 and covariance Σ_k , so that T_k is approximately χ^2 distributed with the degrees of freedom (d.f.) given by the dimension of s_k . If now the true hypotheses $H_{k1}, \dots, H_{kt}$ are considered, $(s_{k1}, \dots, s_{kt})'$ is approximately multivariate normally distributed with mean 0 and covariance $\Sigma_{km} = \sum_{i=1}^n s_{ki}s'_{mi}$ between s_k and s_m , with $k, m = k_1, \dots, k_t$.

The core of the approach is to define standard normally distributed random variables $G_1, \dots, G_n$ independent of the data, and $\tilde{s}_k = \sum_{i=1}^n s_{ki}G_i$ and $\tilde{T}_k = \tilde{s}'_k \Sigma_k^{-1} \tilde{s}_k$. Thus, $\tilde{s}_{k1}, \dots, \tilde{s}_{kt}$ is again multivariate normally distributed with mean 0 and covariance Σ_{km} between $\tilde{s}_k$ and $\tilde{s}_m$, with $k, m = k_1, \dots, k_t$. Conditional on the data, the $\tilde{s}_k$ are weighted sums of independent standard normally distributed random variables. By using Σ_{km} of the observed data in the simulations, the covariance structure is maintained. Thus, the conditional joint distribution of $(\tilde{T}_{k1}, \dots, \tilde{T}_{kt})$ given the data is approximately the same as the unconditional joint distribution of $(T_{k1}, \dots, T_{kt})$, so that the former distribution can be used to approximate the latter.

For this, realizations of the distribution $(\tilde{T}_{k1}, \dots, \tilde{T}_{kt})$ can be obtained by generating the normally distributed random samples $G_1, \dots, G_n$ while holding the data at their observed values. To control the FWER, define the observed test statistics for the $k = 1, \dots, K$ loci as $\tilde{T}_1, \dots, \tilde{T}_K$. In a step-down procedure, starting with hypothesis H_1 , reject H_k if $P(\max_{k \leq m \leq K} \tilde{T}_m \geq \tilde{T}_k) \leq \alpha$, provided that $H_1, \dots, H_{k-1}$ have been tested and rejected. Following this, the FWER adjusted p -value for hypothesis H_k is then $\tilde{p}_k = \min\{\alpha : H_k \text{ is rejected at FWER} = \alpha\}$.

This very general approach has the clear advantage of overcoming the conservativeness of classical procedures. In addition, unlike permutation and other resampling methods, one is only required to simulate independent normally distributed random variables instead of the genotype or phenotype data. Also, repeated analyses of simulated data sets are not needed.

14.4.3 How can adjustments for multiple testing be performed using the effective number of tests?

As an alternative to the Monte-Carlo simulation or permutation approaches described above, one can estimate the number of independent tests, i.e., the *effective number of tests*, e.g., as in Refs. [175, 523]. This allows to subsequently use standard correction methods for independent multiple tests.

A very general approach has been proposed by Gao et al. [235]. They determine the number of effective tests using a principal components analysis (PCA) approach (see Sections 11.4 and 13.4) using the following steps:

1. Estimate the correlation between SNPs using the composite LD (see Chapter 10).
2. The pairwise correlation matrix from the SNP data set is derived.
3. A principal components analysis is performed and the eigenvalues derived.

4. The number of effective tests, M_{eff} is determined by the number of eigenvalues that explain 99.5% of the variation in the data.
5. The point-wise significance level is determined applying the Bonferroni correction through $\alpha_p = \alpha_e / M_{\text{eff}}$.

Using this algorithm, the number of effective tests on the Human1M (Illumina) and the Mapping 500K Array Set (Affymetrix) was estimated to be 378,368 and 208,408, respectively, yielding α_p of $1.3 \cdot 10^{-7}$ and $2.4 \cdot 10^{-7}$ [234].

14.5 HOW DO WE USE ACCUMULATING DATA FROM GENOME-WIDE ASSOCIATION STUDIES?

The data of a GWA study rarely stands on its own. Instead, it has become common practice that most GWA studies are conducted as part of a process that accumulates evidence of association. A typical version is that a GWA study is performed in the first stage to select interesting regions and that these regions are then evaluated in later stages. There is, however, a lot of variation in what these later stages look like. For example, they might be the so-called *in-silico replication* stages. This means that another GWA study had been conducted previously and that the regions can be simply looked up in the available results. Another possibility is that a few or a larger number of SNPs are genotyped separately in probands which were not part of the original GWA study. Then, there is also variation regarding how the evidence across the stages is combined. Some studies use the first GWA stage for screening and the later stages to formally test for statistical significance, whereas others actually combine data to reach a conclusion.

Finally, the actual aim of the procedure may vary from saving cost and time in a multistage design over decreasing the chance of false positive findings in replication or validation of results to increasing power in a meta-analysis. We will therefore sketch these three approaches one by one.

14.5.1 What multistage designs are used in practice?

A *multistage design* is defined by the following facts: First, the assessment of data in later stages is determined by results from previous stages, making them different from meta-analyses. Second, data are combined across stages for the conclusion of the report, which differentiates them from replication.

What makes these designs attractive? A number of reasons have been given, e.g., in the general review of Ref. [185]. The first and obvious reason is cost efficiency: Even though a single-stage design always has greater power, substantial savings might be achieved with restricting the genome-wide genotyping to a subset of the available probands. The second reason is that a multistage design might allow to increase the density of SNPs in interesting regions in later stages, so that these are mapped on a finer scale. Finally, with the different stages utilizing different genotyping technologies, technical artifacts specific to one platform can be avoided.

To apply multistage designs in GWA studies, one might be tempted to select one of the classical group-sequential procedures that are standard for clinical trials. These classical designs have been adapted for GWA studies in three aspects. First, the primary aim in clinical trials usually is to stop a trial early in case of significance. The aim in GWA studies, however, typically is to extract those SNPs in the first stage with promising effects and avoiding expenditure on SNPs that have no prospect of reaching significance [371]. Second, stopping early does not result in the termination of the entire study but merely in the focus on those promising SNPs which are then followed up in the later stages. Finally, we have to consider that the genotyping of a single SNP is much cheaper on a genome-wide platform as opposed to the separate genotyping via TaqMan or Sequenom (Chapter 3). This leads to the specific feature of sequential designs for GWA studies that minimizing the sample size overall cannot be equated with minimizing the costs. Indeed, even minimizing the number of genotypings in the entire study does not minimize the costs.

To accommodate these differences, a number of GWA study approaches have been proposed. Although these largely build on classical group-sequential designs, these allow more flexibility in the context of GWA studies (see, e.g., Refs. [487, 517, 590, 618, 688]). However, to date, we are not aware of any GWA study having actually implemented this group-sequential approach despite the obvious potential to save costs.

14.5.2 How do we replicate results from a genome-wide association study?

In the last but one step in Figure 14.1, results from a GWA study are replicated. Given that a GWA study might result in associations with p -values of less than 10^{-8} , why is independent replication so important? Even in the case of a true association, the estimated effect in an initial study reporting an association is often exaggerated relative to the estimated effect in subsequent studies, a phenomenon known as the *jackpot effect* [297] or *winner's curse* [374]. A classical analysis by Ioannidis and colleagues [334] found that the results of first and following studies on the same association correlate only moderately. This problem is aggravated by weak true effects as typical for GWA studies and relatively small sample sizes in the first study. Related to regression to the mean, it occurs primarily because with a small sample, a weak effect becomes significant only if the effect is overestimated.

An important point to consider, however, is that the goal of a GWA study is not the estimation of the effect *per se*, but the identification of an association. Indeed, given the chance, the sample is compiled to maximize the genetic effect, as described in Section 11.1. Also, effect estimation in a GWA study is difficult anyways since the discovered alleles are most likely only surrogates for the causal variant, and the effect of the association might not mirror the actual clinical effect on a genetic pathway [374].

Nonetheless, replication is the gold standard to distinguish true from false positive effects, and so the question arises when the attempt of a replication might actually

be successful? To set a standard, a Working Group on Replication in Association Studies convened by the National Cancer Institute and the National Human Genome Research Institute defined the following consensus criteria for a valid replication [115]:

1. The replication studies need to have sufficient sample size.
2. The replication should be in independent samples.
3. The phenotype should be the same or very similar.
4. The population should be similar.
5. The magnitude of effect should be similar, and the effect should be in the same direction with the same SNP or an SNP in very high LD.
6. The same genetic model should be used.
7. A joint or combined analysis should lead to a smaller p -value than the initial one.
8. The rationale for the SNP selection should be based on LD, putative function, or findings in the literature.
9. The replication reports should include the same level of detail for design and analysis plans.

It should be noted that issue 5 asks for replication in the same SNP or a very good proxy, as defined by high LD. Success at this is termed *exact replication*. In contrast, one may relax this criterion to allow association signals in a region around the previously selected SNP to yield *regional replication*.

While this is statistically rather straightforward, we would like to note that there is a lot of heterogeneity regarding the terms used, and the terms *replication* and *validation* are usually used interchangeably to describe the confirmation of previously reported associations. We have already suggested how to differentiate these terms and basically reserved replication for situations where the second sample is drawn from the same population as the original one as described above. Validation, on the other hand, might be used to indicate that different populations are included, where differences may concern the ethnic background, the phenotype definition, the recruitment or sampling strategy, and the time point of investigation [327]. As a failed validation in this sense is difficult to interpret, we focus on actual replication of results throughout.

14.5.3 How are meta-analyses of GWA studies conducted?

As described in previous sections, for complex diseases, adequate power is often lacking in a single GWA study to detect small to moderate effect sizes. Since it is often difficult or impossible to increase the sample sizes within a study, many research groups have started to pool their data sets and perform meta-analyses across them. Indeed, meta-analyses of GWA studies have been heralded as the novel gold standard for the detection of new genetic loci [669]. And despite a number of challenges mentioned in the following, the benefit is obvious: larger sample sizes can be investigated, thus increasing the power and reducing the chance of the winner's

curse. In addition, it should be noted that a number of biases that typically pose a problem in meta-analyses of clinical data are not an issue here. For example, publication bias is usually not present.

Different types of summaries and meta-analyses need to be distinguished, and the commonly used classification system is as follows [68]:

Type I: Qualitative summary, the narrative review article.

Type II: Quantitative summary of published data (usually called meta-analysis). Here, aggregated data from different studies are analyzed, e.g., genotype frequencies or log odds ratios (logOR) at every SNP.

Type III: Re-analysis of individual data based on primary studies. This is often called *pooled analysis* in epidemiology.

Type IV: Prospectively planned, pooled analysis of several studies, where pooling is already a part of the protocol. Data collection procedures, definition of variables, questions and hypotheses are standardized for the individual studies.

Sharing individual proband data as in meta-analyses of types III and IV has the clear advantage that analyses on the study level can be highly standardized and that statistical analyses are more flexible. For example, in addition to single SNP analyses one is able to look at haplotype or interaction effects. Although this is an option on the candidate gene level [597], there are ethical and legal restrictions on sharing individual data on the genome-wide level, which make this approach usually not feasible.

Before the data are actually combined, great care needs to be taken regarding the standardized quality control of the data in the single studies and the comparability of the statistical analyses. It is vitally important to use a common *statistical analysis plan* (SAP) that is agreed upon by all participants. Whenever different genotyping platforms are used in the participating studies, imputation plays a crucial role. For example, if not imputed, the overlap between the Genome-Wide Human SNP Array 6.0 by Affymetrix and the Human1M chip by Illumina only includes about 250,000 SNPs [157].

In the pooling of data, one specific challenge lies in making sure that all SNPs are coded in the same way. In general, all SNPs are reported as on the positive strand of the most recent NCBI build and flipped if required [157]. Unfortunately, the strand information is not always available. Furthermore, A/T and C/G SNPs deserve special attention because, e.g., an A/T SNP cannot be distinguished from a T/A SNP. For most of these cases, this problem can be solved by comparing the allele frequencies, but only if the minor allele frequency is not close to 50%.

Finally, the joint analysis of aggregated data can be based on different techniques described in standard textbooks [653, 715] and sketched in the next section.

14.5.3.1 Fixed and random effects models for meta-analyses As before (see Section 11.2.4), we denote a study by $k = 1, \dots, K$ and the logOR of a specific study, i.e., the within-study effect, by θ_k . To model θ_k , either *fixed effect (FE) models* or *random effect (RE) models* may be used. In an FE model, we assume that for every

study, the underlying effect is the same, i.e., $\theta_1 = \theta_2 = \dots = \theta^F$. The effect over all studies, i.e., the combined effect θ^F may be estimated using a linear combination of the study-specific effects

$$\hat{\theta}^F = \frac{\sum_{k=1}^K w_k^F \hat{\theta}_k}{\sum_{k=1}^K w_k^F}$$

with weights w_k^F . These should be proportional to the inverse variance, thus $\hat{w}_k^F = 1/\widehat{\text{Var}}(\theta_k)$. The approach is thus called *inverse variance weighting*. The confidence interval for θ^F can be computed easily from

$$\hat{\theta}^F \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\theta^F)},$$

where $\widehat{\text{Var}}(\theta^F)$ is given by $1/\sum_{k=1}^K \hat{w}_k^F$ and $z_{1-\frac{\alpha}{2}}$ denotes the upper $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution.

In general, the FE model is not adequate because this assumes the existence of a single effect that is common to all studies. Therefore, the only source of variation is the within-study variation that is a function of the sample size and the variance in the single studies. Thus, the basic concept of the FE model is that the set of studies have identical characteristics and effect sizes. In the meta-analysis of GWA studies, a certain degree of heterogeneity is present. This can occur for a number of reasons including the case-control definition and other reasons that are non-specific for GWA studies. In addition to that, the study populations may truly differ genetically, e.g., in the genetic effect, in the LD pattern, or in the use of different genotyping platforms. As a result, RE models should be the method of choice. In applications, however, FE models are commonly reported because these give lower p -values.

In the RE model, we allow between-study heterogeneity. The classical is the *DerSimonian and Laird (DSL) approach* [160] we describe in the following. We assume that differences between the true study-specific effects and the overall effect are normally distributed with mean 0 and variance τ^2 . Then, similar to above, the overall effect is estimated by

$$\hat{\theta}^R = \frac{\sum_{k=1}^K w_k^R \hat{\theta}_k}{\sum_{k=1}^K w_k^R}.$$

In this case, the weights $\hat{w}_k^R = 1/(\widehat{\text{Var}}(\theta_k) + \hat{\tau}^2)$ reflect the heterogeneity. Different methods have been suggested to estimate τ^2 , and the DSL approach is described below.

14.5.3.2 How can heterogeneity between studies be assessed? One classical approach to assess heterogeneity between studies is *Cochran's Q*, estimated by

$$\hat{Q} = \sum_{k=1}^K \hat{w}_k^F (\hat{\theta}_k - \hat{\theta}^F)^2$$

that follows a χ^2 distribution with $K - 1$ d.f. Because this test has low power, a p -value of less than 0.1 is usually considered a significant evidence of heterogeneity. A drawback of this statistic is its dependence on the number of studies.

Using Cochran's Q we can define the variance between studies by

$$\tau^2 = \frac{Q - (K - 1)}{\sum_{k=1}^K w_k^F - \frac{\sum_{k=1}^K (w_k^F)^2}{\sum_{k=1}^K w_k^F}}.$$

An alternative measure is I^2 that expresses the percentage of between-study variability that is attributable to heterogeneity rather than chance. It can be derived from Q by

$$I^2 = \max \left\{ 0, \frac{Q - (K - 1)}{Q} \right\}.$$

It has been suggested that values below 31% denote only mild heterogeneity, and that $I^2 > 56\%$ indicates severe heterogeneity [294]. A disadvantage of this measure is its dependence on sample size [573]. It is therefore suggested not to rely on this statistic alone for a decision on relevant heterogeneity.

14.6 WHAT IS THE CLINICAL IMPACT OF RESULTS FROM A GWA STUDY?

The past years have shown that GWA studies are successful in identifying genetic variants associated with complex diseases. However, what has the clinical impact of these results been? We would like to emphasize that the main goal of GWA studies is not the prediction of individual disease risk but the discovery of biologic pathways underlying polygenic diseases and traits. Therefore, we cannot expect a direct clinical impact or immediate clinical utility of the results. Nonetheless, there is a strong tendency to use results from GWA studies to try to predict individual risk.

14.6.1 How can we evaluate a genetic predictive test?

To evaluate the success of predicting individual risk based on genetic variants, the framework of the ACCE project should be considered. With detailed information on the web site of the project, the ACCE acronym represents the four main criteria for evaluating a predictive genetic test (Figure 14.3):

1. *Analytical validity*: How accurately and reliably does the test measure the genotype of interest? This is largely determined by technical aspects and evaluated within the quality control procedures detailed in Chapter 4. It should be noted, however, that limitations regarding the analytical validity have further implications for the following criteria.
2. *Clinical validity*: How consistently and accurately does the test detect or predict the intermediate or final outcomes of interest [35]? This first depends on the strength of the association between the typed genotypes and the phenotypical

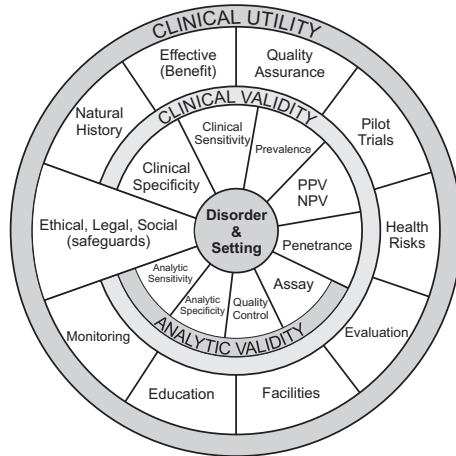


Fig. 14.3 Criteria to evaluate a genetic test. Reproduced from www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm with permission from Muin J. Khoury and Glenn Palomaki, Centers for Disease Control and Prevention, Atlanta, GA, USA.

outcome that is measured using effect sizes as described in Section 11.2. However, on top of that, other statistical measures are required that evaluate how well the genotypes are able to differentiate between cases and controls. These measures are described below. Although advocated by some authors, there are no general thresholds that define a test to be clinically valid. Among others, the availability of alternative prediction models based on classical risk factors, the aim of testing, the burden and costs of disease, and the availability of treatment decide which level of validity is deemed satisfactory.

3. *Clinical utility*: How likely does the test significantly improve patient outcome [36]? For this, the result needs to influence both medical decisions and patient behavior. For evaluation, criteria derived from evidence-based medicine can be considered:
 - (a) Is prevention or early intervention available? The test has clinical utility only if the respective disease could be prevented early.
 - (b) Does a positive test result increase the patient's motivation for a behavioral change? Only if the preventive measures are realized by the patient, the test is clinically useful. On the one hand, this is more often the case in diseases for which an effective intervention is available [446]. On the other hand, the risk conveyed by a genetic test is usually perceived to be less modifiable, so that the patient is less likely to change the behavior [306].
 - (c) Does a negative test result lead to the avoidance of further risks? Finally, if the test is negative, this should not lead to a behavior totally free of care since this only implies a reduced risk but not the exclusion of the disease.
4. *ELSI*: Ethical, legal, and social implications that may arise in the context of using the test. For these, the reader may refer to Refs. [193, 307].

From a statistical point of view, the evaluation of clinical validity merits special attention, and important definitions and standards can be found in guidelines of

the U.S. Food and Drug Administration (FDA) or the European Agency for the Evaluation of Medicinal Products (EMA) (e.g., CPMP/EWP/1119/98). In the following section, we specifically target the prediction based on a single genetic marker and on multiple genetic markers.

14.6.2 How can we evaluate the clinical validity of a test based on a single genetic marker?

If a single SNP is used as a predictive test, classical measures of diagnostic accuracy, such as sensitivity, specificity, and predictive values can be used to evaluate the clinical validity. Technical details on these measures can be found in standard textbooks (e.g., Refs. [527, 757]). In the following, we denote the presence and absence of the disease by $D = 1$ and $D = 0$, respectively, and carriage and non-carriage of the predisposing variant by $G = 1$ and $G = 0$. The *sensitivity* (*sens*) is defined as the probability to be test positive, given that the individual is affected. In the notation from Table 14.3, it is the true positive fraction, $TPF = \text{sens} = P(G = 1|D = 1)$. Similarly, the *specificity* (*spec*) as the probability to be test negative, given that the individual is unaffected, is derived from the false positive fraction, that is, $1 - FPF = \text{spec} = 1 - P(G = 1|D = 0)$. To estimate *sens* from data, we use the proportions of probands which carry the specific genetic variant and have the disease in all diseased patients. Likewise, the proportion of unaffected non-carriers of the specific variant in all unaffected probands estimates the *spec*. Approximate $100(1 - \alpha)\%$ confidence intervals can be obtained analogously to the confidence intervals for risks as given in Section 11.2.3.

Table 14.3 Classification of test results by disease status.

		Disease	
		Yes ($D = 1$)	No ($D = 0$)
Variant	Present ($G = 1$)	True positive	False positive
	Absent ($G = 0$)	False negative	True negative

To further evaluate the predictive worth, we can estimate the positive and negative predictive values. The *positive predictive value* (*ppV*) is defined as the probability for disease given the proband carries the variant, $ppV = P(D = 1|G = 1)$. Similarly, the *negative predictive value* (*npV*) is given as the probability for being a control given that the proband is test negative, so that $npV = P(D = 0|G = 0)$. By using Bayes' theorem, *ppV* and *npV* can be expressed by (Problem 14.1)

$$\begin{aligned}
 ppV &= \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec}) \cdot (1 - \text{prev})}, \\
 npV &= \frac{\text{spec} \cdot (1 - \text{prev})}{\text{spec} \cdot (1 - \text{prev}) + (1 - \text{sens}) \cdot \text{prev}},
 \end{aligned}$$

where $prev$ denotes the prevalence of the disease. Using estimated values of $sens$ and $spec$ as described above leads to $\widehat{ppV}$ and $\widehat{npV}$ estimated from the data. To obtain confidence intervals, it is recommended to use logit transformations [463]:

$$\text{logit}(\widehat{ppV}) = \log \left(\frac{\widehat{sens} \cdot prev}{(1-\widehat{spec})(1-prev)} \right) \quad \text{logit}(\widehat{npV}) = \log \left(\frac{\widehat{spec} \cdot (1-prev)}{(1-\widehat{sens}) \cdot prev} \right)$$

The variances for these expressions are estimated by

$$\begin{aligned} \widehat{\text{Var}}(\text{logit}(\widehat{ppV})) &= \left(\frac{1-\widehat{sens}}{\widehat{sens} \cdot r} \right) + \left(\frac{\widehat{spec}}{(1-\widehat{spec}) \cdot s} \right) \text{ and} \\ \widehat{\text{Var}}(\text{logit}(\widehat{npV})) &= \left(\frac{\widehat{sens}}{(1-\widehat{sens}) \cdot r} \right) + \left(\frac{1-\widehat{spec}}{\widehat{spec} \cdot s} \right), \end{aligned}$$

where r and s denote the numbers of affected and unaffected probands, respectively. From this, approximate $100(1-\alpha)\%$ confidence intervals for $\text{logit}(\widehat{ppV})$ and $\text{logit}(\widehat{npV})$ are given by

$$\text{logit}(\widehat{ppV}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\text{logit}(\widehat{ppV}))} \text{ and } \text{logit}(\widehat{npV}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\text{logit}(\widehat{npV}))},$$

respectively, where $z_{1-\frac{\alpha}{2}}$ is the upper $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. An example is given in Problem 14.2.

14.6.3 How can we evaluate the clinical validity of a test based on multiple genetic markers?

We have already discussed that the effects of single SNPs usually are only small to moderate for complex diseases. Therefore, the predictive value of a single SNP is rather limited. As a result, the interest has turned to the second question of genetic profiling, which is the use of a combination of a number of variants for prediction. In the simplest case, a score is used by counting the number of predisposing variants a single proband carries. More sophisticated and promising approaches introduce a weighting of the various variants depending on the respective genetic effect. Thus, one does no longer use a binary test outcome but a quasi continuous test score. The clinical validity of this can be evaluated using the *receiver operating characteristic (ROC) curve* that plots $sens$ against $(1 - spec)$ for any possible cutoff in the score (see Figure 14.4).

For quantification, a classical measure is its *area under the curve (AUC)*. It is equivalent to the *c-statistic* that is the probability that, among pairs of affected and unaffected probands, the affected proband has the higher score. Several approaches have been proposed to estimate the *c-statistic*, and details on the estimation of the *AUC* and corresponding confidence intervals can be found in the literature [527, 757]. As can be seen from Figure 14.4, the larger the *AUC*, the further the ROC is away from the diagonal, and the higher $sens$ and $spec$ are. An *AUC* = 0.5 corresponds to the case that the ROC is on the diagonal, so that there is no predictive information.

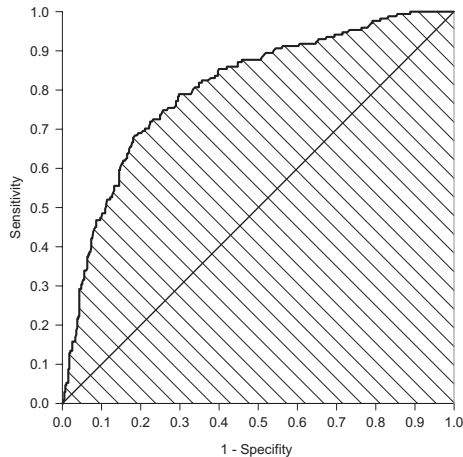


Fig. 14.4 Receiver operating characteristic curve (ROC) for a diagnostic test. The shaded area denotes the area under the ROC curve (AUC).

While simulation studies on the use of genetic profiles were slightly optimistic (see, e.g., Refs. [339, 729]), real-data analyses have painted a different picture, and an example is given below. Finally, we note that clinical validity should not be based on single measures such as the *AUC*, although this is advocated by some colleagues. Instead, they should be complemented, e.g., by the proportion of re-classified patients [142] or other statistics [528, 529]. Importantly, clinical validity of a test depends on the specific clinical situation.

Example 14.2. Zheng and colleagues [756] used five genetic variants to predict prostate cancer. Specifically, they developed a score based on the number of risk genotypes at these variants plus a positive family history (see Table 14.4).

As expected, a higher score was associated with cancer. The odds ratio for cancer for a man with a score of ≥ 2 compared to a lower score was 1.66. Based on these data, we can estimate *sens* and *spec* using this score as a predictive test. If we define a score of ≥ 2 as test positive, $\widehat{sens} = 0.68$. Similarly, $\widehat{spec} = 0.44$. If we now assume a *prev* of about 17% for males, we can estimate *ppV* and *npV* as given above, which are $\widehat{ppV} = 0.20$ and $\widehat{npV} = 0.87$. Therefore, the probability for a prostate cancer without having undergone a test is about 17% for any man. Having a positive test result, that is, having two or more risk variants, increases this probability to about 20%. Interestingly, the second to last column of Table 14.4 gives the frequency of the score in the general population, estimated from the numbers of cases and controls and the overall prevalence. This shows that a risk score of at least 2 can be expected in almost 60% of the population. If the test were negative, the risk would decrease to about $100\% - 87\% = 13\%$. Thus, at this threshold, no extreme changes in risk can be expected.

The respective values using other cutoffs are also shown in the table. From these, we see that the risk for cancer changes more extremely given more extreme risk scores. However, these are also much rarer in the general population. Thus, the score can explain only a very small percentage of cancer cases. This can be quantified by

Table 14.4 Data for a risk score to predict prostate cancer taken from Ref. [756].

Score	Cases	Controls	OR (95% CI)	<i>sens</i> (95% CI)	<i>spec</i> (95% CI)
≥ 0 (all)	2893	1728	1.00	1.00	0.00
≥ 1	2749	1554	2.14 (1.70; 2.69)	0.95 (0.94; 0.96)	0.10 (0.09; 0.12)
≥ 2	1971	973	1.66 (1.47; 1.88)	0.68 (0.66; 0.70)	0.44 (0.41; 0.46)
≥ 3	918	351	1.82 (1.58; 2.10)	0.32 (0.30; 0.33)	0.80 (0.78; 0.82)
≥ 4	276	65	2.70 (2.05; 3.56)	0.10 (0.09; 0.11)	0.96 (0.95; 0.97)
≥ 5	40	5	4.83 (1.90; 12.27)	0.01 (0.01; 0.02)	1.00 (0.99; 1.00)

Score	Freq	<i>ppV</i> (95% CI)	<i>npV</i> (95% CI)	<i>AR</i> (95% CI)
≥ 0 (all)	1.00	0.17	1.00	1.00
≥ 1	0.91	0.18 (0.18; 0.18)	0.91 (0.89; 0.92)	0.51 (0.40; 0.61)
≥ 2	0.58	0.20 (0.19; 0.21)	0.87 (0.86; 0.88)	0.27 (0.22; 0.33)
≥ 3	0.22	0.24 (0.22; 0.26)	0.85 (0.85; 0.86)	0.14 (0.11; 0.17)
≥ 4	0.05	0.34 (0.29; 0.40)	0.84 (0.84; 0.84)	0.06 (0.05; 0.07)
≥ 5	0.00	0.49 (0.28; 0.71)	0.83 (0.83; 0.83)	0.01 (0.01; 0.02)

OR: odds ratio; CI: confidence interval; *sens*: sensitivity; *spec*: specificity; *freq*: frequency of risk score; *ppV*: positive predictive value; *npV*: negative predictive value; *AR*: attributable risk as defined in Section 11.2.3.

the attributable risk as defined in Section 11.2, and the results are given in the last column of the table. They show that an extreme risk score value of 5 or greater increases the risk for prostate cancer to almost 50%, but since it occurs in less than 1% of the male population, it can explain only about 1% of the cases.

This specific example was based on only a handful genetic variants. In addition to that, the quality of the prediction based on known genes has been investigated for a number of complex diseases, and examples are presented in Table 14.5. Here, the AUC based on genetic variants is given (AUC gene), and we see only limited predictive value if only genetic variants are considered. However, there were exceptions not shown in the table for studies on age-related macular degeneration and hypertriglyceridemia with AUC estimates of 0.8. These might be seen as promising examples that were based on a special selection of probands.

Table 14.5 also shows answers to the question of how genetic factors might improve the prediction over only clinical risk factors. Again, the results are rather discouraging at this point.

Can genetic profiling of complex diseases be improved? A plausible answer might be that not enough genetic variants have yet been identified for a given condition, so that prediction will improve as more detections are made. However, it has to be acknowledged that there is a natural upper limit to the predictive values that is defined by the heritability of the trait [130]. Moreover, it has been argued that many genes

Table 14.5 Examples from recent studies on the prediction of complex diseases using clinical risk factors and multiple genes.

Disease	# genes	Clinical risk factors	AUC gene	AUC clin	AUC both	Ref.
CHD	4	Age, triglycerides, cholesterol, systolic blood pressure, smoking	0.62	0.66	0.70	[324]
MI after surgery	3	AXC time, coronary grafts, previous cardiac surgery	0.70	0.70	0.76	[538]
T2D	3	Age, sex, BMI	0.56	0.82	0.82	[676]
T2D	17	Age, sex, BMI	0.60	0.78	0.80	[399]
T2D	17	Age, sex, BMI	0.60	0.66	0.68	[674]

Table adapted from Ref. [340]. # genes: Number of genes included in the risk score calculation; AUC: area under the receiver operating characteristic curve; AUC gene: AUC based on genetic variants; AUC clin: AUC based on clinical risk factors; AUC both: AUC based on both clinical risk factors and genetic variants; CHD: coronary heart disease; MI: myocardial infarction; T2D: type 2 diabetes; BMI: body mass index; AXC: aortic crossclamping.

might not be able to improve the prediction because they are also predisposing the risk factors themselves.

A number of applications are, however, foreseeable, one of which being the area of pharmacogenetics. Here, genetic association studies are carried out much the same as for diseases, but the phenotypes are reactions to drugs such as side effects or therapeutic failure. Finding a strong and reliable association between a genotype and the drug response might then actually lead to an adaptation of the therapy for the individual.

14.7 WHAT MAY WE CONCLUDE FROM GENOME-WIDE ASSOCIATION STUDIES, AND WHAT COMES NEXT?

The introduction of GWA studies undoubtedly has been a great success. Approximately 350 GWA studies have identified more than 250 genetic loci in which common genetic variants occur that are reproducibly associated with complex diseases. For this, a valuable resource is the online catalogue of GWA studies [296].

On top of the identified genetic loci, some interesting specific facts became known [21, 296]:

1. There are associations in genes/regions that were previously unrecognized to have a role in the particular trait.
2. The results imply new mechanistic connections between diseases, e.g., between Morbus Crohn and type 1 diabetes.
3. Many of the identified SNPs reside outside transcription units. Indeed, there are associations in many regions that are currently annotated as being gene

poor. This suggests a greater than anticipated role for noncoding SNPs in common diseases.

4. Overall, the identified SNPs have very modest effects. The odds ratio often is < 1.5 , and the median across studies is 1.33.
5. Strong evidence for epistasis is lacking so far, but this has only rarely been investigated with convincing power.
6. The vast majority of heritability remains unexplained. In contrast, the predictive power of classical, non-genetic, factors usually is much higher.

It has also been found that the frequencies of identified risk alleles were common and mostly well above 5% with a median of 36% [296]. This has been interpreted as support for the common disease—common variant hypothesis (see Chapter 2). However, simulations suggest that these observations are the result of power [329]. Since it has therefore become clear that GWA studies in their typical form do not suffice to elucidate the genetics of common diseases, a number of strategies are likely to be followed.

The first and obvious strategy is to increase the power by including larger sample sizes, thus enabling the detection of yet smaller effects. A specific option is to combine evidence across a number of performed GWA studies by carrying out meta-analyses as outlined in Section 14.5.3.

The second strategy uses the available GWA data but goes beyond the rather simple single SNP association analysis described above. Here, multimarker and haplotype analyses can increase the efficiency for less common alleles, and a successful example of a genome-wide haplotype analysis is the study by Trégouët and colleagues [668]. Also, interactions between SNPs are usually not investigated systematically, but they may provide important insights [441]. Finally, a sensible analysis of subgroups, defined by phenotypic features, age, gender, or ethnicity, has often been neglected.

Third, we may integrate GWA data with other data that might be closer to the interesting biological context. For example, a combination with genome-wide expression studies and other disease-relevant immediate phenotypes, such as proteins or metabolites could help. A review on this is given in Ref. [499].

While the first three strategies use the available GWA data in different ways, the fourth strategy analyzes other forms of genetic variation. Here, CNVs may be looked at. Despite interesting results in candidates, however, large-scale results have been rather discouraging (see Ref. [348]). Another form of structural variation will be in focus, rare genetic variants that afford sequencing. With whole-genome sequencing becoming increasingly available, this is likely to be the next hot topic.

14.8 PROBLEMS

Problem 14.1 For the positive predictive value and the negative predictive value, show that

$$ppV = P(D = 1|G = 1) = \frac{sens \cdot prev}{sens \cdot prev + (1 - spec) \cdot (1 - prev)} \text{ and}$$

$$npV = P(D = 0|G = 0) = \frac{spec \cdot (1 - prev)}{spec \cdot (1 - prev) + (1 - sens) \cdot prev}.$$

Problem 14.2 To illustrate the predictive value of a single SNP, we consider the case of an undoubted association between the SNP rs1333049 on chromosome 9p21 and myocardial infarction [597]. In a meta-analysis, the odds ratio for disease carrying at least variant was estimated to be about 1.4. Based on a frequency of the genotypes of about 0.75 in the general population and a *prev* of myocardial infarction of about 0.1, we estimate that the frequency of carrying the genotypes given that the individual is a case is about 0.79, whereas it is about 0.73 in the controls. Estimate *sens* and *spec* of a test based on this genotype constellation as well as the *ppV* and *npV*.

URLs

1000 GENOMES PROJECT

<http://www.1000genomes.org>

ACCE PROJECT

http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm

THE HAPMAP PROJECT

<http://www.hapmap.org/index.html.en>

IMPUTE

<https://mathgen.stats.ox.ac.uk/impute/impute.html>

MACH

<http://www.sph.umich.edu/csg/abecasis/MACH/>

NATIONAL HUMAN GENOME RESEARCH INSTITUTE (NHGRI) CATALOG OF PUBLISHED GENOME-WIDE ASSOCIATION STUDIES

www.genome.gov/gwastudies

POINTS TO CONSIDER ON THE EVALUATION OF DIAGNOSTIC AGENTS BY THE EUROPEAN AGENCY FOR THE EVALUATION OF MEDICINAL PRODUCTS, CPMP/EWP/1119/98

<http://www.emea.europa.eu/pdfs/human/ewp/111998en.pdf>

POLICY FOR SHARING OF DATA OBTAINED IN NIH SUPPORTED OR CONDUCTED GENOME-WIDE ASSOCIATION STUDIES (GWAS)

<http://www.grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html>

Appendix

Algorithms Used in Linkage Analyses

The Elston–Stewart algorithm and the Lander–Green algorithm are the two standard algorithms for exactly calculating likelihoods in pedigrees. Both have their value in specific situations. The computational time of the Elston–Stewart algorithm increases linearly with the number of subjects in a pedigree but exponentially with the number of markers to be considered jointly. In contrast, the computational time of the Lander–Green algorithm grows linearly with the number of markers and exponentially with the number subjects. Although other approaches, including Markov Chain Monte–Carlo methods, are able to analyze large pedigrees and a large number of markers, they consider the underlying configurations in proportion to their likelihood only. Thus, configurations that are theoretically possible but highly unlikely are often excluded, and the likelihood therefore is only approximate; for a detailed description of this approach, see Refs. [243, 620]. In this appendix, we restrict our attention to the two algorithms providing an exact solution to the likelihood. The fundamental ideas of the Elston–Stewart algorithm are described in Section A.1. The Lander–Green algorithm as introduced by Kruglyak and colleagues [378] and Strauch [641] is discussed in detail in Section A.2. We sketch the Cardon–Fulker algorithm in the last section of this appendix for which the computational time increases linearly in both pedigree size and number of markers. It provides a simple approximation for estimating the proportion of alleles shared IBD in a multipoint setting. For this, single-marker IBD distributions as obtained, e.g., from the application of the Elston–Stewart algorithm may be utilized.

A.1 WHAT IS THE ELSTON–STEWART ALGORITHM?

In this section, we describe the basic ideas of the Elston–Stewart algorithm by rewriting the likelihood of a simple pedigree. Section A.1.1, we sketch the fundamental ideas of the Elston–Stewart algorithm for calculating the joint probability of phenotypes. For these computations, we introduce a latent trait locus. The extension to a trait locus and a linked marker locus is described in Section A.1.2. We refrain from describing how the IBD distribution can be calculated using the Elston–Stewart algorithm, and we refer the interested reader to Ref. [25] instead.

A.1.1 What are the fundamental ideas of the Elston–Stewart algorithm?

The fundamental ideas of the Elston–Stewart algorithm can be best described by rewriting the likelihood of a simple pedigree. We start off with Figure A.1 that shows a three-generation pedigree with seven pedigree members.

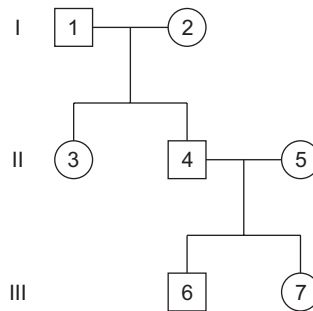


Fig. A.1 Simple pedigree for illustrating the fundamental ideas of the Elston–Stewart algorithm.

We denote the binary phenotypes of the seven individuals in the pedigree by $y_1, \dots, y_7$ and their unobserved genotypes at the trait locus by $g_1, \dots, g_7$. We are interested in calculating the joint probability $P(y_1, \dots, y_7)$ because this probability forms the basis for both model-based and model-free linkage analyses. The first step for calculating the joint probability is the introduction of the latent genotypes at the trait locus $g_1, \dots, g_7$ via the law of total probability. The joint probability may then

be written in terms of genotype probabilities in founders, transition probabilities in non-founders and penetrances, yielding

$$\begin{aligned}
 P(y_1, \dots, y_7) &= \sum_{g_1, \dots, g_7} P(y_1, \dots, y_7, g_1, \dots, g_7) \\
 &= \sum_{g_1, \dots, g_7} P(y_1, \dots, y_7 | g_1, \dots, g_7) P(g_1, \dots, g_7) \\
 &\stackrel{*}{=} \sum_{g_1, \dots, g_7} P(y_1 | g_1) \cdot \dots \cdot P(y_7 | g_7) P(g_1, \dots, g_7) \\
 &\stackrel{**}{=} \sum_{g_1, \dots, g_7} P(y_1 | g_1) \cdot \dots \cdot P(y_7 | g_7) \\
 &\quad \cdot P(g_4) P(g_1 | g_4) P(g_2 | g_1, g_4) P(g_3 | g_1, g_2, g_4) \\
 &\quad \cdot P(g_5 | g_1, g_2, g_3, g_4) P(g_6 | g_1, g_2, g_3, g_4, g_5) \\
 &\quad \cdot P(g_7 | g_1, g_2, g_3, g_4, g_5, g_6) \\
 &\stackrel{***}{=} \sum_{g_1, \dots, g_7} P(y_1 | g_1) \cdot \dots \cdot P(y_7 | g_7) \\
 &\quad \cdot P(g_1) P(g_2) P(g_3 | g_1, g_2) P(g_4 | g_1, g_2) \\
 &\quad \cdot P(g_5) P(g_6 | g_4, g_5) P(g_7 | g_4, g_5) \\
 &= \sum_{g_1, \dots, g_7} \underbrace{\prod_{i=1}^7 P(y_i | g_i)}_{\text{penetrances}} \cdot \underbrace{P(g_1) P(g_2) P(g_5)}_{\text{founder genotypes}} \\
 &\quad \cdot \underbrace{P(g_3 | g_1, g_2) P(g_4 | g_1, g_2) P(g_6 | g_4, g_5) P(g_7 | g_4, g_5)}_{\text{transition probabilities}}. \quad (\text{A.1})
 \end{aligned}$$

* holds since the penetrances are defined as functions for the phenotype of an individual given his/her genotype. The phenotype of an individual is thus assumed to be independent of the genotype of another family member. ** is the result of applying the chain rule for probabilities. Finally, *** is a direct consequence of the independence of founder genotypes and the conditional independence of non-founder genotypes given founder genotypes.

More generally speaking, for an arbitrary pedigree without loops with n subjects one can write the joint probability as

$$P(y_1, \dots, y_n) = \sum_{g_1, \dots, g_n} \prod_{i=1}^n P(y_i | g_i) \prod_{i=1}^f P(g_i) \prod_{i=f+1}^n P(g_i | g_{i,p}, g_{i,m}). \quad (\text{A.2})$$

Here, f denotes the founders, and p and m denote father and mother of the respective offspring.

It is important to note that the first sum in our example pedigree involves g^n different terms. Even if the trait locus in our example pedigree is only diallelic, there

are 3^7 terms because there are seven individuals in the pedigree and three possible genotypes per individual. Thus, for this pedigree, the products are evaluated for all 3^7 genotype combinations separately even if some of them are 0, e.g., because of Mendelian incompatibilities. Therefore the main aim of the Elston–Stewart algorithm is to reduce the calculations by deleting unnecessary terms and re-sorting the likelihood.

The first idea of the Elston–Stewart algorithm involves summation over latent genotypes. The sums are pushed to the right as far as possible. Thereby, the evaluation of the joint probability is more efficient. The idea will be explained using the example pedigree. Using Eq. (A.1), this gives

$$\begin{aligned}
 P(y_1, \dots, y_7) &= \sum_{g_1} P(y_1|g_1)P(g_1) \sum_{g_2} P(y_2|g_2)P(g_2) \\
 &\quad \cdot \left(\sum_{g_3} P(y_3|g_3)P(g_3|g_1, g_2) \right) \\
 &\quad \cdot \left(\sum_{g_4} P(y_4|g_4)P(g_4|g_1, g_2) \left(\sum_{g_5} P(y_5|g_5)P(g_5) \right. \right. \\
 &\quad \left. \left. \cdot \left(\sum_{g_6} P(y_6|g_6)P(g_6|g_4, g_5) \right) \left(\sum_{g_7} P(y_7|g_7)P(g_7|g_4, g_5) \right) \right) \right).
 \end{aligned} \tag{A.3}$$

The reader should note that the expression in the second line of Eq. (A.3) can be separately evaluated given g_1 and g_2 . Furthermore, the sums for g_6 and g_7 can be evaluated independent of each other. In contrast, the sums for g_4 and g_5 are nested, so are the computations of g_1 and g_2 .

Although the presentation above is the standard way for explaining the Elston–Stewart algorithm, it does not make clear the separate evaluation of parts of the pedigree, which is the second idea of Elston and Stewart [187]. Thus, probabilities are conditioned on genotypes, the upper part of a pedigree is conditional independent of its lower part given the genotypes. The conditional independence allows separation of different parts of the pedigree, a process termed *peeling*. The original algorithm of Elston and Stewart was limited in this respect, and the concept of peeling pedigrees has been generalized by Morton and MacLean [481].

For explaining this concept of peeling, we use a different presentation of the likelihood and note that the only connection between the “upper part” individuals 1, 2, and 3 with the “lower part” individuals 5, 6, and 7 of the pedigree is via individual 4. If the genotypes of individual 4 at the trait locus are known, the “upper” and the “lower” part of the pedigree are conditionally independent. And this conditional

independence is the second key issue. As before, we are interested in the joint probability $P(y_1, \dots, y_7)$, and we restart from Eq. (A.1):

$$\begin{aligned}
 P(y_1, \dots, y_7) &= \sum_{g_1, \dots, g_7} P(y_1|g_1) \cdot \dots \cdot P(y_7|g_7) \\
 &\quad \cdot P(g_4)P(g_1|g_4)P(g_2|g_1, g_4)P(g_3|g_1, g_2, g_4) \\
 &\quad \cdot P(g_5|g_1, g_2, g_3, g_4)P(g_6|g_1, g_2, g_3, g_4, g_5) \\
 &\quad \cdot P(g_7|g_1, g_2, g_3, g_4, g_5, g_6) \\
 &= \sum_{g_1, \dots, g_7} P(y_1|g_1) \cdot \dots \cdot P(y_7|g_7) \\
 &\quad \cdot P(g_4)P(g_1|g_4)P(g_2|g_1, g_4)P(g_3|g_1, g_2) \\
 &\quad \cdot P(g_5|g_4)P(g_6|g_4, g_5)P(g_7|g_4, g_5), \tag{A.4}
 \end{aligned}$$

where Eq. (A.4) is a direct consequence of the conditional independence of the “upper part” of the pedigree from the “lower part” of the pedigree and that the sibling genotypes are not required if the parental genotypes are available. Pushing the sums to the right gives

$$\begin{aligned}
 P(y_1, \dots, y_7) &= \sum_{g_4} P(y_4|g_4)P(g_4) \\
 &\cdot \left(\sum_{g_1} P(y_1|g_1)P(g_1|g_4) \left(\sum_{g_2} P(y_2|g_2)P(g_2|g_1, g_4) \left(\sum_{g_3} P(y_3|g_3)P(g_3|g_1, g_2, g_4) \right) \right) \right) \\
 &\cdot \left(\sum_{g_5} P(y_5|g_5)P(g_5|g_4) \left(\sum_{g_6} P(y_6|g_6)P(g_6|g_4, g_5) \right) \left(\sum_{g_7} P(y_7|g_7)P(g_7|g_4, g_5) \right) \right).
 \end{aligned}$$

Finally, we can make the notation convenient by rewriting the second and the third line of the last equation. The second line is identical to

$$\begin{aligned}
 &\sum_{g_1, g_2, g_3} P(y_1|g_1)P(y_2|g_2)P(y_3|g_3)P(g_1|g_4)P(g_2|g_1, g_4)P(g_3|g_1, g_2, g_4) \\
 &= \sum_{g_1, g_2, g_3} P(y_1, y_2, y_3|g_1, g_2, g_3)P(g_1, g_2, g_3|g_4) \\
 &= \sum_{g_1, g_2, g_3} P(y_1, y_2, y_3|g_1, g_2, g_3, g_4) = P(y_1, y_2, y_3|g_4)
 \end{aligned}$$

which involves the individuals of the “upper part” of the pedigree. Only the upper part is still involved if $P(g_4)$ is moved from the first line to the second line

$$a(g_4) = P(g_4)P(y_1, y_2, y_3|g_4),$$

and the resulting expression is termed the *anterior* of individual 4. Similarly, the third line is identical to

$$\begin{aligned} & \sum_{g_5, g_6, g_7} P(y_5|g_5)P(y_6|g_6)P(y_7|g_7)P(g_5|g_4)P(g_6|g_4, g_5)P(g_7|g_4, g_5, g_6) \\ &= \sum_{g_5, g_6, g_7} P(y_5, y_6, y_7|g_5, g_6, g_7)P(g_5, g_6, g_7|g_4) \\ &= \sum_{g_5, g_6, g_7} P(y_5, y_6, y_7|g_4, g_5, g_6, g_7) = P(y_5, y_6, y_7|g_4) = p(g_4) \end{aligned}$$

and termed the *posterior* of individual 4. The structure of both *anterior* and *posterior* is quite analogous. In summary, the joint probability is given by

$$\begin{aligned} P(y_1, \dots, y_7) &= \sum_{g_4} \underbrace{P(g_4)P(y_1, y_2, y_3|g_4)}_{a(g_4)} p(g_4)P(y_4|g_4) \\ &= \sum_{g_4} a(g_4)p(g_4)P(y_4|g_4). \end{aligned}$$

The pedigree is thus split into parts involving all individuals in the “upper” and the “lower part” of the pedigree, and the pedigree is peeled. The second fundamental idea of the Elston–Stewart algorithm is that a single nuclear family can be analyzed at a time. This idea can be generalized quite easily to more complex pedigree structures [212, 644].

To show the application of the Elston–Stewart algorithm on a more complex pedigree, we extended the example pedigree of Figure A.1 by including a third nuclear family. The extended pedigree is displayed in Figure A.2.

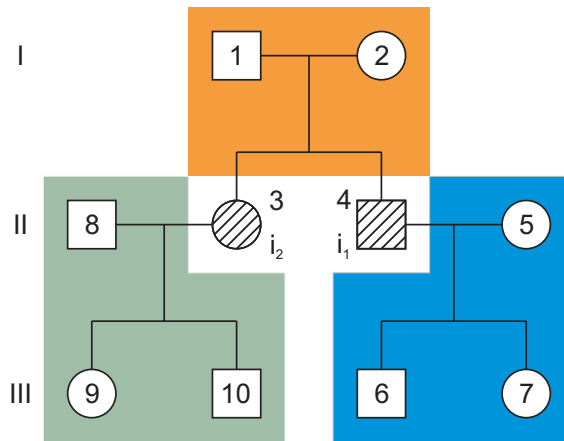


Fig. A.2 Complex pedigree for illustrating the fundamental idea of the Elston–Stewart algorithm.

For calculating the likelihood of this pedigree, we take the following steps:

1. Subdivision of the pedigree in nuclear families: The example pedigree is divided in three sub-pedigrees as marked in Figure A.2.

2. Selection of one nuclear family in the corner of the pedigree: A nuclear family is selected and the corresponding partial likelihood $L_{i_1}^*(g_{i_1})$ is calculated. Within this nuclear family, i_1 is the individual that connects this first nuclear family with the remaining pedigree, and i_1 is called *pivot element*. For calculating $L_{i_1}^*(g_{i_1})$, the genotypes g_{i_1} of the pivot element are assumed as fixed. Starting point in the example pedigree is, e.g., the nuclear family comprising subjects 4, 5, 6, and 7. In this case, the pivot element i_1 is individual 4, and the first nuclear family is collapsed on i_1 . To the overall likelihood, individual i_1 therefore provides its own contribution and its posterior.
3. Calculation of $L_{i_2}^*(g_{i_2})$ of the adjacent nuclear family: $L_{i_2}^*(g_{i_2})$ again refers to its own pivot element i_2 that connects the second nuclear family with the rest of the pedigree. Individual 3 is the appropriate second pivot element.
4. After completion of the second nuclear family, the third nuclear family is incorporated. The corresponding adjacent nuclear family in the example pedigree consists in individuals 1, 2, 3, and 4. In the third nuclear family comprising individuals 3, 8, 9, and 10, the pivot element is arbitrary.
5. Finally, the overall likelihood is calculated by summing the partial likelihoods on all genotypes of the last pivot element.

This procedure results in the following formulae:

$$\begin{aligned}
 P(y_1, \dots, y_{10}) &= \sum_{g_4} a(g_4) p(g_4) P(y_4 | g_4) \\
 &= \sum_{g_4} \underbrace{P(g_4) P(y_1, y_2, y_3, y_8, y_9, y_{10} | g_4)}_{a(g_4)} \underbrace{P(y_5, y_6, y_7 | g_4)}_{p(g_4)} P(y_4 | g_4) \\
 &= \sum_{g_4} \left(\sum_{g_1} P(g_1) P(y_1 | g_1) \left(\sum_{g_2} P(g_2) P(y_2 | g_2) \right. \right. \\
 &\quad \left. \left. \left(\sum_{g_3} P(g_3 | g_1, g_2) P(y_3 | g_3) p(g_3) \right) P(g_4 | g_1, g_2) \right) \right) p(g_4) P(y_4 | g_4) \\
 &= \sum_{g_4} P(g_4) P(y_4 | g_4) p(g_4) \left(\sum_{g_3} P(g_3 | g_4) P(y_3 | g_3) \right),
 \end{aligned}$$

where $p(g_3)$ is the posterior of subject 3, i.e., $p(g_3) = P(y_8, y_9, y_{10} | g_4)$.

The ideas described above can be applied to extended pedigrees of arbitrary size without loops. And for the case of loops, the reader may refer to Refs. [212, 644].

In summary, the Elston–Stewart algorithm consists in the following two fundamental ideas:

1. The algorithm involves summation over latent genotypes. The sums are pushed to the right as far as possible.
2. Probabilities are conditioned on genotypes. This leads to a conditional independence of the upper part of a pedigree from its lower part given the genotypes. The conditional independence allows separation of different parts of the pedigree, a process termed *peeling*.

A.1.2 How can a single linked marker be incorporated in the Elston–Stewart algorithm?

In the previous section, the Elston–Stewart algorithm has been introduced using a binary phenotype y . This may, however, be easily extended to quantitative traits, where the penetrance function is replaced by an appropriate density for y given genotype g .

The next step, however, is the computation of a joint probability for a single autosomal trait locus and a single linked marker locus. Here, we follow the excellent presentation of Stricker and colleagues [645]. As before, we assume that the phenotypic values y_i are conditionally independent given the genotypes at the trait locus. Thus, the conditional probability of the phenotypic values $\mathbf{y} = (y_1, \dots, y_n)$ given the vector $\mathbf{z}$ of the haplotypes at the trait and the marker locus can be written as

$$P(\mathbf{y}|\mathbf{z}) = P(\mathbf{y}|\mathbf{g}) = \prod_{i=1}^n P(y_i|g_i),$$

with $\mathbf{g} = (g_1, \dots, g_n)$ denoting the genotypes at the trait locus (see Eq. (A.2)). Analogously, $P(\mathbf{z}) = \prod_{i=1}^f P(z_i) \prod_{i=f+1}^n P(z_i|z_{i,p}, z_{i,m})$. Let T_a, T_b be alleles at the trait locus and M_c, M_d alleles at the marker locus with genotype frequencies $P(T_a, T_b)$ and $P(M_c, M_d)$, respectively. Then, the probability of the haplotypes z_i of subject i can be computed as

$$P(z_i) = P\left(\begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix}\right) = c P(T_a, T_b) P(M_c, M_d),$$

where $c = 1$ if $T_a = T_b$ or if $M_c = M_d$, and $\frac{1}{2}$ otherwise.

Before we finally can formulate the joint probability $P(y_1, \dots, y_n)$, we need to define the transition probabilities for the joint genotypes. Here, we use sex-specific recombination fractions. Let $\tau_s\left(\begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix} \middle| \begin{smallmatrix} T_e & T_f \\ M_h & M_i \end{smallmatrix}\right)$ denote the probability that a parent of gender s with haplotypes $\begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix}$ transmits haplotype $\begin{smallmatrix} T_e & T_f \\ M_h & M_i \end{smallmatrix}$ to the offspring. The transmission probabilities can be computed as

$$\begin{aligned} \tau_s\left(\begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix} \middle| \begin{smallmatrix} T_e & T_f \\ M_h & M_i \end{smallmatrix}\right) &= \frac{1}{2}(1 - \theta_s)(\delta_{T_a T_e} \delta_{M_c M_h} + \delta_{T_b T_e} \delta_{M_d M_h}) \\ &\quad + \frac{1}{2}\theta_s(\delta_{T_a T_e} \delta_{M_d M_h} + \delta_{T_b T_e} \delta_{M_c M_h}), \end{aligned}$$

where θ_s is the sex-specific recombination fraction between the trait and the marker locus and δ_{ab} is Kronecker's delta, equaling 1 if $a = b$, and 0 otherwise. With these transmission probabilities, the transition probability $P(z_i|z_{i,p}, z_{i,m})$ can be calculated as

$$\begin{aligned} P(z_i|z_{i,p}, z_{i,m}) &= P\left(\begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix} \middle| \begin{smallmatrix} T_e & T_k \\ M_h & M_l \end{smallmatrix}, \begin{smallmatrix} T_o & T_r \\ M_q & M_u \end{smallmatrix}\right) \\ &= \begin{cases} \tau_p\left(\begin{smallmatrix} T_e & T_k \\ M_h & M_l \end{smallmatrix} \rightarrow \begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix}\right) \cdot \tau_m\left(\begin{smallmatrix} T_o & T_r \\ M_q & M_u \end{smallmatrix} \rightarrow \begin{smallmatrix} T_b & T_d \\ M_d & M_d \end{smallmatrix}\right), & \text{if } T_a = T_b \text{ and } M_c = M_d \\ \tau_p\left(\begin{smallmatrix} T_e & T_k \\ M_h & M_l \end{smallmatrix} \rightarrow \begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix}\right) \cdot \tau_m\left(\begin{smallmatrix} T_o & T_r \\ M_q & M_u \end{smallmatrix} \rightarrow \begin{smallmatrix} T_b & T_d \\ M_d & M_d \end{smallmatrix}\right) \\ \quad + \tau_p\left(\begin{smallmatrix} T_e & T_k \\ M_h & M_l \end{smallmatrix} \rightarrow \begin{smallmatrix} T_b & T_d \\ M_d & M_d \end{smallmatrix}\right) \cdot \tau_m\left(\begin{smallmatrix} T_o & T_r \\ M_q & M_u \end{smallmatrix} \rightarrow \begin{smallmatrix} T_a & T_b \\ M_c & M_d \end{smallmatrix}\right), & \text{otherwise,} \end{cases} \end{aligned}$$

yielding

$$P(y_1, \dots, y_n) = \sum_{z_1, \dots, z_n} \prod_{i=1}^n P(y_i | g_i) \prod_{i=1}^f P(z_i) \prod_{i=f+1}^n P(z_i | z_{i,p}, z_{i,m})$$

as joint probability. It is straightforward to extend this approach to the situation of more than one linked marker locus. The calculation of IBD probabilities from the Elston–Stewart algorithm is, however, complicated, and the reader is therefore referred to Ref. [25].

A.2 WHAT IS THE LANDER–GREEN ALGORITHM?

In this section, the Lander–Green algorithm [392] is described in detail, and we borrow heavily from the excellent German description of the algorithm by Strauch [641]. The algorithm utilizes the conditional independence along a chromosome that holds if interference is neglected. In contrast to the Elston–Stewart algorithm, where the key issue is the conditional independence of individuals within pedigrees, the Lander–Green algorithm considers all individuals jointly for every chromosomal position.

The algorithm is divided into two parts. In the first step, the complete inheritance information—which can be used, e.g., for deriving the IBD distribution—is obtained using the information at a single marker. In the second step, inheritance information is combined from different markers. This, in turn, can be used to estimate the IBD distribution or to conduct model-based linkage analysis at any chromosomal position along the chromosome.

In the following sections, we consider only one family at a time because different families can be assumed to be independent. They thus contribute additively to the log likelihood.

A.2.1 How can the inheritance distribution be estimated at a single genetic marker?

A.2.1.1 What is an inheritance vector? The first part of the algorithm consists in determining the inheritance pattern for every single genetic marker. As explained in Chapter 8, a meiosis can be described by a bit where 1 denotes the maternally inherited allele and 0 denotes the paternally inherited allele. All inheritance bits in a pedigree are collected in the 1×2^b dimensional *inheritance vector*, where $b = 2n$ is the number of meioses in a pedigree with n non-founders.

In practice, the inheritance vector cannot be determined unambiguously but has to be estimated from the data: if an individual is homozygous at a genetic marker, it is unclear which allele has been inherited by his/her offspring. This also holds true for untyped individuals where the genotype cannot be reconstructed from the pedigree. In both cases, the bit describing the meiosis cannot be determined. Furthermore, the parental origin of the founder alleles also remains indefinite. Because the bit for the

first offspring of a founder cannot be determined, it can be either 0 or 1. However, all other bits for meioses of this founder depend on the first bit. As a consequence, one has to consider the *inheritance distribution* at a genetic marker j , which is the distribution V_j over all 2^b possible inheritance vectors. This probability distribution is the key to determining the joint likelihood and determines the IBD distribution of a pair of relatives.

The concept of the inheritance vector and the inheritance distribution is depicted in Figure A.3 for a sib-pair with genotyped parents. Figure A.3.a) displays marker data of a notional locus. The founder alleles of both the father and the mother cannot be determined. Furthermore, the mother is homozygous with the 3 allele so that meioses 3 and 4 are not informative. Figure A.3.b) shows the true inheritance pattern of a notional completely polymorphic locus where the parental origin of the founder alleles is known. The meioses are numbered from 1 to 4. The first meiosis is the paternal meiosis to individual 3 and the final meiosis is the maternal meiosis to individual 4. Paternally inherited alleles are always listed first. Thus, a_3 denotes the paternally inherited allele in the mother. With this notation, it can be seen that the true inheritance vector of the second pedigree is 0011. The first offspring has received a_1 paternally. a_1 is the paternal allele in the father so that the first bit in the inheritance vector is a 0. Analogously, the third bit in the inheritance vector indicates the paternal transmission to the second offspring. This time the maternally inherited allele in the father a_2 has been transmitted to the offspring so that the third bit is a 1.

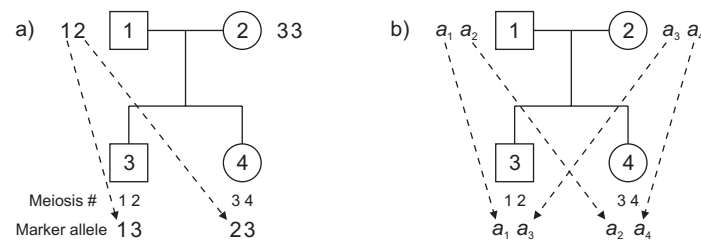


Fig. A.3 Sib-pair with genotyped parents for illustrating the concept of the inheritance vector and the inheritance distribution. Pedigree a) displays marker data of a notional locus. Here, the founder alleles cannot be determined and the maternal meioses are not informative. Pedigree b) shows the true inheritance pattern of a notional completely polymorphic locus where the parental origin of the founder alleles is known.

Table A.1 shows the 16 possible inheritance vectors, the a priori inheritance distribution that is obtained in the absence of marker data, and the a posteriori inheritance distribution of the two pedigrees from Figure A.3 given the marker data. A priori, all 16 inheritance distributions are equally likely. However, only one inheritance vector is compatible with the pedigree in Figure A.3.b so that the true a posteriori inheritance distribution is a degenerated one with probability 1 for the inheritance vector 0011. For the pedigree in Figure A.3.a, it can be seen that the father has transmitted the 1 allele to his son and the 2 allele to his daughter. Therefore, bits one and three of the inheritance vector have to be different. However, the parental

origin of the two alleles remains unclear. Therefore, all constellations 0, 1 and 1, 0 for bits one and three are possible. Because the mother is homozygous at the marker locus, bits two and four can take all possible combinations for these bits. Given the genotype data, eight different inheritance vectors are possible that are equally likely with probability $\frac{1}{8}$ according to the third Mendelian law.

Table A.1 Summary of the inheritance distributions of the two pedigrees displayed in Figure A.3.

Inheritance vector	P_a priori	Figure A.3.a P_a posteriori	Figure A.3.b P_{true}
0000	1/16	0	0
0001	1/16	0	0
0010	1/16	1/8	0
0011	1/16	1/8	1
0100	1/16	0	0
0101	1/16	0	0
0110	1/16	1/8	0
0111	1/16	1/8	0
1000	1/16	1/8	0
1001	1/16	1/8	0
1010	1/16	0	0
1011	1/16	0	0
1100	1/16	1/8	0
1101	1/16	1/8	0
1110	1/16	0	0
1111	1/16	0	0

The first column lists all 16 possible inheritance vectors and the second column gives the a priori probability for these inheritance vectors. Bits one and two represent the paternal and maternal meioses, respectively, for the first offspring. These are denoted by meioses 1 and 2 in Figure A.3. Bits three and four analogously stand for the paternal and maternal meiosis, respectively, for the second offspring. The third and fourth columns contain the a posteriori probability of the pedigrees shown in Figure A.3.a and A.3.b given the marker data.

A.2.1.2 How can the a posteriori inheritance distribution be determined for a single genetic marker using a graph theoretical approach? Let $P_{\text{Marker } j}(v) = P_{a \text{ posteriori}, j}(v)$ denote the conditional probability for inheritance vector v given the genotypes at marker j . This probability is given by

$$P_{\text{Marker } j}(v) = P(V_j = v | M_j) = \frac{P(M_j, V_j = v)}{P(M_j)}$$

and can be rewritten using the Bayes' formula as

$$P_{\text{Marker } j}(v) = \frac{P(M_j | V_j = v) P(V_j = v)}{\sum_{w \in V} P(M_j | V_j = w) P(V_j = w)}. \quad (\text{A.5})$$

Here, $\mathcal{V}$ denotes the set of all 2^b inheritance vectors. Because $P(V_j = v)$ is identical to the a priori probability that is $\frac{1}{2^b}$, the last term in the numerator and the denominator of Eq. (A.5) cancel out. Therefore, together with the notation $q_j = P(M_j|V_j = v_j)$, Eq. (A.5) reduces to

$$P_{\text{Marker } j}(v) = \frac{q_j(v)}{\sum_{w \in \mathcal{V}} q_j(w)}. \quad (\text{A.6})$$

An extreme situation occurs if the family is completely uninformative at a marker. This happens if all typed individuals in the pedigree are homozygous with the same genotype. In this case, the a posteriori distribution given marker data and the a priori distribution without any marker information are identical. However, in some families it is possible to determine the inheritance bits of untyped individuals using information from their relatives and Mendelian laws. This can substantially reduce the number of possible inheritance vectors and subsequently leads to a greater concentration of probabilities.

The general case for determining the inheritance distribution given a single marker is quite complicated. The key to this is the probability $q_j(v) = P(M_j|V_j = v)$ that is used in the Bayes formula. It can be determined with an algorithm that is based on a graph theoretical process with the following steps:

1. A single graph is generated to describe a specific inheritance vector.
2. The number of nodes is identical to the number $2f$ of founder alleles. Here, f is the number of founders.
3. Instead of the observed alleles, unambiguous placeholder alleles $a_l, l = 1, 2, \dots, 2f$ are used to denote the nodes.
4. Every individual in the pedigree is represented by an edge connecting his/her two placeholder alleles based on the specific inheritance vector v . Note that two nodes are connected by several edges if the inherited paternal genotypes of several offspring are identical.
5. The observed alleles are matched to the placeholder alleles, at the same time complying with Mendelian laws. If an individual is genotyped, the edge is marked with the two observed alleles of the individual's genotype. For these individuals, the observed allele has to appear at all edges that are connected to the respective node. This assignment determines the orientation of the genotype to its neighboring edge. It also determines the node of the second allele of the genotype.
6. If an assignment leads to an incompatibility, the orientation of the genotype has to be altered in the first step. If the altered orientation also leads to an incompatibility at any position in the graph, the inheritance vector has zero probability given the marker data.

These steps are taken for all theoretically possible inheritance vectors, and the probability $q_j(v)$ is determined by using all compatible inheritance vectors. However, in-

stead of using equal weights for all compatible inheritance vectors, they are weighted by their allele frequencies:

$$q_j(v) = P(M_j | V_j = v) = \sum_{\substack{\text{founder alleles} \\ \text{compatible with } v \\ \text{and all individual genotypes}}} \prod_{l=1}^{2f} p_j(A(a_l))$$

Here, for all founders $l = 1, \dots, 2f$, $A(a_l)$ denotes the relationship between the placeholder allele a_l and the observed alleles A_i .

Example A.1. The algorithm is exemplified by using the pedigrees from Figure A.3. This nuclear pedigree has two founders. Therefore, there are $2 \cdot 2$ placeholder alleles a_l . Figure A.4.a shows the corresponding nodes for the placeholder alleles a_1 – a_4 of the graph. The parents have genotypes a_1, a_2 and a_3, a_4 , respectively, and Figure A.4.b displays the nodes and the parental edges, which are denoted by M and F . Figure A.4.c shows all nodes and edges for the 16 theoretically possible inheritance vectors using the placeholder alleles a_1 – a_4 . O_1 and O_2 along the edges denote offspring 1 and 2, respectively.

We already know from Table A.1 that the first two inheritance vectors compatible with the pedigree in Figure A.3.b are $v_3 = (0010)$ and $v_4 = (0011)$, while $v_1 = (0000)$ and $v_2 = (0001)$ are incompatible with the marker configuration. Figure A.4.d displays the observed marker alleles placed along the edges of the graph for inheritance vectors v_1 – v_4 . This yields $a_1 = 1, a_2 = 2, a_3 = 3$, and $a_4 = 4$ for both v_3 and v_4 . Similarly, the incompatibility is illustrated for v_1 and v_2 .

For the compatible inheritance vectors v_3 and v_4 , we finally determine q_j as $q_j(v_3) = q_j(v_4) = \prod_{i=1}^4 p_j(A(a_i)) = p_{j,1} \cdot p_{j,2} \cdot p_{j,3} \cdot p_{j,4}$, where $p_{j,i}$ denotes the allele frequency of allele A_i at marker j .

In this example, all eight compatible inheritance vectors have the same probability q_j , so that $P_{\text{Marker } j}(v) = \frac{1}{8}$. In general, the probabilities q_j can differ with the inheritance vector, e.g., if genotypes are missing.

A.2.2 How can the inheritance distribution be estimated at an arbitrary chromosomal position given all genetic markers?

The algorithm described in the previous section can be used to estimate the probability distribution $P_{\text{Marker } j}(v) = P(V_j = v | M_j)$ of the inheritance vector at any marker position j for all inheritance vectors v . We are, however, interested in the inheritance distribution at an arbitrary chromosomal position x :

$$P_{\text{complete}, x}(v) = P(V(x) = v | M_1, \dots, M_J)$$

for markers $M_1, \dots, M_J$, which are numbered from $j = 1$ to J according to their order on the chromosome. In Section A.2.2.1, we will consider only chromosomal positions that are identical to genetic marker positions so that $x = j$ for $j = 1, \dots, J$.

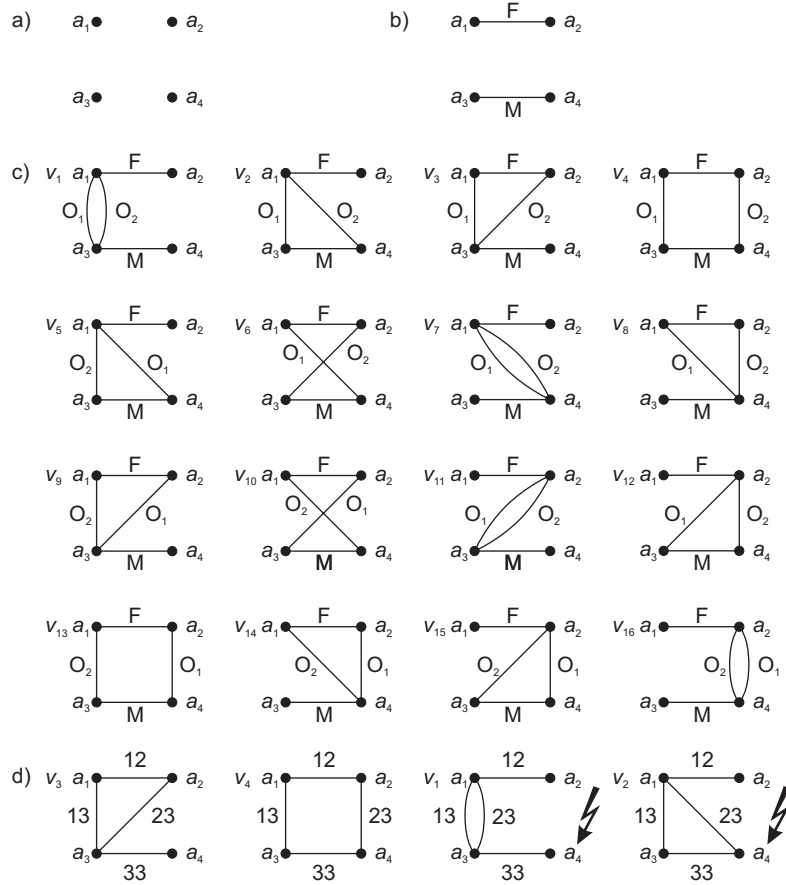


Fig. A.4 Illustration of the graph theoretical algorithm. (a) The four placeholder alleles of the two founders as nodes. (b) Nodes and maternal M as well as paternal F edges. (c) All nodes and edges for the 16 theoretically possible inheritance vectors using the placeholder alleles. (d) The compatibility of inheritance vectors v_3 and v_4 with the observed marker data as well as the incompatibility of inheritance vectors v_1 and v_2 .

In the last section of this appendix, we generalize the concept to a genetic position that is not identical to that of a genetic marker position.

A.2.2.1 How can the inheritance distribution be estimated at a chromosomal position that is identical to the position of a genetic marker? For calculating the probability, we will use a Markov property in this section. The Markov property relies on the assumption of absence of interference. With this assumption, the inheritance pattern of a family at all chromosomal positions can be described by an inhomogeneous Markov chain along the genetic markers of a chromosome. The observed states of the Markov chain correspond to the marker genotypes. The inher-

itance vectors are, however, hidden. Therefore, this Markov chain can be interpreted as a *hidden Markov model* where the hidden states, i.e., the inheritance vectors, can be partly reconstructed using the marker genotypes. This idea for connecting observed and hidden states has already been illustrated for the single-marker case, where the probabilities $q_j(v)$ have been calculated using the graph theoretic approach. In the multi-marker case, one considers inheritance vectors v_j and v_{j+1} at markers j and $j + 1$. If the inheritance vectors v_j and v_{j+1} are different, at least one transition has occurred between markers j and $j + 1$. Specifically, the difference in a single bit between neighboring markers indicates a change in the inheritance pattern that is identical to a recombination in the interval between the two markers. The Hamming distance $H(v_j, v_{j+1})$ measures the total number of recombinations for all meioses b in a pedigree for the inheritance vectors v_j and v_{j+1} . Thus, the H counts the number of bits being different between v_j and v_{j+1} . The transition probability from inheritance vector v_j at marker j to v_{j+1} at marker $j + 1$ is subsequently given by

$$\begin{aligned} T_{v_j, v_{j+1}}(\theta_{j, j+1}) &= P(V_{j+1} = v_{j+1} | V_j = v_j) \\ &= (\theta_{j, j+1})^{H(v_j, v_{j+1})} (1 - \theta_{j, j+1})^{c-H(v_j, v_{j+1})}, \end{aligned} \quad (\text{A.7})$$

where $\theta_{j, j+1}$ denotes the recombination fraction between markers j and $j + 1$. The probabilities (A.7) for all pairs (v_j, v_{j+1}) form the $2^b \times 2^b$ transition matrix $T(\theta_{j, j+1})$ from marker j to $j + 1$, with j and $j + 1$ denoting rows and columns, respectively.

In the following, let p_j^L denote the $2^b \times 1$ vector of probabilities $p_j^L(v_j)$ at marker j conditional on all observed genotypes of markers $1, \dots, j$:

$$[p_j^L]_{v_j} = p_j^L(v_j) = P(V_j = v_j | M_1, \dots, M_j) \quad (\text{A.8})$$

The superscript L indicates that the probabilities at the marker j are conditioned to the “left.” For the first marker, one specifically obtains $[p_1^L]_{v_1} = P(V_1 = v_1 | M_1) = P_{\text{Marker 1}}(v_1)$.

Furthermore, let q_j be the $2^c \times 1$ column vector of probabilities $q_j(v_j)$ at marker j :

$$[q_j]_{v_j} = q_j(v_j) = P(M_j | V_j = v_j) \quad (\text{A.9})$$

Note that q_j is conditioned on $V_j = v_j$ in contrast to Eq. (A.8), where the condition is on the markers $M_1, \dots, M_j$.

With the notations from above and with $\circ$ denoting the *Hadamard product* of element-by-element multiplication [47], p_j^L can be calculated for $j = 2, \dots, J$ recursively as [392, 641]

$$p_{j+1}^L = \frac{[(p_j^L)^T T(\theta_{j, j+1})]^T \circ q_{j+1}}{(p_j^L)^T T(\theta_{j, j+1}) q_{j+1}} \quad (\text{A.10})$$

The validity of Eq. (A.10) can be proven by writing $\mathbf{p}_{j+1}^L$ element-wise:

$$\begin{aligned}
 [\mathbf{p}_{j+1}^L]_{v_{j+1}} &\stackrel{!}{=} \frac{[(\mathbf{p}_j^L)^T \mathbf{T}(\theta_{j,j+1})]_{v_{j+1}} [\mathbf{q}_{j+1}]_{v_{j+1}}}{(\mathbf{p}_j^L)^T \mathbf{T}(\theta_{j,j+1}) \mathbf{q}_{j+1}} \\
 &\stackrel{*}{=} \frac{\sum_{w_j \in V} q_{j+1}(v_{j+1}) T_{w_j, v_{j+1}}(\theta_{j,j+1}) p_j^L(w_j)}{\sum_{w_j, w_{j+1} \in V} q_{j+1}(w_{j+1}) T_{w_j, w_{j+1}}(\theta_{j,j+1}) p_j^L(w_j)} \\
 &\stackrel{**}{=} \frac{\sum_{w_j \in V} P(M_{j+1}|V_{j+1}=v_{j+1}) P(V_{j+1}=v_{j+1}|V_j=w_j) P(V_j=w_j | M_1, \dots, M_j)}{\sum_{w_j, w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}) P(V_{j+1}=w_{j+1}|V_j=w_j) P(V_j=w_j | M_1, \dots, M_j)} \\
 &\stackrel{***}{=} \frac{P(M_{j+1}|V_{j+1}=v_{j+1}) \sum_{w_j \in V} P(V_{j+1}=v_{j+1}, V_j=w_j | M_1, \dots, M_j)}{\sum_{w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}) \sum_{w_j \in V} P(V_{j+1}=w_{j+1}, V_j=w_j | M_1, \dots, M_j)} \\
 &\stackrel{****}{=} \frac{P(M_{j+1}|V_{j+1}=v_{j+1}) P(V_{j+1}=v_{j+1} | M_1, \dots, M_j)}{\sum_{w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}) P(V_{j+1}=w_{j+1} | M_1, \dots, M_j)} \\
 &= \frac{P(M_{j+1}|V_{j+1}=v_{j+1}, M_1, \dots, M_j) P(V_{j+1}=v_{j+1} | M_1, \dots, M_j)}{\sum_{w_{j+1} \in V} P(M_{j+1}|V_{j+1}=w_{j+1}, M_1, \dots, M_j) P(V_{j+1}=w_{j+1} | M_1, \dots, M_j)} \\
 &= P(V_{j+1} = v_{j+1} | M_1, \dots, M_{j+1})
 \end{aligned}$$

Here, “!” denotes “has to be shown.” Furthermore, * stand for element-wise writing of the vectors; **: employ definition of probabilities; ***: use Markov property $P(V_{j+1} = v_{j+1} | V_j = v_j, M_1, \dots, M_{j+1}) = P(V_{j+1} = v_{j+1} | V_j = v_j)$ because v_{j+1} depends on genotypes of markers 1 to j only via v_j ; ****: $P(M_{j+1} | V_{j+1} = v_{j+1}) = P(M_{j+1} | V_{j+1} = v_{j+1}, M_1, \dots, M_j)$ because M_{j+1} depends only on v_{j+1} .

Analogously to $\mathbf{p}_j^L$, we now define a $2^b \times 1$ column vector $\mathbf{p}_j^R$ comprising the probabilities $p_j^R(v_j)$ for all inheritance vectors at marker j conditional on the observed data at markers $j, \dots, J$:

$$[\mathbf{p}_j^R]_{v_j} = p_j^R(v_j) = P(V_j = v_j | M_j, \dots, M_J) \quad (\text{A.11})$$

Here, the superscript R indicates that the probabilities at the marker j are conditioned to the “right.” For the rightmost marker J , one specifically obtains $[\mathbf{p}_J^R]_{v_J} = P_{\text{marker } J}(v_J)$, and the recursion formula, which is analogous to Eq. (A.10), is

$$\mathbf{p}_{j+1}^R = \frac{\mathbf{q}_{j-1} \circ \mathbf{T}(\theta_{j-1,j}) \mathbf{p}_j^R}{(\mathbf{q}_{j-1})^T \mathbf{T}(\theta_{j-1,j}) \mathbf{p}_j^R}.$$

In total, one obtains the following $2^b \times 1$ column vector:

$$[\mathbf{P}_{\text{complete},j}]_{v_j} = P_{\text{complete},j}(v_j) = P(V_j = v_j | M_1, \dots, M_J)$$

of the inheritance vectors at marker j given the genotypes at all markers as

$$\mathbf{P}_{\text{complete},j} = \frac{[(\mathbf{p}_{j-1}^L)^T \mathbf{T}(\theta_{j-1,j})]^T \circ \mathbf{p}_j^R}{(\mathbf{p}_{j-1}^L)^T \mathbf{T}(\theta_{j-1,j}) \mathbf{p}_j^R}. \quad (\text{A.12})$$

The validity of Eq. (A.12) can be proven as follows:

$$\begin{aligned}
[\mathbf{P}_{\text{complete},j}]_{v_j} &= \frac{[(\mathbf{p}_{j-1}^L)^T \mathbf{T}(\theta_{j-1,j})]_{v_j} [\mathbf{p}_j^R]_{v_j}}{(\mathbf{p}_{j-1}^L)^T \mathbf{T}(\theta_{j-1,j}) \mathbf{p}_j^R} \\
&= \frac{\sum_{w_{j-1} \in \mathcal{V}} \mathbf{p}_j^R(v_j) T_{w_{j-1}, v_j}(\theta_{j-1,j}) \mathbf{p}_{j-1}^L(w_{j-1})}{\sum_{w_{j-1}, w_j \in \mathcal{V}} \mathbf{p}_j^R(w_j) T_{w_{j-1}, w_j}(\theta_{j-1,j}) \mathbf{p}_{j-1}^L(w_{j-1})} \\
&= \frac{\sum_{w_{j-1} \in \mathcal{V}} P(V_j = v_j | M_j, \dots, M_J) P(V_j = v_j | V_{j-1} = w_{j-1}) P(V_{j-1} = w_{j-1} | M_1, \dots, M_{j-1})}{\sum_{w_{j-1}, w_j \in \mathcal{V}} P(V_j = w_j | M_j, \dots, M_J) P(V_j = w_j | V_{j-1} = w_{j-1}) P(V_{j-1} = w_{j-1} | M_1, \dots, M_{j-1})} \\
&= \frac{P(V_j = v_j | M_j, \dots, M_J) \sum_{w_{j-1} \in \mathcal{V}} P(V_j = v_j, V_{j-1} = w_{j-1} | M_1, \dots, M_{j-1})}{\sum_{w_j \in \mathcal{V}} P(V_j = w_j | M_j, \dots, M_J) \sum_{w_{j-1} \in \mathcal{V}} P(V_j = w_j, V_{j-1} = w_{j-1} | M_1, \dots, M_{j-1})} \\
&= \frac{\frac{P(M_j, \dots, M_J | V_j = v_j) P(V_j = v_j)}{M_j, \dots, M_J} P(V_j = v_j | M_1, \dots, M_{j-1})}{\sum_{w_j \in \mathcal{V}} \frac{P(M_j, \dots, M_J | V_j = w_j) P(V_j = w_j)}{M_j, \dots, M_J} P(V_j = w_j | M_1, \dots, M_{j-1})} \\
&= \frac{P(M_j, \dots, M_J | V_j = v_j, M_1, \dots, M_{j-1}) P(V_j = v_j | M_1, \dots, M_{j-1})}{\sum_{w_j \in \mathcal{V}} P(M_j, \dots, M_J | V_j = w_j, M_1, \dots, M_{j-1}) P(V_j = w_j | M_1, \dots, M_{j-1})} \\
&= P(V_j = v_j | M_1, \dots, M_J) \tag{A.13}
\end{aligned}$$

In Eq. (A.13) we use as before both $P(V_j = v_j | V_{j-1} = v_{j-1}) = P(V_j = v_j | V_{j-1} = v_{j-1}, M_1, \dots, M_{j-1})$ and $P(M_j, \dots, M_k | V_j = v_j) = P(M_j, \dots, M_k | V_j = v_j, M_1, \dots, M_{j-1})$. Furthermore, $P(V_j = v_j)$ is the a priori probability for inheritance vector v_j at marker j and given by $1/2^b$ for all inheritance vectors. Instead of using Eq. (A.12), $\mathbf{P}_{\text{complete},j}$ can also be obtained by

$$\mathbf{P}_{\text{complete},j} = \frac{\mathbf{p}_j^L \circ \mathbf{T}(\theta_{j,j+1}) \mathbf{p}_{j+1}^R}{(\mathbf{p}_j^L)^T \mathbf{T}(\theta_{j,j+1}) \mathbf{p}_{j+1}^R}. \tag{A.14}$$

Equations (A.12) and (A.14) are the original formulations for calculating $\mathbf{P}_{\text{complete},j}$. The recursive nature of these formulae is, however, somehow awkward. As with most algorithms, the recursion can be dissolved and a direct computation can be used instead [377]. This forward approach requires only the transition matrices $\mathbf{T}(\theta_{j,j+1})$ and a posteriori probabilities of the inheritance vectors $q_j(v_j)$ at markers $j, j = 1, \dots, J$.

In the following, let $\mathbf{Q}_j = \text{diag}(q_j(v))$ be the $2^b \times 2^b$ diagonal matrix at marker $j, j = 1, \dots, J$. Furthermore, let $\mathbf{1} = (1, \dots, 1)^T$ denote the $2^b \times 1$ one-vector. The probability $\mathbf{P}_{\text{complete},j}$ can then also be written as:

$$\mathbf{P}_{\text{complete},j} = \frac{[\mathbf{1}^T \mathbf{Q}_1 \mathbf{T}(\theta_{1,2}) \mathbf{Q}_2 \dots \mathbf{T}(\theta_{j-1,j})]^T \mathbf{Q}_j \mathbf{T}(\theta_{j,j+1}) \dots \mathbf{Q}_{J-1} \mathbf{T}(\theta_{J-1,J}) \mathbf{Q}_J \mathbf{1}}{\mathbf{1}^T \mathbf{Q}_1 \mathbf{T}(\theta_{1,2}) \mathbf{Q}_2 \dots \mathbf{T}(\theta_{j-1,j}) \cdot \mathbf{Q}_j \mathbf{T}(\theta_{j,j+1}) \dots \mathbf{Q}_{k-1} \mathbf{T}(\theta_{k-1,k})} \tag{A.15}$$

The Hadamard product could also be used on the right side of $\mathbf{Q}_j$ because $\mathbf{Q}_j$ is diagonal. For proving the equality, we first consider the denominator:

$$\begin{aligned}
& \mathbf{1}^T \mathbf{Q}_1 \mathbf{T}(\theta_{1,2}) \mathbf{Q}_2 \dots \mathbf{T}(\theta_{j-1,j}) \cdot \mathbf{Q}_j \mathbf{T}(\theta_{j,j+1}) \dots \mathbf{Q}_{J-1} \mathbf{T}(\theta_{J-1,J}) \\
&= \sum_{v_1, \dots, v_J \in \mathcal{V}} q_J(v_J) T_{v_{J-1}, v_J}(\theta_{J-1,J}) q_{J-1}(v_{J-1}) \dots q_2(v_2) T_{v_1, v_2}(\theta_{1,2}) q_1(v_1) \\
&= \sum_{v_1, \dots, v_J \in \mathcal{V}} P(M_J | V_J = v_J) P(V_J = v_J | V_{J-1} = v_{J-1}) P(M_{J-1} | V_{J-1} = v_{J-1}) \\
&\quad \cdot \dots \cdot P(M_2 | V_2 = v_2) P(V_2 = v_2 | V_1 = v_1) P(M_1 | V_1 = v_1) \\
&= \sum_{v_1, \dots, v_J \in \mathcal{V}} P(M_J | V_J = v_J) P(M_{J-1} | V_{J-1} = v_{J-1}) \dots \cdot P(M_1 | V_1 = v_1) \\
&\quad \cdot P(V_J = v_J | V_{J-1} = v_{J-1}) P(V_2 = v_2 | V_1 = v_1) \\
&\stackrel{*}{=} \sum_{v_1, \dots, v_J \in \mathcal{V}} P(M_J | M_1, \dots, M_{J-1}, V_1 = v_1, \dots, V_J = v_J) \\
&\quad \cdot P(M_{J-1} | M_1, \dots, M_{J-2}, V_1 = v_1, \dots, V_J = v_J) \\
&\quad \cdot \dots \cdot P(M_2 | M_1, V_1 = v_1, \dots, V_J = v_J) P(M_1 | V_1 = v_1, \dots, V_J = v_J) \\
&\quad \cdot P(V_J = v_J | V_1 = v_1, \dots, V_{J-1} = v_{J-1}) \\
&\quad \cdot P(V_{J-1} = v_{J-1} | V_1 = v_1, \dots, V_{J-2} = v_{J-2}) \\
&\quad \cdot \dots \cdot P(V_2 = v_2 | V_1 = v_1) P(V_1 = v_1) \frac{1}{P(V_1 = v_1)} \\
&= \sum_{v_1, \dots, v_J \in \mathcal{V}} P(M_1, \dots, M_J | V_1 = v_1, \dots, V_J = v_J) P(V_1 = v_1, \dots, V_J = v_J) 2^b \\
&= \sum_{v_1, \dots, v_J \in \mathcal{V}} P(M_1, \dots, M_J, V_1 = v_1, \dots, V_J = v_J) 2^b = 2^b P(M_1, \dots, M_J)
\end{aligned}$$

* holds because of a conditional independence argument: the marker data M_j depend on genotypes and inheritance vectors at other markers only via v_j so that $P(M_j | V_j) = P(M_j | M_1, \dots, M_J, V_1, \dots, V_J)$. In this step, we also use the Markov property $P(V_j = v_j | V_{j-1} = v_{j-1}) = P(V_j = v_j | V_1 = v_1, \dots, V_{j-1} = v_{j-1})$.

The numerator is identical to the denominator except for the sum over v_j from marker j :

$$\begin{aligned}
& [\mathbf{1}^T \mathbf{Q}_1 \mathbf{T}(\theta_{1,2}) \mathbf{Q}_2 \dots \mathbf{T}(\theta_{j-1,j})]_{v_j} \cdot [\mathbf{Q}_j \mathbf{T}(\theta_{j,j+1}) \dots \mathbf{Q}_{J-1} \mathbf{T}(\theta_{J-1,J}) \mathbf{Q}_J \mathbf{1}]_{v_j} \\
&= \sum_{v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_J \in \mathcal{V}} P(M_1, \dots, M_J, V_1 = v_1, \dots, V_J = v_J) 2^b \\
&= 2^b P(M_1, \dots, M_J, V_j = v_j),
\end{aligned}$$

yielding in total

$$\begin{aligned}
& [P_{\text{complete},j}]_{v_j} \\
&= \frac{[1^T Q_1 T(\theta_{1,2}) Q_2 \dots T(\theta_{j-1,j})]_{v_j} [Q_j T(\theta_{j,j+1}) \dots Q_{J-1} T(\theta_{J-1,J}) Q_J 1]_{v_j}}{1^T Q_1 T(\theta_{1,2}) Q_2 \dots T(\theta_{j-1,j}) \cdot Q_j T(\theta_{j,j+1}) \dots Q_{J-1} T(\theta_{J-1,J}) Q_J 1} \\
&= \frac{2^b P(M_1, \dots, M_J, V_j = v_j)}{2^b P(M_1, \dots, M_J)} = P(V_j = v_j | M_1, \dots, M_J).
\end{aligned}$$

A.2.2.2 How can the inheritance distribution be estimated at a chromosomal position that is not identical to the position of a genetic marker?

In the previous section, we formulated $P_{\text{complete},j}(v)$ for chromosomal positions that are identical to genetic marker positions. In this section, we extend the algorithm so that we can compute $P_{\text{complete},x}(v)$ at an arbitrary chromosomal position x that may be different from a marker position j . The idea is simple: a non-informative genetic marker at position x is added to the Markov chain. This, subsequently, allows application of the results from the previous section.

In the first step, we assume that the chromosomal position of interest is between markers j and $j + 1$. The transition matrix between j and $j + 1$ then becomes

$$T(\theta_{j,j+1}) = T(\theta_{j,x}) T(\theta_{x,j+1}) = T(\theta_{j,x}) I T(\theta_{x,j+1}) = T(\theta_{j,x}) Q_x T(\theta_{x,j+1})$$

by using the Chapman–Kolmogorov equations [120], which assume the availability of a multilocus feasible map function. Here, $T(\theta_{j,x})$ and $T(\theta_{x,j+1})$ denote the transition matrices from marker j to position x and from x to marker $j + 1$. Furthermore, I denotes the $2^b \times 2^b$ identity matrix that is identical to Q_x because no marker information is available at position x . Therefore, we obtain for $P_{\text{complete},x}(v)$ using Eq. (A.15):

$$P_{\text{complete},x} = \frac{[1^T Q_1 T(\theta_{1,2}) Q_2 \dots T(\theta_{j,x})]^T \circ T(\theta_{x,j+1}) \dots Q_{J-1} T(\theta_{J-1,J}) Q_J 1}{1^T Q_1 T(\theta_{1,2}) Q_2 \dots T(\theta_{j,x}) T(\theta_{x,j+1}) \dots Q_{J-1} T(\theta_{J-1,J}) Q_J 1}$$

If the inheritance distribution $P(V(x) = v)$ has to be determined for x outside of the interval of all markers $1-J$, either the transition matrix $T(\theta_{x,1})$ or $T(\theta_{J,x})$ is required. If x is left of marker 1, $P_{\text{complete},x}(v)$ is given by

$$P_{\text{complete},x} = \frac{1 \circ T(\theta_{x,1}) Q_1 T(\theta_{1,2}) Q_2 \dots Q_{J-1} T(\theta_{J-1,J}) Q_J 1}{1^T T(\theta_{x,1}) Q_1 T(\theta_{1,2}) Q_2 \dots Q_{J-1} T(\theta_{J-1,J}) Q_J 1}.$$

A similar formula is obtained for a chromosomal position on the right side of the markers.

If the chromosomal position x and the genetic markers are unlinked, one obtains $P(V(x) = v) = P_{\text{a priori}} = 1/2^b$, which is easily seen from the transition matrix: $T(\theta_{x,1})$ at $\theta_{x,1} = \frac{1}{2}$. In this case, all entries of T have the same value $1/2^b$ because $T_{v,v_1}(\theta_{x,1}) = (\frac{1}{2})^{H(v,v_1)} (1 - \frac{1}{2})^{b-H(v,v_1)} = \frac{1}{2^b}$. If an arbitrary column vector

is multiplied from the right to this matrix, the resulting inheritance distribution has components with identical values for all possible inheritance vectors.

Finally, we note that the difference between $P_{\text{marker } j}(v)$ and $P(V(j) = v) = P_{\text{complete}, j}(v)$ at x can be interpreted as follows. $P_{\text{marker } j}(v)$ represents the inheritance distribution at marker j given only the genotypes at marker j . In contrast, $P_{\text{complete}, j}(v)$ represents the inheritance distribution at the same marker but now containing the information from all available markers. In this multi-marker or multipoint approach, genotype data from a linked genetic marker influence the inheritance distribution at a chromosomal position. Because differences between the inheritance distributions at two loci decrease with the recombination fraction between them, missing inheritance information at a marker can be compensated by neighboring genetic markers (see Problem 7.6). This is termed the *spillover* of inheritance information.

A.3 WHAT IS THE CARDON–FULKER ALGORITHM?

In this section, the basic idea of the Cardon–Fulker approach is described. For this, we work with a single sib-pair family but several chromosomal positions. Therefore, we drop the family index in the following and use a slightly different notation compared to the other sections in this chapter. $\hat{\tau}_j$ now denotes the observed proportion of alleles shared IBD by a single sib-pair at marker M_j , and τ_t is the proportion of alleles shared IBD at the chromosomal position of interest.

We start by considering two genetic markers M_j and $M_{j'}$ as well as a chromosomal position of interest M_t in the interval M_j – $M_{j'}$. Recall that $\mathbb{E}(\hat{\tau}_j) = \frac{1}{2}$ and $\text{Var}(\hat{\tau}_j) = \frac{1}{8}$ under the null hypothesis of no linkage. We estimate the IBD proportion τ_t at position t using linked markers in a simple linear regression. For this, we require the following formulae ([183]; see Problem A.1)

$$\begin{aligned} \text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'}) &= 8 \text{Var}(\hat{\tau}_j) \text{Var}(\hat{\tau}_{j'}) (1 - 2\theta_{jj'})^2 = \frac{1}{8} (1 - 2\theta_{jj'})^2, \\ \text{Cov}(\hat{\tau}_j, \tau_t) &= \text{Var}(\hat{\tau}_j) (1 - 2\theta_{jt})^2 = \frac{1}{8} (1 - 2\theta_{jt})^2, \end{aligned}$$

where $\theta_{jj'}$ denotes the recombination fraction between markers M_j and $M_{j'}$. Note that τ_t is the probability of interest, and thus is considered fixed, while the other probabilities are estimated from the data. τ_t is estimated from the data as a linear function of $\hat{\tau}_j$ and $\hat{\tau}_{j'}$ by

$$\hat{\tau}_t = \hat{\alpha} + \hat{\beta}_j \hat{\tau}_j + \hat{\beta}_{j'} \hat{\tau}_{j'}, \quad (\text{A.16})$$

and the regression coefficients are the solutions of

$$\begin{pmatrix} \hat{\beta}_j \\ \hat{\beta}_{j'} \end{pmatrix} = \begin{pmatrix} \text{Var}(\hat{\tau}_j) & \text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'}) \\ \text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'}) & \text{Var}(\hat{\tau}_{j'}) \end{pmatrix}^{-1} \begin{pmatrix} \text{Cov}(\hat{\tau}_j, \tau_t) \\ \text{Cov}(\hat{\tau}_{j'}, \tau_t) \end{pmatrix}. \quad (\text{A.17})$$

Subsequently, the regression coefficients are [229]

$$\begin{aligned}\hat{\beta}_j &= \frac{(1 - 2\theta_{jt})^2 - (1 - 2\theta_{tj'})^2(1 - 2\theta_{jj'})^2}{1 - (1 - 2\theta_{jj'})^4}, \\ \hat{\beta}_{j'} &= \frac{(1 - 2\theta_{tj'})^2 - (1 - 2\theta_{jt})^2(1 - 2\theta_{jj'})^2}{1 - (1 - 2\theta_{jj'})^4}, \\ \hat{\alpha} &= (1 - \hat{\beta}_j - \hat{\beta}_{j'})/2.\end{aligned}$$

The following example has been taken from Ref. [229] and illustrates the simplicity of this approach.

Example A.2. Tables A.2 and A.3 display estimated IBD proportions $\hat{\tau}_t$ for five chromosomal positions of interest assuming equally spaced intervals. The distance between the flanking markers is assumed to be 20 and 100 cM, respectively. Conversions from map distances to recombination fractions have been done using the Haldane map function. It can be seen that when the position of t is identical to that of a marker, $\hat{\tau}_t$ is identical to that of the marker. If t is different from a marker position, $\hat{\tau}_t$ is similar to a weighted average of $\hat{\tau}_j$ and $\hat{\tau}_{j'}$ with weights proportional to the distance between markers and t . Table A.3 shows that the midpoint, which is 50cM, thus corresponding to $\theta \approx 0.46$, always tends to 0.50.

Table A.2 Interval estimates for τ_t as derived from $\hat{\tau}_j$ and $\hat{\tau}_{j'}$ assuming a sib-pair with unambiguous identical by descent status at both marker loci.

$\hat{\tau}_j$	$\hat{\tau}_{j'}$	Location				
		1	2	3	4	5
1	1	1.00	0.97	0.96	0.97	1.00
1	$\frac{1}{2}$	1.00	0.86	0.73	0.61	0.50
1	0	1.00	0.75	0.50	0.25	0.00
$\frac{1}{2}$	1	0.50	0.61	0.73	0.86	1.00
$\frac{1}{2}$	$\frac{1}{2}$	0.50	0.50	0.50	0.50	0.50
$\frac{1}{2}$	0	0.50	0.39	0.27	0.14	0.00
0	1	0.00	0.25	0.50	0.75	1.00
0	$\frac{1}{2}$	0.00	0.14	0.27	0.39	0.50
0	0	0.00	0.03	0.04	0.03	0.00
$\hat{\beta}_j$		1.00	0.72	0.46	0.23	0.00
$\hat{\beta}_{j'}$		0.00	0.23	0.46	0.72	1.00

The flanking markers j and j' are separated by 20 cM. Mappings have been done with the Haldane map function assuming equal spacings for the chromosomal positions of interest. Table adopted from Ref. [229], Table 1.

Table A.3 Interval estimates for τ_t as derived from $\hat{\tau}_j$ and $\hat{\tau}_{j'}$ assuming a sib-pair with unambiguous identical by descent status at both marker loci.

$\hat{\tau}_j$	$\hat{\tau}_{j'}$	Location				
		1	2	3	4	5
1	1	1.00	0.71	0.63	0.71	1.00
1	$\frac{1}{2}$	1.00	0.68	0.57	0.52	0.50
1	0	1.00	0.66	0.50	0.34	0.00
$\frac{1}{2}$	1	0.50	0.52	0.57	0.68	1.00
$\frac{1}{2}$	$\frac{1}{2}$	0.50	0.50	0.50	0.50	0.50
$\frac{1}{2}$	0	0.50	0.48	0.43	0.32	0.00
0	1	0.00	0.34	0.50	0.66	1.00
0	$\frac{1}{2}$	0.00	0.32	0.43	0.48	0.50
0	0	0.00	0.29	0.37	0.29	0.00
	$\hat{\beta}_j$	1.00	0.37	0.13	0.04	0.00
	$\hat{\beta}_{j'}$	0.00	0.04	0.13	0.37	1.00

The flanking markers j and j' are unlinked and separated by 100 cM. Mappings have been done with the Haldane map function assuming equal spacings for the chromosomal positions of interest. Table adopted from Ref. [229], Table 2.

When terms involving powers and products are ignored, the regression coefficients can be approximated for small inter-marker distances, i.e., $\theta_{jj'} \leq 0.1$:

$$\hat{\beta}_j \approx \frac{\theta_{jt}}{\theta_{jj'}} \quad \hat{\beta}_{j'} \approx \frac{\theta_{j't}}{\theta_{jj'}}$$

For the general case of J markers, the sib-pair approach is easily generalized by using $\hat{\tau}_t = \hat{\alpha} + \sum_{j=1}^J \hat{\beta}_j \hat{\tau}_j$ instead of Eq. (A.16) and the corresponding generalization of the regression estimator Eq. (A.17). Because of its simplicity, robustness, and ease of computation, the Fulkner–Cardon method has been extended to arbitrary unilineal relationships by Almasy and Blangero [18].

A.4 PROBLEM

Problem A.1 Show the validity of the formula $\text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'}) = \frac{1}{8}(1 - 2\theta_{jj'})^2$.

Solutions to Study Problems

Solution 1.1

Solution 1.1.1. The structure of DNA and RNA can be described as follows:

1. Backbone of sugar and phosphate residues, bases adenine and guanine (purines), cytosine and thymine (pyrimidines)
2. Hydrogen bonding between two opposing bases of the two strands: thymine with adenine (two hydrogen bonds), cytosine with guanine (three hydrogen bonds)
3. RNA is smaller, single-stranded, and contains uracil instead of thymine and ribose instead of deoxyribose

Solution 1.1.2. The terms are defined as follows:

1. *Nucleotide* = one nucleoside plus phosphate; *nucleoside* = one sugar plus attached base
2. *Diploid* = state of a cell with a double set of chromosomes; *haploid* = state of a cell with a single set of chromosomes
3. *Gene* = functional unit of DNA encoding a product; *genotype* = overall genetic constitution of an individual; *phenotype* = observable characteristics of an individual
4. *Introns* = noncoding DNA that is transcribed and then removed from RNA by splicing; *exons* = DNA that is transcribed and represented in mRNA; *promoters* = DNA segments including binding sites for the RNA polymerases initiating the transcription

Solution 1.1.3. The translation of the DNA strand is

1. Complementary DNA strand

A T G T T A C T A G A C T G C T A A

2. RNA strand after transcription

A U G U U A C U A G A C U G C U A A

3. Result of translation

Start (Met) Leu Leu Asp Cys Termination

Solution 1.2

Solution 1.2.1. The process of meiosis is described in detail in Section 1.2.

Solution 1.2.2. For the different situations, recombinations are observed as follows:

1. *Recombinations* between *A* and *C*, *A* and *D*, *A* and *E*, *B* and *C*, *B* and *D*, and *B* and *E*; *no recombination* between *A* and *B* and between *C*, *D*, and *E*
2. *Recombinations* between *A* and *B*, *A* and *C*, *B* and *D*, *B* and *E*, *C* and *D*, and *C* and *E*; *no recombination* between *A*, *D*, and *E*

Solution 1.3

Solution 1.3.1. The terms are defined as follows:

1. *Locus* = specific DNA segment defined by its sequence of bases; *allele* = variant at a locus
2. *Homozygous* = state of an individual at a locus showing the same alleles; *heterozygous* = state of an individual at a locus showing different alleles
3. *Mutation* = state that permanently changes the function of a gene; *polymorphism* = variation with a frequency of at least 1% in at least one human population

Solution 1.3.2. The effects of modifying the original template strand are

1. Insertion of the single base adenine at position 8:

New strand: T A C A A T G A A T C T G A C G A T T

Complementary DNA strand: A T G T T A C T T A G A C T G C T A A

Complementary RNA strand: A U G U U A C U U A G A C U G C U A A

Result of translation: Start (Met) Leu Leu Arg Leu Leu

2. Substitution of the single base thymine at position 9 with the base adenine:

New strand: T A C A A T G A A C T G A C G A T T

Complementary DNA strand: A T G T T A C T T G A C T G C T A A

Complementary RNA strand: A U G U U A C U U G A C U G C U A A

Result of translation: Start (Met) Leu Leu Asp Cys Termination

3. Deletion of bases adenine, thymine, and cytosine at positions 8, 9, and 10, respectively:

New strand: T A C A A T G T G A C G A T T

Complementary DNA strand: A T G T T A C A C T G C T A A

Complementary RNA strand: A U G U U A C A C U G C U A A

Result of translation: Start (Met) Leu His Cys Termination

Solution 1.3.3. The techniques are described in depth in Section 1.3.2.

Solution 1.4 For $n = 1000$, the probabilities of not observing both variants is 0.9048 when only homozygous subjects are observed. It is 0.8187 under HWE. With $n = 10,000$, the two probabilities are 0.3679 and 0.1353, and for $n = 20,000$ they are 0.1353 and 0.0183. Finally, for $n = 50,000$ they are 0.0067 and $4.5377 \cdot 10^{-5}$. Thus, only a sample size exceeding 20,000 is sufficient for detecting both variants with 99% probability when HWE holds.

Solution 2.1

Solution 2.1.1. X-chromosomal recessive. However, autosomal recessive cannot be excluded.

Solution 2.1.2. Autosomal dominant with reduced penetrance because great-grandparents are not affected.

Solution 2.1.3. X-chromosomal recessive. Again, autosomal recessive cannot be excluded. Autosomal dominant would be possible with a reduced penetrance in females.

Solution 2.1.4. X-chromosomal recessive.

Solution 2.1.5. Autosomal dominant with reduced penetrance or gender-specific expression.

Solution 2.2

Solution 2.2.1.

1. *Penetrance* = conditional probability of being affected with a disease given a specific genotype; *phenocopy* = affection without carriage of the disease-causing genotype

2. *Variable expressivity* = variation in the degree of disease manifestation; *anticipation* = special case of variable expressivity where the severity of the expression becomes greater and manifests earlier in successive generations; *imprinting* = expression depending on origin of disease-causing allele
3. *Genetic heterogeneity* = different mutations cause the same disease in different individuals; *pleiotropy* = the same mutation causes variable symptoms or unrelated effects; *epistasis* = interaction between different loci

Solution 2.2.2.

1. Variable expressivity
2. Incomplete dominance, probably additive genetic effect
3. Genetic heterogeneity

Solution 2.2.3. With the frequency of sporadic cases in the general population of about 1 in 10,000, $f_0 = 0.0001$. The penetrance in homozygous aa individuals is $f_2 = 0.98$. Distinguishing between paternal and maternal origin of the a allele, $f_{1P} = 0.65$ and $f_{1M} = 0.35$. Hence, $\gamma_2 = 0.98/0.0001 = 9800$, $\gamma_{1P} = 0.65/0.0001 = 6500$, and $\gamma_{1M} = 0.35/0.0001 = 3500$.

Solution 2.3**Solution 2.3.1.**

- No random mating in the investigated population
- Selection or migration
- Mutations
- Population stratification
- Small population size

Solution 2.3.2. Following Hardy–Weinberg equilibrium, with a diallelic disease locus, the frequency of individuals carrying at least one disease-causing allele p_1 is $p_1^2 + 2 \cdot p_1 \cdot p_2$, which is equal 0.01. Because $p_2 = 1 - p_1$, this equation can be solved for p_1 and renders $p_1 = 0.005$, which is the frequency of the disease-causing allele. Given an autosomal dominant inheritance pattern, the frequency of carriers equals the prevalence, which is 0.01.

Solution 3.1

1. Most commonly used are short tandem repeats (STRs or microsatellites) and single nucleotide polymorphisms (SNPs).
2. SNPs are more frequent and less informative than STRs. The mutation frequency of STRs is higher, and the ease of ascertainment is lower than in SNPs.
3. The degree of functionality of STRs is somewhat unclear. For SNPs, the functionality depends on the exact location and the specific kind of base substitution (see Table 3.3).

Solution 3.2 For marker *A*, the measures of informativity are $HET = 1 - 4 \cdot 0.25^2 = 0.7500$, $LIC = \frac{(4-1)(2 \cdot 4^2 - 1)}{2 \cdot 4^3} = 0.7266$, and $PIC = 1 - 4 \cdot 0.25^2 - 12 \cdot 0.25^4 = 0.7031$. For marker *B*, they are

$$\begin{aligned} HET &= 1 - (0.8^2 + 0.07^2 + 0.07^2 + 0.06^2) = 0.3466, \\ LIC &= 1 - (0.8^2 + 0.07^2 + 0.07^2 + 0.06^2) \\ &\quad + \frac{1}{2}(0.8^4 + 0.07^4 + 0.07^4 + 0.06^4) \\ &\quad - \frac{1}{2}(0.8^2 + 0.07^2 + 0.07^2 + 0.06^2)^2 = 0.3380, \\ PIC &= 1 - (0.8^2 + 0.07^2 + 0.07^2 + 0.06^2) \\ &\quad - (0.8^2 \cdot 0.07^2 + 0.8^2 \cdot 0.07^2 + 0.8^2 \cdot 0.06^2 \\ &\quad + 0.07^2 \cdot 0.8^2 + 0.07^2 \cdot 0.07^2 + 0.07^2 \cdot 0.06^2 \\ &\quad + 0.07^2 \cdot 0.8^2 + 0.07^2 \cdot 0.07^2 + 0.07^2 \cdot 0.06^2 \\ &\quad + 0.06^2 \cdot 0.8^2 + 0.06^2 \cdot 0.07^2 + 0.06^2 \cdot 0.07^2) = 0.3293. \end{aligned}$$

Solution 3.3 In the first step, the number m of equifrequent alleles that corresponds to a PIC of 0.6 is calculated to be $m \approx 3.0508$. In the second step, m is used to compute LIC with $LIC = \frac{(3.0508-1)(2 \cdot 3.0508^2 - 1)}{2 \cdot 3.0508^3} \approx 0.6361$.

Solution 4.1

Solution 4.1.1. The pedigree is not compatible with Mendelian laws. The genotype “132 134” is not possible in Lucy. One possible reason is that a genotyping error has occurred at that marker, maybe caused by stutter bands.

Solution 4.1.2. The inconsistency with Mendelian inheritance could also be resolved if Lucy is a daughter of Brutus, not of Popeye.

Solution 4.2 The pedigree is not compatible with Mendelian inheritance. Kelly is homozygous at the second marker with allele 146, though her father Al Bundy was heterozygous with alleles 128 and 134. It is very likely that the paternally transmitted allele of the second marker has not been detected.

Solution 4.3

Solution 4.3.1. The pedigree is compatible with Mendelian inheritance.

Solution 4.3.2. Though the pedigree is compatible with Mendelian inheritance, further investigation is needed. Marcie’s genotypes could well include a double recombinant although she is heterozygous at all three marker loci. If the displayed genetic markers are given in their order on a single chromosome and if the distance between the markers is low, a genotyping error becomes more plausible.

Solution 4.3.3. Peggy’s haplotypes cannot be determined, and Kelly’s and Marcie’s genotypes are plausible even if the markers are tightly linked. Single recombinations in Kelly and Marcie can explain the observed genotypes even if the markers are closely linked. The observed genotypes are more plausible than those of Figure 4.20 because the genotype pattern can be explained by two single recombinations. Because of positive interference (see Chapter 5), two single recombinations are more plausible than a double recombination.

Solution 4.3.4. If the markers are closely linked and given in the correct order, the pedigree is plausible if Peggy has, e.g., haplotypes $\begin{array}{c|c} 130 & 133 \\ 130 & 133 \\ 133 & 130 \end{array}$ and if there is a single recombination in both Kelly and Marcie.

Solution 4.4 The allele frequency is estimated as $\hat{p} = \widehat{P(A_1)} = (2 \cdot 20 + 40)/(2 \cdot 140) \approx 0.2857$ and $\hat{q} = \widehat{P(A_2)} = 1 - \hat{p} \approx 0.7143$. In the next step, the expected genotype probabilities assuming HWE can be computed (see Table S.1). For example, $\hat{P}_{HWE}(A_1A_1) = \widehat{P(A_1)}^2 \approx 0.2857^2 \approx 0.0816$. This in turn can be used to determine the expected genotype distribution under HWE (see Table S.1). For example, the expected number of A_1A_1 genotypes is $n \cdot \hat{P}_{HWE}(A_1A_1) \approx 140 \cdot 0.0816 \approx 11.4286$. Subsequently, the χ^2 statistic of the HWE goodness-of-fit test is $\chi_{HW}^2 \approx \frac{(20-11.4286)^2}{11.4286} + \frac{(40-57.1429)^2}{57.1429} + \frac{(80-71.4286)^2}{71.4286} \approx 12.6000$. This corresponds to a p -value of 0.0004 using the χ^2 distribution with 1 d.f. The conclusion is that HWE is violated.

Table S.1 Observed and expected genotype distribution for the data of Problem 4.4.

	Genotype			Σ
	A_1A_1	A_1A_2	A_2A_2	
Observed count	20	40	80	140
Expected genotype probabilities	0.0816	0.4082	0.5102	
Expected count	11.4286	57.1429	71.4286	140

Solution 4.5 The allele frequency is estimated as $\hat{p} = \widehat{P(A_1)} = (2 \cdot 32 + 52)/(2 \cdot 150) \approx 0.4333$ and $\hat{q} = \widehat{P(A_2)} = 1 - \hat{p} \approx 0.5667$. The expected genotype probabilities assuming HWE are obtained by $\hat{P}_{HWE}(A_1A_1) = \widehat{P(A_1)}^2 \approx 0.4333^2 \approx 0.1878$ (see Table S.2). The expected number of A_1A_1 genotypes is $n \cdot \hat{P}_{HWE}(A_1A_1) \approx 150 \cdot 0.1878 \approx 28.1667$ (see Table S.2). Finally, the χ^2 statistic of the HWE goodness-of-fit test yields 1.6247, corresponding to $p \approx 0.2024$ using the χ_1^2 distribution. The data do not provide evidence for a deviation from HWE.

Table S.2 Observed and expected genotype distribution for the data of Problem 4.5.

	Genotype			Σ
	A_1A_1	A_1A_2	A_2A_2	
Observed count	32	66	52	150
Population frequency	0.1878	0.4911	0.3211	1
Expected count	28.1667	73.6667	48.1667	150

The estimated REH is $\widehat{\text{REH}} = 0.8090$. The asymptotic 95% confidence interval (CI) for the REH is $[0.5836; 1.1215]$. The recommended equivalence margin reaches from 0.7143 to 1.4. The CI for REH is not completely included in the equivalence margin. Thus, compatibility with HWE cannot be established.

Solution 4.6 The number of required simulations to estimate a p -value with a precision of $e = 0.01$ at the 95% confidence level is approximately $N \approx \left(\frac{1.96}{2 \cdot 0.01}\right)^2 \approx 9604 \approx 10,000$.

Solution 4.7 The number of required simulations to estimate a p -value around 0.05 with a precision of $e = 0.01$ at the 99% confidence level is approximately $N \approx \left(\frac{1.96 \cdot 0.05(1-0.95)}{0.01}\right)^2 \approx 3162 \approx 4000$.

Solution 4.8 In a sample of n subjects, the estimators of the genotype frequencies are $\hat{p}_{11} = n_{11}/n$ and $\hat{p}_{12} = n_{12}/n$. Using the standard results for the multinomial distribution, their variances and the covariance are $\text{Var}(\hat{p}_{11}) = p_{11}(1 - p_{11})/n$, $\text{Var}(\hat{p}_{12}) = p_{12}(1 - p_{12})/n$, and $\text{Cov}(\hat{p}_{11}, \hat{p}_{12}) = -p_{11}p_{12}/n$.

Solution 4.8.1. The estimator of the allele frequency $p = P(A_1)$ is given by $\hat{p} = \hat{p}_{11} + \hat{p}_{12}/2 = (2n_{11} + n_{12})/(2n)$, and its expected value is $\mathbb{E}(\hat{p}) = \mathbb{E}(\hat{p}_{11}) + \frac{1}{2}\mathbb{E}(\hat{p}_{12}) = p_{11} + p_{12}/2 = p$. The variance is

$$\begin{aligned}\text{Var}(\hat{p}) &= \text{Var}(\hat{p}_{11}) + \text{Var}(\hat{p}_{12}/2) + 2\text{Cov}(\hat{p}_{11}, \hat{p}_{12}/2) \\ &= \frac{p_{11}(1 - p_{11})}{n} + \frac{p_{12}(1 - p_{12})}{4n} - \frac{p_{11}p_{12}}{n},\end{aligned}$$

which gives the desired result if the variance is expressed in terms of p and p_{11} using $p_{12} = 2(p - p_{11})$.

Solution 4.8.2. The general expression $\text{Var}(\hat{p}) = \frac{p(1-p)}{2n} + \frac{p_{11}-p^2}{2n} = \frac{p(1-p)}{2n} + \frac{\mathcal{D}_{A_1}}{2n}$ reduces to the desired result because of HWE if and only if $\mathcal{D}_{A_1} = 0$.

Solution 4.8.3. If $p_{11} = 0$, $p = p_{12}/2$ so that $\text{Var}(\hat{p}) = p_{12}(1 - p_{12})/(4n) = 2p(1 - p)/(4n) = (p(1 - p) - p^2)/(2n)$. This variance is lower than the variance of $\hat{p}$ when HWE is fulfilled.

Solution 4.9 Let $\mathbf{p} = (p_{11}, p_{12}, p_{22})'$. The central limit theorem implies that $\hat{\mathbf{p}} = (\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{22})'$ are jointly asymptotically normal. The function $\kappa = \kappa(\mathbf{p}) = \ln(p_{12}/2\sqrt{p_{11}p_{22}})$ is continuously differentiable with respect to $\mathbf{p}$ in a neighborhood of the true genotype probability. According to the multivariate delta method, $\hat{\kappa} = \kappa(\hat{\mathbf{p}})$ is asymptotically normal, in detail $\hat{\kappa} = \ln \hat{\omega} \stackrel{a}{\sim} N(\kappa = \ln \omega, \sigma_{\ln \omega}^2)$, with

$$\sigma_{\ln \omega}^2 = \frac{1}{n} \frac{\partial \kappa(\mathbf{p})}{\partial \mathbf{p}'} \text{Var}(\sqrt{n}\hat{\mathbf{p}}) \frac{\partial \kappa(\mathbf{p})}{\partial \mathbf{p}} = \frac{1}{n} \boldsymbol{\zeta}' \boldsymbol{\Sigma}_p \boldsymbol{\zeta}.$$

To derive $\sigma_{\ln \omega}^2$, we require the derivatives of κ with respect to $\mathbf{p}$, and these are $\boldsymbol{\zeta} = \partial \kappa / \partial \mathbf{p} = \left(-\frac{1}{2p_{11}}, \frac{1}{p_{12}}, -\frac{1}{2p_{22}}\right)'$. Using the standard result for the covariance

matrix $\Sigma_p = \text{Var}(\sqrt{n}\hat{\mathbf{p}})$ of $\sqrt{n}\hat{\mathbf{p}}$ from the multinomial distribution, we can write $\Sigma_p = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}'$, where

$$\text{diag}(\mathbf{p}) = \begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 0 & p_{22} \end{pmatrix}.$$

By making the appropriate substitutions in $\sigma_{\ln \omega}^2 = \frac{1}{n} \boldsymbol{\zeta}' \Sigma_p \boldsymbol{\zeta}$, the desired result is obtained.

Solution 4.10 In the general case, $\sigma_{\ln \hat{\omega}}^2 = \frac{1}{n} \left(\frac{1-p_{12}}{4p_{11}p_{22}} + \frac{1}{p_{12}} \right)$ according to Eq. (4.8). Furthermore, we can write $p_{12} = 2\sqrt{p_{11}}(1 - \sqrt{p_{11}})$ under HWE because under HWE $p_{11} = (p_{11} + p_{12}/2)^2 \Rightarrow p_{12} = 2\sqrt{p_{11}}(1 - \sqrt{p_{11}})$. Using this result we obtain after doing some algebra

$$\sigma_{\ln \hat{\omega}}^2 = \sigma_{\ln \hat{p}_{11}}^2 = \frac{1}{n} (2\sqrt{p_{11}}(1 - \sqrt{p_{11}}))^{-2} = \frac{1}{n} \xi^2$$

for $\xi = (2\sqrt{p_{11}}(1 - \sqrt{p_{11}}))^{-1}$ so that $\ln \hat{\omega} \stackrel{a}{\sim} N(0, \frac{1}{n} \xi^2)$ under HWE.

We consider $\ln \omega$ instead of ω . Thus, for the hypothesis $H: \omega \leq 1 - \varepsilon_1 \vee \omega \geq 1 + \varepsilon_2$ the equivalence boundaries are $\ln(1 - \varepsilon_1)$ and $\ln(1 + \varepsilon_2)$. The test of H , whose power against $\omega = 1$ we are interested in, rejects H if and only if

$$\ln \hat{\omega} - z_{1-\alpha} \sigma_{\ln \hat{p}_{11}} > \ln(1 - \varepsilon_1) \wedge \ln \hat{\omega} + z_{1-\alpha} \sigma_{\ln \hat{p}_{11}} < \ln(1 + \varepsilon_2).$$

With $\tilde{\varepsilon}_1 = -\ln(1 - \varepsilon_1)$ and $\tilde{\varepsilon}_2 = \ln(1 + \varepsilon_2)$, this event can be written as

$$\begin{aligned} z_{1-\alpha} \sigma_{\ln \hat{p}_{11}} - \tilde{\varepsilon}_1 &< \ln \hat{\omega} < -z_{1-\alpha} \sigma_{\ln \hat{p}_{11}} + \tilde{\varepsilon}_2 \\ z_{1-\alpha} \frac{\xi}{\sqrt{n}} - \tilde{\varepsilon}_1 &< \ln \hat{\omega} < -z_{1-\alpha} \frac{\xi}{\sqrt{n}} + \tilde{\varepsilon}_2 \\ z_{1-\alpha} - \frac{\sqrt{n}}{\xi} \tilde{\varepsilon}_1 &< Z = \sqrt{n} \frac{\ln \hat{\omega}}{\xi} < -z_{1-\alpha} + \frac{\sqrt{n}}{\xi} \tilde{\varepsilon}_2. \end{aligned}$$

Here, Z indicates an asymptotically standard normally distributed random variable. Therefore, the event has a probability that converges to

$$\Phi\left(\frac{\sqrt{n}}{\xi} \tilde{\varepsilon}_2 - z_{1-\alpha}\right) - \Phi\left(z_{1-\alpha} - \frac{\sqrt{n}}{\xi} \tilde{\varepsilon}_1\right) = \Phi\left(\frac{\sqrt{n}}{\xi} \tilde{\varepsilon}_2 - z_{1-\alpha}\right) + \Phi\left(\frac{\sqrt{n}}{\xi} \tilde{\varepsilon}_1 - z_{1-\alpha}\right),$$

and this probability is the type II error β , thus $\beta = 2\Phi\left(\frac{\sqrt{n}}{\xi} \tilde{\varepsilon} - z_{1-\alpha}\right)$ for an equivalence margin $\pm \varepsilon$. By solving this equation for n and by replacing ξ , the sample size formula is obtained.

Solution 5.1

Solution 5.1.1. $\theta_{AB} = 0.05, \Rightarrow x_{AB} \approx 0.0527 \text{ M} = 5.27 \text{ cM}$.

Solution 5.1.2. $x_{AB} = 2 \text{ cM} = 0.02 \text{ M}, \Rightarrow \theta_{AB} \approx 0.0196$.

Solution 5.1.3. $\theta_{AB} = 0.05, \theta_{BC} = 0.085 \Rightarrow x_{AB} = 0.0527 \text{ M}, x_{BC} = 0.0932 \text{ M} \Rightarrow x_{AC} = 0.1459 \Rightarrow \theta_{AC} = 0.1265$

Solution 5.1.4. $\theta_{AB} = 0.05, \theta_{AC} = 0.085, \Rightarrow x_{AB} = 0.0405 \Rightarrow \theta_{AB} = 0.0389$.

Solution 5.2 The four conditions for connecting two-point recombination fractions θ with joint probabilities are

$$\theta_{AB} = g_{11} + g_{10}, \quad \theta_{BC} = g_{11} + g_{01}, \quad \theta_{AC} = g_{01} + g_{10}, \quad \text{and} \quad \sum_{i,j} g_{ij} = 1.$$

Note that θ_{AC} describes the event that a recombination is observed either in the interval A – B (without a recombination between B and C) or in the interval B – C (without a recombination between A and B). Straightforward inversion leads to

$$\begin{aligned} g_{11} &= \frac{1}{2}(\theta_{AB} + \theta_{BC} - \theta_{AC}) && \text{with restriction} && \theta_{AC} \leq \theta_{AB} + \theta_{BC}, \\ g_{10} &= \frac{1}{2}(\theta_{AB} - \theta_{BC} + \theta_{AC}) && \text{with restriction} && \theta_{BC} \leq \theta_{AB} + \theta_{AC}, \\ g_{01} &= \frac{1}{2}(-\theta_{AB} + \theta_{BC} + \theta_{AC}) && \text{with restriction} && \theta_{AB} \leq \theta_{BC} + \theta_{AC}, \\ g_{00} &= 1 - g_{11} - g_{10} - g_{01} = 1 - \frac{1}{2}(\theta_{AB} + \theta_{BC} + \theta_{AC}). \end{aligned}$$

Solution 5.3

Solution 5.3.1. We have $n = 3$ and $\epsilon = (1, 1)'$. Four combinations are possible leading to

$$\Delta' = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

First, we compute all s_i : $s_1 = \delta'_1 \epsilon = (1 \ 1)' \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2$. Similarly, we obtain $s_2 = s_3 = 1$ and $s_4 = 0$. With $\theta_{\delta_4} = 0$ denoting the recombination event of $\delta'_4 = (0, 0)$, we finally have

$$\begin{aligned} P(\epsilon) = P((1, 1)') &= \frac{1}{4} \sum_{i=1}^4 (-1)^{s_i} (1 - 2r_{\delta_i}) \\ &= \frac{1}{4} \left((-1)^2 (1 - 2r_{\delta_1}) + (-1)^1 (1 - 2r_{\delta_2}) \right. \\ &\quad \left. + (-1)^1 (1 - 2r_{\delta_3}) + (-1)^0 (1 - 2r_{\delta_4}) \right) \\ &= \frac{1}{2} (-r_{\delta_1} + r_{\delta_2} + r_{\delta_3} - r_{\delta_4}) = \frac{1}{2} (\theta_{AB} + \theta_{BC} - \theta_{AC}). \end{aligned}$$

Solution 5.3.2. In this exercise, $n = 4$ and $\epsilon = (1, 1, 1)'$. Eight combinations are possible, yielding

$$\Delta' = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

The corresponding s_i are $s_1 = 3$, $s_2 = s_3 = s_4 = 2$, and $s_5 = s_6 = s_7 = 1$. Finally,

$$\begin{aligned} P(\epsilon) &= P((1, 1, 1)') = \frac{1}{8} \sum_{i=1}^4 (-1)^{s_i} (1 - 2r_{\delta_i}) \\ &= \frac{1}{4} (r_{\delta_1} - r_{\delta_2} - r_{\delta_3} - r_{\delta_4} + r_{\delta_5} + r_{\delta_6} + r_{\delta_7} - r_{\delta_8}) \\ &= \frac{1}{4} (\theta_{AD} - \theta_{AC} - \theta_{AB,CD} - \theta_{BD} + \theta_{CD} + \theta_{BC} + \theta_{AB}) . \end{aligned}$$

Solution 5.3.3. Given are $\theta_{AB} = 0.1$, $\theta_{BC} = 0.05$, and $\theta_{CD} = 0.2$. Find $P((1, 1, 1)')$ using the Kosambi map function. Recombination fractions θ and corresponding map distances are given in Table S.3. Applying the result from Problem 5.3.2 yields $P((1, 1, 1)') = \frac{1}{4} (0.1 + 0.05 + 0.2 - 0.147 - 0.24 - 0.278 + 0.31) = -0.0012$.

Solution 5.4 The third derivative of the Kosambi map function is

$$M^{(3)}(x) = \frac{e^{4x}(64 - 256e^{4x} + 64e^{8x})}{(1 + e^{4x})^4},$$

yielding $M^{(3)}(0) = -8$. Because $(-1)^3 M^{(3)}(0) = +8 > 0$, the Kosambi map function is not multilocus feasible.

Solution 6.1 Without loss of generality, we assume $\mu = 0$. The population mean is then given by $p^2a + 2pq \cdot 0 + q^2(-a) = (p^2 - q^2)a = (p + q)(p - q)a = (p - q)a$ because $p + q = 1$. The phenotypic variance of individual X that is due to the trait locus subsequently is $\sigma_g^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = (p^2a^2 + 2pq \cdot 0^2 + q^2(-a)^2) - ((p - q)a)^2$. By solving the binomial and factorizing a^2 , we furthermore obtain $\sigma_g^2 = (p^2 + q^2 - p^2 + 2pq - q^2)a^2 = 2pqa^2$.

Solution 6.2 Starting point is the linear model $y_o = \mu_o + \beta(y_p - \mu_p) + \varepsilon$, where μ_o and μ_p are the general means in offspring and parents, respectively. β is the regression coefficient of interest, and ε is the error term. β summarizes the effect of the parental phenotype on the offspring's trait value.

The expected value of the offspring's phenotype given the parental trait value is $\mathbb{E}(y_o|y_p) = \mu_o + \beta(y_p - \mu_p)$, and the variance of this conditional mean is given by $\mathbb{V}\text{ar}(\mathbb{E}(y_o|y_p)) = \mathbb{V}\text{ar}(\beta(y_p - \mu_p)) = \beta^2 \mathbb{V}\text{ar}(y_p)$.

For calculating the total variance $\mathbb{V}\text{ar}(y_o)$ among offspring, we need the variance decomposition formula

$$\mathbb{V}\text{ar}(y_o) = \mathbb{V}\text{ar}(\mathbb{E}(y_o|y_p)) + \mathbb{E}(\mathbb{V}\text{ar}(y_o|y_p)).$$

Table S.3 Intervals, recombination fractions θ , and map distance x for Solution 5.3.3.

Interval	θ	x
A-B	0.1	0.1012
B-C	0.05	0.0502
C-D	0.2	0.2118
A-C	0.147	0.1514
B-D	0.24	0.2620
A-B, C-D	0.278	0.3130
A-D	0.31	0.3632

With the assumptions $\mu_O = \mu_P = \mu$ and $\text{Var}(y_O) = \text{Var}(y_P) = \text{Var}(y)$, the coefficient of determination R^2 is given by

$$\begin{aligned} R^2 &= 1 - \frac{\mathbb{E}(\text{Var}(y_O|y_P))}{\text{Var}(y_O)} = \frac{\text{Var}(y_O) - \mathbb{E}(\text{Var}(y_O|y_P))}{\text{Var}(y_O)} \\ &= \frac{\text{Var}(\mathbb{E}(y_O|y_P)) + \mathbb{E}(\text{Var}(y_O|y_P)) - \mathbb{E}(\text{Var}(y_O|y_P))}{\text{Var}(y_O)} \\ &= \frac{\text{Var}(\mathbb{E}(y_O|y_P))}{\text{Var}(y_O)} = \frac{\beta^2 \text{Var}(y_P)}{\text{Var}(y_O)} = \beta^2. \end{aligned}$$

The coefficient of determination R^2 thus equals the squared of the regression coefficient β . Thus, β can be used for estimating the familial resemblance between a parent and an offspring. Specifically, 2β is the so-called *empirical heritability*; see, e.g., [337, 353]. Note that this empirical heritability is not a measure of genetic effects alone but also includes environmental factors.

If both parental phenotypes are available, the empirical heritability can be deduced as described above if a) the midparent value $(y_M + y_F)/2$ is used and b) parental phenotypes are assumed to be independent. Therefore, no inbreeding may be present, and the family may not be selected through a proband not being the parents. If these assumptions are fulfilled, the regression coefficient β estimated from the linear model $y_O = \mu_O + \beta \frac{y_F + y_M}{2} + \varepsilon$ is the empirical heritability coefficient.

Solution 7.1 Both grandparents are homozygous at all three marker loci. The mother therefore has haplotypes $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \bigg| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}$ regardless of the marker order. The father is homozygous at all three markers and thus non-informative for linkage. There are six informative meioses in the offspring. Four recombinations are observed between markers M_1 and M_2 , while only two recombinations are observed between markers M_1 and M_3 and between markers M_2 and M_3 , respectively. Therefore, the most plausible order of the markers is M_2 – M_3 – M_1 . Because no information about p – q orientation of the chromosome is available, M_1 – M_3 – M_2 is an identical marker order.

Solution 7.2 Both parents are heterozygous at both marker loci, and thus informative for linkage analysis.

Solution 7.2.1. The grandparental meiosis are used to determine the phases in both parents. There are 24 meioses from the parents to their children. All of these are informative.

Solution 7.2.2. The maternal haplotypes are $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}$ and $\begin{smallmatrix} 2 \\ 2 \end{smallmatrix}$. The paternal haplotypes have been determined in Section 7.1.1. Two recombinations have occurred in the father, and none in the mother. The kernel of the likelihood thus is $L(\theta) = \theta^2(1-\theta)^{22}$.

Solution 7.2.3. θ is estimated for the pedigree as $\hat{\theta} = \frac{2}{24} \approx 0.0833$.

Solution 7.2.4. The maximum LOD score for this pedigree is

$$LOD \approx \lg \frac{0.0833^2(1 - 0.0833)^{22}}{0.5^{24}} \approx 4.2350.$$

Subsequently, $LRT \approx 2 \ln(10) LOD \approx 19.5029$ and $p \approx 5.0264 \cdot 10^{-6}$.

Solution 7.3 The calculations for this pedigree are more complicated than those of the previous problem.

Solution 7.3.1. At the beginning, we note that the paternal haplotypes are $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \big| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}$, as can be determined from the grandparental genotypes. The mother has the two possible haplotype combinations $H_1 = \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \big| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}$ and $H_2 = \begin{smallmatrix} 1 \\ 2 \end{smallmatrix} \big| \begin{smallmatrix} 2 \\ 1 \end{smallmatrix}$, each with 50% probability. Four different genotype combinations are observed in the offspring: $6 \times O_1 = \begin{smallmatrix} 11 \\ 11 \end{smallmatrix}$, $4 \times O_2 = \begin{smallmatrix} 12 \\ 12 \end{smallmatrix}$, $1 \times O_3 = \begin{smallmatrix} 11 \\ 12 \end{smallmatrix}$, and $1 \times O_4 = \begin{smallmatrix} 12 \\ 11 \end{smallmatrix}$, where O is the abbreviation for offspring. If the phase of the mother is H_1 , the kernel of the likelihood is

$$L_1(\theta) = \underbrace{(1-\theta)^{12}}_{O_1} \cdot \underbrace{(0.5\theta^2 + 0.5(1-\theta)^2)^4}_{O_2} \cdot \underbrace{\theta(1-\theta)}_{O_3} \cdot \underbrace{\theta(1-\theta)}_{O_4}.$$

This result is obtained by noting that in none of the maternal and paternal meioses recombinations occur for O_1 . The haplotype combinations in the 4 O_2 offspring are either $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \big| \begin{smallmatrix} 2 \\ 2 \end{smallmatrix}$ or $\begin{smallmatrix} 1 \\ 2 \end{smallmatrix} \big| \begin{smallmatrix} 2 \\ 1 \end{smallmatrix}$, both with equal probabilities. While the first combination corresponds to eight non-recombinants, the second combination corresponds to eight recombinants. O_3 and O_4 must exhibit exactly one recombination and one non-recombination.

$L_2(\theta)$ can be determined analogously for H_2 :

$$L_2(\theta) = \underbrace{\theta^6(1-\theta)^6}_{O_1} \cdot \underbrace{\theta^4(1-\theta)^4}_{O_2} \cdot \underbrace{(0.5\theta^2 + 0.5(1-\theta)^2)}_{O_3} \cdot \underbrace{(0.5\theta^2 + 0.5(1-\theta)^2)}_{O_4}.$$

H_1 and H_2 are equally likely, and thus the kernel of the joint likelihood function is given by $L(\theta) = 0.5 L_1(\theta) + 0.5 L_2(\theta)$. The calculations in this fairly simple pedigree demonstrate the clear need for efficient computer programs that can be used for estimating the likelihood in pedigrees.

Solution 7.3.2. The *LOD* score table is given in Table S.4.

Table S.4 *LOD* score table for Problem 7.3.

θ	0.0	0.01	0.05	0.1	0.2	0.3	0.4
<i>LOD</i> score	$-\infty$	1.62	2.63	2.73	2.29	1.56	0.71

Solution 7.4 The phase in the father depends on the haplotypes of the grandmother. We therefore first determine the haplotype combinations in the grandmother. Second, we derive the kernels of the likelihood for each possible haplotype combination. Third, we calculate the probability of each haplotype combination, and, finally, we combine all three parts to the joint likelihood.

Solution 7.4.1. Because the father is heterozygous for the 1 allele at both marker loci, the grandmother must have at least one 1 allele at both marker loci. Therefore,

only the following 5 different haplotype combinations out of 10 possible haplotype combinations are admissible:

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad H_3 = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad H_5 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Solution 7.4.2. Haplotype combination H_5 implies an extra recombination from the grandmother to the father, while H_4 exhibits no recombination. The other haplotype combinations H_1 to H_3 are not informative for linkage because they show homozygosity for at least one marker. Thus, there are $n = 12$ informative meioses for combinations H_1 to H_3 . Pedigrees resulting from H_4 and H_5 have $n = 13$ informative meioses. One can now easily count the number of recombinants and non-recombinants: $k = 2$ for H_1 to H_4 , and $k = 3$ for H_5 . The kernels of the likelihoods corresponding to H_1 and H_5 thus are $L_1(\theta) = L_2(\theta) = L_3(\theta) = \theta^2(1 - \theta)^{10}$, $L_4(\theta) = \theta^2(1 - \theta)^{11}$, and $L_5(\theta) = \theta^3(1 - \theta)^{10}$.

Solution 7.4.3. To derive the kernel of the joint likelihood, the probabilities of the five possible haplotype combinations need to be calculated. By assuming linkage equilibrium between marker loci and Hardy-Weinberg equilibrium at both marker loci, one obtains identical conditional probabilities for the possible haplotype combinations as in Eq. (7.6).

Solution 7.4.4. The kernel of the joint likelihood function is given by

$$L(\theta) = \sum_{i=1}^5 P(H_i) \cdot L_i(\theta),$$

with $P(H_i)$ and $L_i(\theta)$ given in Eq. (7.6) and Solution 7.4.2, respectively. With $p_1 = p_2 = 0.5$, giving $P(H_1) \approx 0.1111$, $P(H_2) = P(H_3) = P(H_4) = P(H_5) \approx 0.2222$, the kernel of the joint likelihood is approximately

$$\begin{aligned} L(\theta) &= 0.5555 \theta^2 (1 - \theta)^{10} + 0.2222 \theta^2 (1 - \theta)^{11} + 0.2222 \theta^3 (1 - \theta)^{10} \\ &= \theta^2 (1 - \theta)^{10} (0.5555 + 0.2222 \theta + 0.2222 (1 - \theta)) \\ &\propto \theta^2 (1 - \theta)^{10}. \end{aligned}$$

This result seems to be surprising: The kernel of the joint likelihood for the pedigree displayed in Figure 7.4 without grandmaternal data is identical to the joint likelihood with grandmaternal data. However, the grandfather is homozygous at both marker loci; therefore, the phase in the father can be determined even without the genotypes of the grandmother.

Solution 7.5 The pedigree depicted in Figure 7.18 represents the simplest case for a pedigree following an autosomal recessive mode of inheritance with complete penetrance, no phenocopies, and no mutation.

Solution 7.5.1. The mother is not informative for linkage. Her haplotypes are $\frac{1}{D} \mid \frac{1}{D}$. Instead, the father is informative for linkage with haplotypes $\frac{1}{D} \mid \frac{2}{N}$.

Solution 7.5.2. One recombination has occurred in the last offspring, i.e., the unaffected girl is homozygous 1 at the marker locus.

Solution 7.5.3. The kernel of the likelihood function is $\theta(1 - \theta)^7$, and the LOD score function subsequently is $\text{LOD}(\theta) = \lg(\theta(1 - \theta)^7)/(0.5^8)$. The ML estimate for the recombination fraction is $\hat{\theta} = \frac{1}{8}$, giving a maximum *LOD* score of 1.0992.

Solution 7.6 As outlined in the text, information can be gained by considering the markers jointly.

Solution 7.6.1. Only the left part of the pedigree is informative for linkage at marker 1. Specifically, only the affected son is informative for linkage. The affected mother from the right part of the pedigree is homozygous 4, and thus not informative.

Solution 7.6.2. Only the right part of the pedigree is informative for linkage at marker 2. Specifically, the affected mother on the right part is informative for linkage. Her affected brother is homozygous at the marker 2, thus not informative.

Table S.5 Two-point *LOD* score table for Problem 7.6.3.

Recombination fraction θ	0.0	0.1	0.2	0.3	0.4
<i>LOD</i> score for marker M_1 at θ	0.6	0.51	0.41	0.29	0.16
<i>LOD</i> score for marker M_2 at θ	0.9	0.77	0.61	0.44	0.24

Solution 7.6.3. Tables S.5 and S.6 show the two-point and multipoint *LOD* scores, respectively. In the multipoint situation the trait locus can be either on one side of the two markers or between the two genetic markers. Figure S.1 also shows the haplotypes of the three-generation pedigree as determined with GENEHUNTER.

Table S.6 Multipoint *LOD* score table for Problem 7.6.3.

$T \quad M_1$	0.5	0.4	0.3	0.2	0.1	0
$M_1 \quad M_2$	0.005	0.005	0.005	0.005	0.005	0.005
<i>LOD</i> score	0.0000	0.3937	0.7269	1.0157	1.2706	1.4986
$M_1 \quad T^a$	0.000	0.001	0.002	0.003	0.004	0.005
$T \quad M_2$	0.005	0.004	0.003	0.002	0.001	0.000
<i>LOD</i> score	1.4986	1.4990	1.4995	1.4999	1.5003	1.5008
$M_1 \quad M_2$	0.005	0.005	0.005	0.005	0.005	0.005
$M_2 \quad T$	0.0	0.1	0.2	0.3	0.4	0.5
<i>LOD</i> score	1.5008	1.2725	1.0173	0.7281	0.3944	0.0000

In each block, the recombination fraction between the two pairs of loci is given in the first two rows. The third row displays the multipoint *LOD* score. The assumed locus order is $T-M_1-M_2$ in the first block of the table, with T denoting the trait locus. It is M_1-T-M_2 and M_1-M_2-T in the second and third block, respectively.

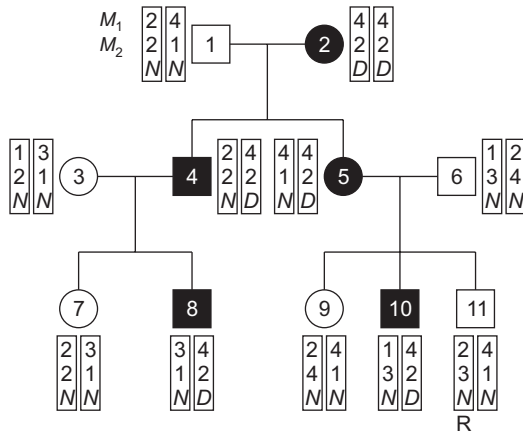


Fig. S.1 Haplotypes for the three-generation pedigree of Problem 7.6. The displayed solution is the printout from GENEHUNTER. At least one other solution is possible. Offspring 9 and 10 could show a recombination, while offspring 11 does not.

Solution 7.7

Solution 7.7.1. (a) All meioses are phase-known. We can identify III_1-III_6 and III_8 unambiguously as being non-recombinant and III_7 as being recombinant. (b) The same family as (a), now phase-unknown. The mother, II_2 , could have inherited either marker allele 1 or marker allele 3 with the disease. She is therefore phase-unknown. Subsequently, either individuals III_1-III_6 and III_8 are nonrecombinant and III_7 is recombinant or III_1-III_6 and III_8 are recombinant and III_7 is non-recombinant.

Solution 7.7.2. (c) The same family as in (b) after further tracing of relatives. The new part of the pedigree (c) makes analyses more complicated by introducing a dependency on marker allele frequencies. Nevertheless, pedigree (c) is more informative than (b) because it is likely that the disease runs with allele 1 in the family.

Solution 7.7.3. In pedigree (c), III_9 and III_{10} have also inherited marker allele 1 along with the disease from their father. However, we cannot be sure whether their father's allele 1 is IBD (see Chapter 8) to allele 1 in his sister II_2 . There could well be two copies of allele 1 among the four grandparental marker alleles. The likelihood of this depends on the allele frequency.

Solution 7.8 The expected value of the LRT, denoted by Λ equals 1 under H_0 . Then, on the null hypothesis, the inequality $\Lambda > a$ for $a > 1$ cannot occur with probability greater than $1/a$, since if it did, this in itself would be enough to raise $\mathbb{E}(\Lambda)$ to 1. Therefore, the occurrence of a value of Λ greater than a is at least as strong evidence against H_0 as a test level of $1/a$.

Solution 8.1 In the first pedigree, both IBS and IBD values are 1. In the second pedigree, they are 2.

Solution 8.2 Because the disease is a simple autosomal recessive with complete penetrance, no phenocopies and no mutations, both parents are heterozygous DN at the trait locus, and the offspring are homozygous DD . We aim at calculating the probability that the ASP shares j alleles IBD at a marker locus. We start off with z_2 ,

the probability that $IBD = 2$. There are four different genotype combinations in the affected offspring giving $IBD_m = 2$ at the marker locus, thus

$$P(IBD_m = 2|FD) = P(1\ 3, 1\ 3|FD) + P(1\ 4, 1\ 4|FD) \\ + P(2\ 3, 2\ 3|FD) + P(2\ 4, 2\ 4|FD),$$

where $P(1\ 3, 2\ 4)$ denotes P (“marker genotype of one offspring is 1 3”, “marker genotype of the other offspring is 2 4”), and “FD” stands for the additional information on the family data that is available. Here, FD means that both offspring are affected. They are thus homozygous D at the trait locus, yielding

$$P(IBD_m = 2|FD) = P(1D\ 3D, 1D\ 3D|FD) + P(1D\ 4D, 1D\ 4D|FD) \\ + P(2D\ 3D, 2D\ 3D|FD) + P(2D\ 4D, 2D\ 4D|FD). \quad (S.1)$$

Four parental phases are possible, as displayed in Figure S.2. If parental phase is as given in family a), an offspring genotype $1D\ 3D$ corresponds to two recombinations. Therefore, the contribution of one affected offspring to the probability of interest is θ^2 . Since the family consists of two affected offspring with identical genotypes, their joint contribution is $\theta^{2^2} = \theta^4$.

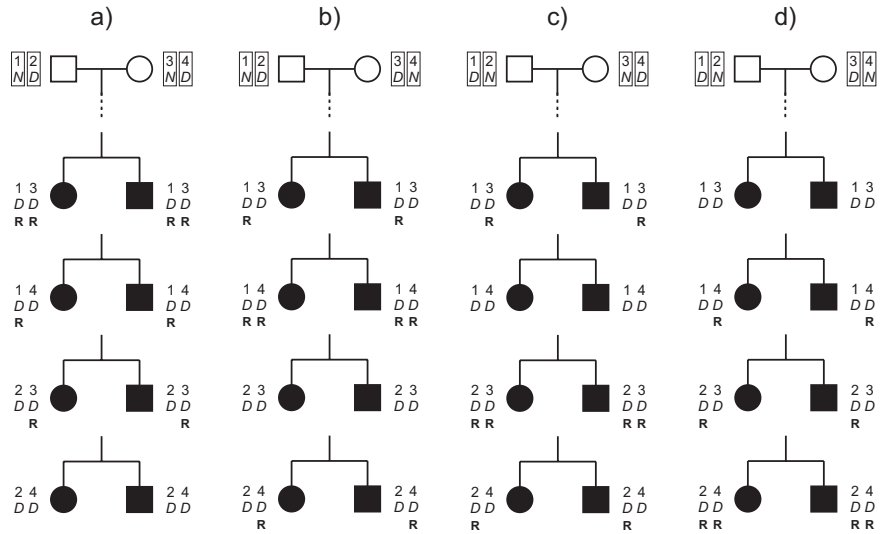


Fig. S.2 The four possible parental phases for the affected sib-pair family of Figure 8.6. Furthermore, all possible offspring genotypes are shown given the parental genotypes and an identical by descent (IBD) value of 2. The letter R indicates a recombination for the respective meiosis. It can be seen that all four possible parental phases, there are four possible genotype combinations. One, two, and one of these per genotype combination show none, two, and four recombinations, respectively.

Similarly, $2D\ 4D$ corresponds to two non-recombinants, thus $(1 - \theta)^2$ for one affected offspring. Finally, $1D\ 4D$ and $2D\ 3D$ both correspond to one recombination

and one non-recombination, yielding $\theta(1 - \theta)$ for one affected offspring. In all cases, the family has two affected offspring with identical genotypes, and Eq. (S.1) may be rewritten as

$$\begin{aligned} P(\text{IBD}_m = 2|\text{FD, const. a}) &= \theta^4 + \theta^2(1 - \theta)^2 + \theta^2(1 - \theta)^2 + (1 - \theta)^4 \\ &= \theta^4 + 2\theta^2(1 - \theta)^2 + (1 - \theta)^4 \end{aligned} \quad (\text{S.2})$$

for the family depicted in Figure S.2.a. Here, “const. a)” denotes “family constellation a)”. For the other constellations b)–d) of Figure S.2, it can be seen that $P(\text{IBD}_m = 2|\text{FD, const. x})$ is always identical to the expression given in Eq. (S.2). In summary, there are four parental phases that are equally likely. All four yield identical probabilities so that, finally,

$$P(\text{IBD}_m = 2|\text{FD}) = \theta^4 + 2\theta^2(1 - \theta)^2 + (1 - \theta)^4. \quad (\text{S.3})$$

For $\text{IBD}_m = 1$, one analogously obtains

$$\begin{aligned} P(\text{IBD}_m = 1|\text{FD}) &= P(1D\ 3D, 2D\ 3D|\text{FD}) + P(1D\ 3D, 1D\ 4D|\text{FD}) \\ &\quad + P(2D\ 3D, 2D\ 4D|\text{FD}) + P(1D\ 4D, 2D\ 4D|\text{FD}) \\ &= 2\theta^3(1 - \theta) + 2\theta^3(1 - \theta) + 2\theta(1 - \theta)^3 + 2\theta(1 - \theta)^3 \\ &= 4\theta(1 - \theta)[\theta^2 + (1 - \theta)^2] \\ &= 4\theta(1 - \theta)\Psi. \end{aligned} \quad (\text{S.4})$$

The term $\Psi = \theta^2 + (1 - \theta)^2$ points to the fact that either none or two recombinations must occur per parent in the offspring. Finally, $P(\text{IBD}_m = 0|\text{FD}) = 1 - P(\text{IBD}_m = 2|\text{FD}) - P(\text{IBD}_m = 1|\text{FD}) = 4\theta^2(1 - \theta)^2$.

Solution 8.3 One parent and both offspring are affected in the nuclear family. The other parent is supposed to be unaffected, thus not informative for linkage. As the disease is rare, the affected parent can be assumed to be heterozygous DN . Two phases are possible for the affected mother: $\begin{smallmatrix} 3 \\ D \end{smallmatrix} | \begin{smallmatrix} 4 \\ N \end{smallmatrix}$ and $\begin{smallmatrix} 3 \\ N \end{smallmatrix} | \begin{smallmatrix} 4 \\ D \end{smallmatrix}$, each with probability $\frac{1}{2}$.

First, we compute $P(\text{IBD}_m = 2|\text{FD})$. Both offspring need to have the same marker genotypes, and four equally likely offspring constellations are compatible with this. If the maternal phase is $\begin{smallmatrix} 3 \\ D \end{smallmatrix} | \begin{smallmatrix} 4 \\ N \end{smallmatrix}$, the offspring genotypes 1 3 and 2 3 exhibit no recombination. The constellation of the maternal phase being $\begin{smallmatrix} 3 \\ D \end{smallmatrix} | \begin{smallmatrix} 4 \\ N \end{smallmatrix}$ and offspring genotypes 1 3 therefore occurs with probability

$$\underbrace{\frac{1}{2}}_{\text{phase}} \cdot \underbrace{\frac{1}{4}}_{\text{constellation}} \cdot \underbrace{(1 - \theta)^2}_{\text{no recombination}}.$$

The same probability is obtained for the genotype 2 3. If the offspring genotypes are 1 4 or 2 4, a recombination must have occurred in both offspring. This happens with probability $\frac{1}{2} \cdot \frac{1}{4} \cdot \theta^2$ if the maternal phase is $\begin{smallmatrix} 3 \\ D \end{smallmatrix} | \begin{smallmatrix} 4 \\ N \end{smallmatrix}$. In summary,

$$P(\text{IBD}_m = 2|\text{FD, maternal phase} = \begin{smallmatrix} 3 \\ D \end{smallmatrix} | \begin{smallmatrix} 4 \\ N \end{smallmatrix}) = \frac{1}{4} \cdot \frac{1}{2} \cdot \theta^2 \cdot 2 + \frac{1}{4} \cdot \frac{1}{2} \cdot (1 - \theta)^2 \cdot 2.$$

The same probability is obtained if maternal phase is $\frac{3}{N} \mid \frac{4}{D}$, yielding

$$P(\text{IBD}_m = 2|\text{FD}) = 2\left(\frac{1}{4} \cdot \theta^2 + \frac{1}{4} \cdot (1 - \theta)^2\right) = \frac{1}{2}(\theta^2 + (1 - \theta)^2) = \frac{1}{2}\Psi$$

in total with $\Psi = \theta^2 + (1 - \theta)^2$. Analogously, one obtains $P(\text{IBD}_m = 0|\text{FD}) = \theta(1 - \theta)$. This last probability can also be explained intuitively. If the IBD value needs to be 0, exactly one recombinant and one non-recombinant need to be observed for the informative maternal meiosis. The paternal meiosis does not contribute. As the recombination may take place in either the first or the second offspring, the resulting expression needs to be twice as large as the probability for sharing two alleles IBD. Finally, $P(\text{IBD}_m = 1|\text{FD}) = 1 - P(\text{IBD}_m = 2|\text{FD}) - P(\text{IBD}_m = 0|\text{FD}) = \frac{1}{2}$.

Solution 8.4 The data are from two-point analysis, resulting in seven possible IBD constellations. The update formula can therefore be simplified.

Solution 8.4.1. The update formulae for this example are

$$\begin{aligned}\hat{z}_0^{(k+1)} &= \frac{1}{81} \left(4 + 7 \frac{\frac{1}{3}\hat{z}_0^{(k)}}{\frac{1}{3}\hat{z}_0^{(k)} + \frac{2}{3}\hat{z}_1^{(k)}} + 5 \frac{\frac{1}{2}\hat{z}_0^{(k)}}{\frac{1}{2}\hat{z}_0^{(k)} + \frac{1}{2}\hat{z}_2^{(k)}} + 6 \hat{z}_0^{(k)} \right), \\ \hat{z}_1^{(k+1)} &= \frac{1}{81} \left(13 + 25 \frac{\frac{2}{3}\hat{z}_1^{(k)}}{\frac{2}{3}\hat{z}_1^{(k)} + \frac{1}{3}\hat{z}_2^{(k)}} + 7 \frac{\frac{2}{3}\hat{z}_1^{(k)}}{\frac{1}{3}\hat{z}_0^{(k)} + \frac{2}{3}\hat{z}_1^{(k)}} + 6 \hat{z}_1^{(k)} \right), \\ \hat{z}_2^{(k+1)} &= \frac{1}{81} \left(21 + 25 \frac{\frac{1}{3}\hat{z}_2^{(k)}}{\frac{2}{3}\hat{z}_1^{(k)} + \frac{1}{3}\hat{z}_2^{(k)}} + 5 \frac{\frac{1}{2}\hat{z}_2^{(k)}}{\frac{1}{2}\hat{z}_0^{(k)} + \frac{1}{2}\hat{z}_2^{(k)}} + 6 \hat{z}_2^{(k)} \right).\end{aligned}$$

Solution 8.4.2. The results are given in Table S.7.

Table S.7 Updates from the expectation maximization algorithm for estimating the identical by descent (IBD) probabilities in the sample of the 81 affected sib-pairs given in Table 8.2.

Update step	IBD probability estimates			Update step	IBD probability estimates		
	$\hat{z}_0$	$\hat{z}_1$	$\hat{z}_2$		$\hat{z}_0$	$\hat{z}_1$	$\hat{z}_2$
0	0.10530	0.3421	0.5526	3	0.0689	0.4824	0.4485
1	0.07858	0.4314	0.4899	4	0.0684	0.4875	0.4439
2	0.07095	0.4685	0.4605	5	0.0683	0.4894	0.4421

Solution 8.5 First, z_0 trivially is ≥ 0 . Second, $z_1 \leq \frac{1}{2}$ because $\frac{1}{2}\lambda_S \geq 0$, $(2\Psi - 1)^2 \geq 0$ and $\lambda_S - \lambda_O \geq 0$. To finally show $2 \cdot z_0 \leq z_1$ we first note that this inequality is equivalent to $\frac{3}{2}z_1 + z_2 \geq 1$ because of $\sum_{j=0}^2 z_j = 1$. Straightforward calculations give

$$\begin{aligned}\frac{3}{2}z_1 + z_2 &= 1 - \frac{2\Psi-1}{4\lambda_S} 3(2\Psi-1)(\lambda_S - \lambda_O) + \frac{2\Psi-1}{4\lambda_S} ((\lambda_S - 1) + 2\Psi(\lambda_S - \lambda_O)) \\ &= 1 + \frac{2\Psi-1}{4\lambda_S} ((\lambda_S - 1) + (3 - 4\Psi)(\lambda_S - \lambda_O)).\end{aligned}\quad (\text{S.5})$$

$(3 - 4\Psi)(\lambda_S - \lambda_O)$ cannot be smaller than $-(\lambda_S - \lambda_O) = -\frac{1}{4K^2}\sigma_d^2$ because $\Psi \leq 1$. $\lambda_S - 1 = \frac{1}{4K^2}\sigma_d^2 + \frac{1}{2K^2}\sigma_a^2$. Thus, $(\lambda_S - 1) + (3 - 4\Psi)(\lambda_S - \lambda_O) \geq 0$. Because

$2\Psi \geq 1$, we have shown that the second term of Eq. (S.5) ≥ 0 , which completes the proof.

Solution 8.6 The second one answers, “It’s okay, but you lose a degree of freedom!”

Solution 8.7 This problem does not comprise a sufficient number of notional families for the asymptotic results for the test statistics of interest to be applied and is included only for training on the computational skills.

Solution 8.7.1. The estimate of the IBD distribution for the total sample prior to application of the EM algorithm gives $\hat{z}_0 = 0.0007$, $\hat{z}_1 = 0.2894$, and $\hat{z}_2 = 0.7099$. We can easily verify that this is a point in the possible triangle. First, $\hat{z}_1 = 0.2894$, thus < 0.5 . Second, $2\hat{z}_0 < \hat{z}_1$.

Table S.8 Weights at the marker locus for the 10 affected sib-pair families of Table 8.8 from Problem 8.7.

Family	$\hat{w}_{i0}$	$\hat{w}_{i1}$	$\hat{w}_{i2}$	Family	$\hat{w}_{i0}$	$\hat{w}_{i1}$	$\hat{w}_{i2}$
1	0	0	1	6	0	$\frac{2}{3}$	$\frac{1}{3}$
2	0	0	1	7	$\frac{1}{3}$	$\frac{1}{3}$	0
3	0	0	1	8	$\frac{1}{3}$	$\frac{1}{3}$	0
4	0	0	1	9	$\frac{1}{2}$	0	$\frac{1}{2}$
5	0	$\frac{2}{3}$	$\frac{1}{3}$	10	$\frac{1}{2}$	0	$\frac{1}{2}$

Solution 8.7.2. Table S.9 shows the required values for calculating both the score-type and the Wald-type versions of the mean test. From the table it can be seen that $\bar{\hat{\tau}} = 0.7$ in the sample. Furthermore, $\sum_{i=1}^{10} (\hat{\tau}_i - \bar{\hat{\tau}})^2 = 0.85$. Subsequently, $T_{W-M} = (\bar{\hat{\tau}} - 0.5) / \sqrt{\text{Var}(\bar{\hat{\tau}})} = (0.7 - 0.5) / \sqrt{\frac{1}{10.9} \sum_{i=1}^{10} (\hat{\tau}_i - \bar{\hat{\tau}})^2} \approx 0.2 / 0.0972 = 2.0580$. This gives an *LOD* score of approximately 0.92 and an asymptotic one-sided *p*-value from the *t*-distribution with 8 d.f. of approximately 0.0370. Similarly, we obtain $T_{S-M} = (\bar{\hat{\tau}} - 0.5) / \sqrt{1/(8n)} = 0.2 / \sqrt{1/80} \approx 1.7888$, corresponding to an *LOD* of approximately 0.6949 and an asymptotic one-sided *p*-value from the *t*-distribution with 8 d.f. of approximately 0.0557.

Table S.9 Working table for the affected sib-pair families of Problem 8.7.

Family	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{\tau}_i$	$(\hat{\tau}_i - \bar{\hat{\tau}})^2$	Family	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{\tau}_i$	$(\hat{\tau}_i - \bar{\hat{\tau}})^2$
1	0	1	1	0.09	6	0.5	0.5	0.75	0.0025
2	0	1	1	0.09	7	0.5	0.5	0.25	0.2025
3	0	1	1	0.09	8	0.5	0	0.25	0.2025
4	0	1	1	0.09	9	0	0.5	0.5	0.04
5	0.5	0.5	0.75	0.0025	10	0	0.5	0.5	0.04

Solution 8.8 The power is the probability of rejecting the null hypothesis H_0 if the alternative hypothesis H_1 is true. The mean test rejects H_0 if $T_{S-M} = \sqrt{8n}(\hat{\tau} - \frac{1}{2}) > z_{1-\alpha}$. The power of the mean test is thus given by $P_{H_1}(T_{S-M} = \sqrt{8n}(\hat{\tau} - \frac{1}{2}) > z_{1-\alpha})$, where, as before, P_{H_1} denotes the probability being evaluated under H_1 .

With $\mathbb{E}_{H_1}(\hat{\tau}) = \tau_m = \frac{1}{2} + \delta_m$ and $\mathbb{V}\text{ar}_{H_1}(\hat{\tau}) = \frac{1}{2n}\tau_m(1 - \tau_m) = \frac{1}{2n}(\frac{1}{4} - \delta_m^2)$, one subsequently obtains

$$\begin{aligned} \text{Power} &= P_{H_1}\left(\hat{\tau} > \frac{1}{2} + \frac{z_{1-\alpha}}{\sqrt{8n}}\right) = P_{H_1}\left(\underbrace{\frac{\hat{\tau} - \mathbb{E}_{H_1}(\hat{\tau})}{\sqrt{\mathbb{V}\text{ar}_{H_1}(\hat{\tau})}}}_{\stackrel{\sim}{\sim} N(0,1) \text{ under } H_1} > \frac{\frac{1}{2} + \frac{z_{1-\alpha}}{\sqrt{8n}} - \mathbb{E}_{H_1}(\hat{\tau})}{\sqrt{\mathbb{V}\text{ar}_{H_1}(\hat{\tau})}}\right) \\ &\approx 1 - \Phi\left(\frac{\frac{z_{1-\alpha}}{\sqrt{8n}} - (\mathbb{E}_{H_1}(\hat{\tau}) - \frac{1}{2})}{\sqrt{\mathbb{V}\text{ar}_{H_1}(\hat{\tau})}}\right) = 1 - \Phi\left(\frac{\frac{z_{1-\alpha}}{\sqrt{8n}} - \delta_m}{\sqrt{(\frac{1}{4} - \delta_m^2)/\sqrt{2n}}}\right) \\ &= 1 - \Phi\left(\frac{\frac{1}{2}z_{1-\alpha} - \delta_m\sqrt{2n}}{\sqrt{\frac{1}{4} - \delta_m^2}}\right). \end{aligned}$$

The test has power $1 - \beta$ if $\beta = \Phi\left(\left(\frac{1}{2}z_{1-\alpha} - \delta_m\sqrt{2n}\right)/\left(\sqrt{\frac{1}{4} - \delta_m^2}\right)\right)$, which is identical to

$$z_\beta = \frac{\frac{1}{2}z_{1-\alpha} - \delta_m\sqrt{2n}}{\sqrt{\frac{1}{4} - \delta_m^2}}.$$

If this equation is solved for n , we obtain that at least $n \geq \left(\frac{1}{2}z_{1-\alpha} + z_{1-\beta}\sqrt{\left(\frac{1}{4} - \delta_m^2\right)}\right)^2/(2\delta_m^2)$ ASP are required in the mean test to have a power of $1 - \beta$ with a significance level of α at a fully informative marker locus. Furthermore, because δ_m is generally close to 0, we obtain $\sqrt{\frac{1}{4} - \delta_m^2} \approx \frac{1}{2}$. Therefore, we can approximate the sample size formula by

$$n \geq \frac{\left(\frac{1}{2}z_{1-\alpha} + \frac{1}{2}z_{1-\beta}\right)^2}{2\delta_m^2} \approx \frac{\frac{1}{4}(z_{1-\alpha} + z_{1-\beta})^2}{2\delta_m^2} \approx \frac{\left(\frac{1}{4} - \delta_m^2\right)(z_{1-\alpha} + z_{1-\beta})^2}{2\delta_m^2}.$$

It is easily seen that $\delta_m = \tau_m - \frac{1}{2}$ and $\frac{1}{4} - \delta_m^2 = \tau_m(1 - \tau_m)$. This completes the proof.

Solution 9.1

$$\begin{aligned} \mathbb{E}(y_i|z_{i1}) &= \sum_{g_{i1}, g_{i2}} P(g_{i1}, g_{i2}|z_{i1}) \cdot \mathbb{E}(y_i|g_{i1}, g_{i2}) \\ &= p^3\mathbb{E}(\varepsilon_i^2) + pq\mathbb{E}(\varepsilon_i^2) + q^3\mathbb{E}(\varepsilon_i^2) \\ &\quad + p^2q\mathbb{E}((a - d + \varepsilon_i)^2) + p^2q\mathbb{E}((d - a + \varepsilon_i)^2) \\ &\quad + pq^2\mathbb{E}((a + d + \varepsilon_i)^2) + pq^2\mathbb{E}((-a - d + \varepsilon_i)^2). \end{aligned}$$

With $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \mathbb{V}\text{ar}(\varepsilon_i) = \sigma_\varepsilon^2$, one obtains

$$\begin{aligned}\mathbb{E}((\pm d \pm a + \varepsilon_i)^2) &= \mathbb{E}((\pm d \pm a)^2 + 2(\pm d \pm a)\varepsilon_i + \varepsilon_i^2) \\ &= (\pm d \pm a)^2 + 2(\pm d \pm a)\mathbb{E}(\varepsilon_i) + \mathbb{E}(\varepsilon_i^2) \\ &= (\pm d \pm a)^2 + \sigma_\varepsilon^2,\end{aligned}$$

so that

$$\begin{aligned}\mathbb{E}(y_i|z_{i1}) &= \underbrace{(p^3 + pq + q^3 + 2p^2q + 2pq^2)}_{=1 \text{ since } pq=p^2q+q^2p} \sigma_\varepsilon^2 \\ &\quad + p^2q(a-d)^2 + p^2q(d-a)^2 + pq^2(a+d)^2 + pq^2(-a-d)^2 \\ &= \sigma_\varepsilon^2 + 2p^2q(a^2 - 2ad + d^2) + 2pq^2(a^2 + 2ad + d^2) \\ &= \sigma_\varepsilon^2 + 2pq(a-d(p-q))^2 + 2pqd^2(1-(p-q)^2) \\ &= \sigma_\varepsilon^2 + 2pqd^2(4pq) = \sigma_\varepsilon^2 + \sigma_a^2 + 2\sigma_d^2\end{aligned}$$

because $(1-(p-q)^2) = (p^2 + 2pq + q^2) - p^2 + 2pq - q^2 = 4pq$.

$\mathbb{E}(y_i|z_{i0})$ can be determined analogously:

$$\begin{aligned}\mathbb{E}(y_i|z_{i0}) &= \sum_{g_{i1}, g_{i2}} P(g_{i1}, g_{i2}|z_{i1}) \cdot \mathbb{E}(y_i|g_{i1}, g_{i2}) \\ &= p^4\mathbb{E}(\varepsilon_i^2) + 4p^2q^2\mathbb{E}(\varepsilon_i^2) + q^4\mathbb{E}(\varepsilon_i^2) \\ &\quad + 2p^3q\mathbb{E}((a-d+\varepsilon_i)^2) + 2p^3q\mathbb{E}((-a+d+\varepsilon_i)^2) \\ &\quad + 2pq^3\mathbb{E}((a+d+\varepsilon_i)^2) + 2pq^3\mathbb{E}((-a-d+\varepsilon_i)^2) \\ &\quad + p^2q^2\mathbb{E}((2a+\varepsilon_i)^2) + p^2q^2\mathbb{E}((-2a+\varepsilon_i)^2) \\ &= \sigma_\varepsilon^2 + 4p^3q(a-d)^2 + 4pq^3(a+d)^2 + 8p^2q^2a^2 = \dots \\ &= \sigma_\varepsilon^2 + 2\sigma_a^2 + 4pq(2pqd^2) = \sigma_\varepsilon^2 + 2\sigma_a^2 + 2\sigma_d^2.\end{aligned}$$

Solution 9.2 Table S.10 contains the relevant data extracted from the 10 nuclear families and Table S.11 contains additional data for estimating the regression coefficients. Finally, Figure S.3 displays all squared trait differences of the sib-pairs

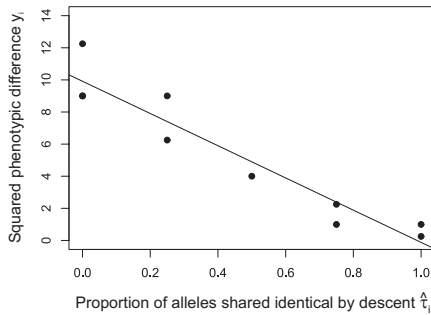


Fig. S.3 Scatterplot and Haseman-Elston regression line for the 10 nuclear families from Problem 9.2. Plotted is the squared trait difference of a notional quantitative trait against the proportion of alleles shared identical by descent of a sib-pair at a notional genetic marker locus.

Table S.10 Quantitative traits for the individual siblings, the squared trait difference of the sib pairs, the estimated identical by descent (IBD) distributions at the marker locus, and the estimated proportion of alleles shared IBD at the marker locus for the 10 families displayed in Figure 9.3.

Family	$\hat{z}_{i0}$	$\hat{z}_{i1}$	$\hat{z}_{i2}$	$\hat{\tau}_i$	x_{i1}	x_{i2}	y_i
1	0	0	1	1	3	2	1
2	0	0	1	1	3.5	4	0.25
3	1	0	0	0	4	1	9
4	0.5	0.5	0	0.25	5	2	9
5	1	0	0	0	3.5	0	12.25
6	0	1	0	0.5	3.5	1	6.25
7	0	0.5	0.5	0.75	2	1	1
8	0.5	0.5	0	0.25	2.5	0	6.25
9	0	0.5	0.5	0.75	2.5	4	2.25
10	0.5	0	0.5	0.5	3	5	4
Mean				0.5			5.125

Table S.11 Additional data for estimating the regression slope in the 10 families displayed in Figure 9.3.

Family	$\hat{\tau}_i - \bar{\tau}$	$(\hat{\tau}_i - \bar{\tau})^2$	$y_i - \bar{y}$	$(\hat{\tau}_i - \bar{\tau})(y_i - \bar{y})$	$\hat{\varepsilon}_i$	$\hat{\varepsilon}_i^2$
1	0.50	0.2500	-4.125	-2.0625	1.075	1.1556
2	0.50	0.2500	-4.875	-2.4375	0.325	0.1056
3	-0.50	0.2500	3.875	-1.9375	-1.325	1.7556
4	-0.25	0.0625	3.875	-0.9688	1.275	1.6256
5	-0.50	0.2500	7.125	-3.5625	1.925	3.7056
6	0.00	0.0000	1.125	0.0000	1.125	1.2656
7	0.25	0.0625	-4.125	-1.0313	-1.525	2.3256
8	-0.25	0.0625	1.125	-0.2813	-1.475	2.1756
9	0.25	0.0625	-2.875	-0.7188	-0.275	0.0756
10	0.00	0.0000	-1.125	0.0000	-1.125	1.2656
Sum	0	1.25	0	-13	0	15.4563

against the estimated proportion of alleles shared IBD together with the resulting regression line.

Solution 9.2.1. The IBD distribution of the sib-pairs and the estimated proportion of alleles shared IBD by the sib-pairs is given in Table S.10.

Solution 9.2.2. The squared trait difference is given in Table S.10.

Solution 9.2.3. The relevant data for estimating the regression coefficient $\hat{\beta}$ for the HE regression are given in the last row of Table S.11 with $\bar{\hat{\tau}} = 0.45$ and $\bar{y} = 5.125$. Subsequently,

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^n (\hat{\tau}_i - \bar{\hat{\tau}})(y_i - \bar{y})}{\sum_{i=1}^n (\hat{\tau}_i - \bar{\hat{\tau}})^2} = \frac{-13}{1.25} \approx -10.4, \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{\hat{\tau}} \approx 5.125 + 10.4 \cdot 0.5 = 10.325, \\ \hat{\sigma}_\varepsilon^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 \approx \frac{1}{8} 15.4563 = 1.9320, \\ \widehat{\text{Var}}(\hat{\beta}) &= \frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (\hat{\tau}_i - \bar{\hat{\tau}})^2} \approx \frac{1.9320}{1.25} \approx 1.5456.\end{aligned}$$

The test statistic finally is

$$T = \hat{\beta} / \sqrt{\widehat{\text{Var}}(\hat{\beta})} \approx \frac{-10.4}{\sqrt{1.5456}} \approx \frac{-10.4}{1.2432} \approx -8.3653.$$

T is distributed as $t_{n-2} = t_8$ under H_0 , and the corresponding one-sided p -value is approximately $1.6 \cdot 10^{-5}$. T corresponds to an LOD of $(-8.3653)^2 (2 \ln(10)) \approx 15.20$. This result is exceptionally large and has been chosen only for illustration.

Solution 9.3 We calculate $\text{Var}(y_{i,2}) = \text{Var}((x_{i1} - x_{i2})^2 | \tau_i = 1)$ under the assumption that marker and trait locus are identical using an elegant method. A straightforward more cumbersome approach is used below for the solution of Problem 9.4. We assume that $(x_{i1}, x_{i2})'$ are bivariate normally distributed with $\mathbb{E}(x_{it}) = \mu_t$, $\text{Var}(x_{it}) = \sigma_e^2$, and $\text{Cov}(x_{i1}, x_{i2}) = \sigma_e^2 \varrho_{SS}$. With the notation of Section 9.2 we have $(x_{i1} - x_{i2}) \sim N((\mu_1 - \mu_2), 2\sigma_e^2(1 - \varrho_{SS}))$ so that $\frac{x_{i1} - x_{i2}}{\sqrt{2\sigma_e^2(1 - \varrho_{SS})}} \sim N(\mu_1 - \mu_2, 1)$. Subsequently, $\frac{(x_{i1} - x_{i2})^2}{2\sigma_e^2(1 - \varrho_{SS})} \sim \chi_{1, (\mu_1 - \mu_2)^2}^2$. By utilizing a standard result from the non-central χ^2 distribution, it follows that

$$\text{Var}\left(\frac{(x_{i1} - x_{i2})^2}{2\sigma_e^2(1 - \varrho_{SS})}\right) = 2 + 4(\mu_1 - \mu_2)^2.$$

$\mu_1 = \mu_2$ if $\tau_i = 1$, thus $\mu_1 - \mu_2 = 0$. We therefore obtain

$$\text{Var}\left(\frac{y_{i,2}}{2\sigma_e^2(1 - \varrho_{SS})}\right) = 2 \quad \Rightarrow \quad \text{Var}(y_{i,2}) = 8\sigma_e^4(1 - \varrho_{SS})^2.$$

Solution 9.4 For the derivation of the formula for $\text{Var}_{H_1}(y_{i,0})$, we use a standard technique and a more elegant method for $\text{Var}_{H_1}(y_{i,2})$. In the following, we omit the subscript H_1 . We start by calculating $\text{Var}(y_{i,0})$. To keep notations simple, we use

the following abbreviation: $C = \mathbb{E}((x_{i,1} - x_{i,2})^2 | \tau_i = 0) = \mathbb{E}(y_{i,0}) = \sigma_e^2 + 2\sigma_g^2$. Furthermore, we note that $P(g_{i1}, g_{i2} | \tau_i = 0)$ has been given in Table 9.2.

$$\begin{aligned}
 \text{Var}(y_{i,0}) &= \mathbb{E}\left(\sum_{k=1}^9 P(g_{i1}, g_{i2} | \tau_i = 0) ((x_{i,1} - x_{i,2})^2 - C)^2\right) \\
 &= \mathbb{E}\left((p^4 + q^4 + 4p^2q^2)(\varepsilon_i^2 - C)^2\right. \\
 &\quad + 2p^3q(((a-d+\varepsilon_i^2)^2 - C)^2 + ((-a+d+\varepsilon_i^2)^2 - C)^2) \\
 &\quad + 2pq^3(((a+d+\varepsilon_i^2)^2 - C)^2 + ((-a-d+\varepsilon_i^2)^2 - C)^2) \\
 &\quad \left.+ p^2q^2(((2a+\varepsilon_i^2)^2 - C)^2 + ((-2a+\varepsilon_i^2)^2 - C)^2)\right) = \dots \\
 &= \mathbb{E}((\varepsilon_i^4 - 2\varepsilon_i^2C + C^2) + pq(24a^2\varepsilon_i^2 - 8a^2C) + 32p^2q^2a^4 \\
 &\quad + 4p^3q((a-d)^4 - 12ad\varepsilon_i^2 + 4adC + 6d^2\varepsilon_i^2 - 2d^2C) \\
 &\quad + 4pq^3((a+d)^4 + 12ad\varepsilon_i^2 - 4adC + 6d^2\varepsilon_i^2 - 2d^2C)) \\
 &= \underbrace{4pq(p^2(a-d)^4 + q^2(a+d)^4 + 8pqa^4)}_{=: \lambda} \\
 &\quad + \mathbb{E}(\varepsilon_i^4) - 2C\mathbb{E}(\varepsilon_i^2) + C^2 \\
 &\quad + \mathbb{E}(pq(24a^2\varepsilon_i^2 - 8a^2C) \\
 &\quad - 48p^3qad\varepsilon_i^2 + 16p^3qadC + 24p^3qd^2\varepsilon_i^2 - 8p^3qd^2C \\
 &\quad + 48p^3qad\varepsilon_i^2 - 16p^3qadC + 24pq^3d^2\varepsilon_i^2 - 8pq^3d^2C)).
 \end{aligned}$$

We now need to determine $\mathbb{E}(\varepsilon_i^2)$ and $\mathbb{E}(\varepsilon_i^4)$. This can be achieved by utilizing standard formulae for normally distributed random variables and by noting that $\mathbb{E}(\varepsilon_i) = 0$ (see, e.g., http://en.wikipedia.org/wiki/Normal_distribution#Moments): $\mathbb{E}(\varepsilon_i^4) = 3\sigma_\varepsilon^4$ and $\mathbb{E}(\varepsilon_i^2) = \sigma_\varepsilon^2$. Subsequently, we obtain

$$\begin{aligned}
 \text{Var}(y_{i,0}) &= \dots \text{ Then a miracle occurs } \dots \\
 &= \lambda + 2\sigma_\varepsilon^4 + 4\sigma_g^4 + pq(\sigma_\varepsilon^2 - \sigma_g^2) \cdot \\
 &\quad (16a^2 - 32p^2ad + 16p^2d^2 + 32q^2ad + 16q^2d^2).
 \end{aligned}$$

Solution 10.1 We first observe $p_{22} \leq \min\{p_A, p_B\}$ for fixed allele frequencies p_A and p_B because haplotype frequencies cannot be larger than the underlying allele frequencies. Furthermore, $p_A + p_B - 1 = (p_{21} + p_{22}) + (p_{12} + p_{22}) - (p_{11} + p_{12} + p_{21} + p_{22}) = p_{22} - p_{11} \leq p_{22}$. Furthermore, $p_{22} \geq 0$ so that $p_{22} \geq \max\{0, p_A + p_B - 1\}$. With these two inequalities, it is easy to derive the restrictions for $\mathcal{D}_{AB}$. Thus, $\mathcal{D}_{AB} = p_{22} - p_{APB} \leq \min\{p_A - p_{APB}, p_B - p_{APB}\} = \min\{p_A q_B, p_B q_A\}$. The other restrictions can be derived analogously.

Solution 10.2

Solution 10.2.1. The result directly follows from

$$\begin{aligned}
 1 - p_A - p_B + p_{APB} + \mathcal{D}_{AB} &= 1 - p_A - p_B + p_{APB} + p_{22} - p_{APB} \\
 &= 1 - p_{22} - p_{12} - p_{21} = p_{11}.
 \end{aligned}$$

Solution 10.2.2. The result is obtained using straightforward algebra.

$$\begin{aligned}
 & 2p_A q_A p_B q_B + \mathcal{D}_{AB}(1 - 2p_A)(1 - 2p_B) + 2\mathcal{D}_{AB}^2 \\
 &= 2p_A q_A p_B q_B + (p_{22} - p_A p_B)(1 - 2p_A)(1 - 2p_B) + 2(p_{22} - p_A p_B)^2 \\
 &= p_A p_B + p_{22} - 2p_{22}p_A - 2p_{22}p_B + 2p_{22}^2 \\
 &= (p_{21} + p_{22})(p_{12} + p_{22}) + p_{22} - 2p_{22}(p_{12} + p_{22}) - 2p_2(p_{21} + p_{22}) + 2p_{22}^2 \\
 &= \dots \\
 &= p_{22}p_{11} + p_{12}p_{21}.
 \end{aligned}$$

Solution 10.3

$$y = \frac{(\text{OR} - 1)}{(\text{OR} + 1)} = \frac{\frac{p_{11}p_{22}}{p_{12}p_{21}} - 1}{\frac{p_{11}p_{22}}{p_{12}p_{21}} + 1} = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}}.$$

Solution 10.4

Solution 10.4.1. Given that $\mathcal{D}' = 1$, $\mathcal{D}_{AB}$ has to equal $\mathcal{D}_{max} = \min\{p_A q_B, q_A p_B\}$. We first assume that $\mathcal{D}_{max} = p_A q_B$. Then,

$$\Delta^2 = \frac{\mathcal{D}_{AB}^2}{p_A p_B q_A q_B} = \frac{p_A^2 q_B^2}{p_A p_B q_A q_B} = \frac{p_A(1 - p_B)}{p_B(1 - p_A)} = \frac{p_A - p_A p_B}{p_B - p_A p_B}.$$

The last formulation can equal 1 only if $p_A = p_B$. An analogous formulation is found if $\mathcal{D}_{max} = q_A p_B$.

Solution 10.4.2. Again, first assuming that $\mathcal{D}_{max} = p_A q_B$, we have

$$\mathcal{D}' = \frac{\mathcal{D}_{AB}}{\mathcal{D}_{max}} = \frac{p_{22} - p_A p_B}{p_A q_B} = \frac{p_{22} - p_A p_B}{p_A - p_A p_B}.$$

The last formulation shows that in this case, $\mathcal{D}'$ can equal 1 only if $p_{22} = p_A$, meaning that $p_{21} = 0$ and therefore the A allele always occurs in the presence of B . The same holds for $\mathcal{D}_{max} = q_A p_B$.

Solution 10.4.3. A simple example is $h_3 = 0$, $h_1 = h_4 = \frac{1}{4}$, $h_2 = \frac{1}{2}$.

Solution 10.5 First, Δ_{AB} is given by

$$\Delta_{AB} = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{(p_{11} + p_{12})(p_{11} + p_{21})(p_{12} + p_{22})(p_{21} + p_{22})}}. \quad (\text{S.6})$$

Second, the joint genotype distribution of (X, Y) of Table 10.2 can be represented as $(X, Y) = (A_1 + A_2, B_1 + B_2)$, where A_1 and B_1 (A_2 and B_2) count the number of A and B alleles on the paternally (maternally) inherited chromosome, respectively. Under HWE, (A_1, B_1) and (A_2, B_2) are independently and identically distributed implying

$$\begin{aligned}
 \mathbb{V}\text{ar}(X) &= \mathbb{V}\text{ar}(A_1) + \mathbb{V}\text{ar}(A_2) = 2p_A q_A = 2(p_{21} + p_{22})(p_{11} + p_{12}), \\
 \mathbb{V}\text{ar}(Y) &= \mathbb{V}\text{ar}(B_1) + \mathbb{V}\text{ar}(B_2) = 2p_B q_B = 2(p_{12} + p_{22})(p_{11} + p_{21}).
 \end{aligned}$$

The covariance $\mathcal{D}_{XY}$ is similarly obtained as the sum of the covariance of two dependent binomials:

$$\mathcal{D}_{XY} = \text{Cov}(X, Y) = \text{Cov}(A_1, B_1) + \text{Cov}(A_2, B_2) = 2(p_{11}p_{22} - p_{12}p_{21})$$

The genotype-based Pearson product-moment correlation between the two SNPs is thus given by

$$\Delta_{XY} = \text{Corr}(X, Y) = \frac{2(p_{11}p_{22} - p_{12}p_{21})}{4(p_{11} + p_{12})(p_{21} + p_{22})(p_{11} + p_{21})(p_{12} + p_{22})}$$

which is identical to Eq. (S.6).

Solution 10.6 Assuming the absence of mutations, an $\frac{A}{B}$ haplotype can be observed in generation $t + 1$ only if the parent in generation t has at least one A allele and one B allele. Therefore, only those parental haplotype combinations are possible that are displayed together with their occurrence probabilities in Figure S.4. These probabilities are added up, with p_{22} denoting $p_{22,t}$ in generation t :

$$\begin{aligned} p_{22,t+1} &= \underbrace{p_{22}p_{22}}_{\text{Case 1}} + \underbrace{2p_{21}p_{22}\frac{1}{2}}_{\text{Case 2}} + \underbrace{2p_{12}p_{22}\frac{1}{2}}_{\text{Case 3}} \\ &\quad + \underbrace{2p_{11}p_{22}\frac{1}{2}(1-\theta)}_{\text{Case 4}} + 0 + 0 + \underbrace{2p_{12}p_{21}\frac{1}{2}\theta}_{\text{Case 5}} \\ &= p_{22} - \theta p_{11}p_{22} + \theta p_{12}p_{21} \\ &= p_{22} - \theta(p_{22} - p_{AB}) = p_{22} - \theta \mathcal{D}_t \end{aligned} \quad (\text{S.7})$$

In Eq. (S.7), $p_{22,t+1}$ denotes the probability of haplotype $\frac{A}{B}$ in generation $t + 1$. The haplotype probabilities without the time index are based on generation t . It follows that $\mathcal{D}_{t+1} - \mathcal{D}_t = p_{22,t+1} - p_{22} = -\theta \mathcal{D}_t$ so that finally $\mathcal{D}_{t+1} = (1 - \theta)\mathcal{D}_t$.

Solution 10.7

Solution 10.7.1. The genotype frequencies for the entire group are shown in Table S.12. From this, the allele frequencies can be estimated by

$$\begin{aligned} \hat{p}_G &= \frac{2 \cdot 5 + 7}{30} \approx 0.5667, & \hat{p}_A &= \frac{7 + 2 \cdot 3}{30} \approx 0.4333, \\ \hat{p}_C &= \frac{2 \cdot 4 + 9}{30} \approx 0.5667, & \hat{p}_T &= \frac{9 + 2 \cdot 2}{30} \approx 0.4333. \end{aligned}$$

Solution 10.7.2. The haplotypes can be determined for all individuals who are homozygous at all but one locus. Hence, in these data, the three individuals who are doubly heterozygous have an unknown haplotype. Disregarding these individuals, the haplotypes can be determined (see Table S.13), and the haplotype frequency estimates are

$$\begin{aligned} \hat{p}_{GC} &= \frac{9}{24} \approx 0.3750, & \hat{p}_{AC} &= \frac{5}{24} \approx 0.2083, \\ \hat{p}_{GT} &= \frac{5}{24} \approx 0.2083, & \hat{p}_{AT} &= \frac{5}{24} \approx 0.2083. \end{aligned}$$

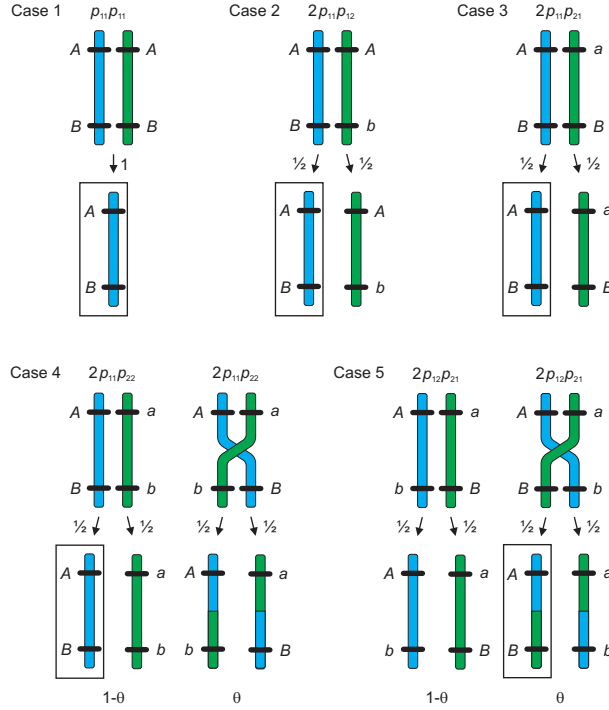


Fig. S.4 Decay of linkage disequilibrium from one generation to the other. The figure displays the possible paternal haplotype combinations that could yield an $\begin{smallmatrix} A \\ B \end{smallmatrix}$ offspring haplotype. The occurrence probabilities of the parental haplotypes are also given. Furthermore, it is shown whether or not a recombination event results in an $\begin{smallmatrix} A \\ B \end{smallmatrix}$ offspring haplotype.

Solution 10.7.3. For determining allelic LD measures, we have set up Table S.14, which includes the gametic frequencies in the body and the allele frequencies in the row and column totals. Using the numbers of this table, different allelic LD measures can be estimated, for example,

$$\begin{aligned}\hat{D}_{GC} &\approx 0.3750 - 0.5833 \cdot 0.5833 = 0.0347, \\ \hat{D}_{\max} &\leq \min\{0.5833 \cdot 0.4167, 0.4167 \cdot 0.5833\} = 0.2431, \\ \hat{D}' &\approx 0.0347/0.2431 = 0.1428, \\ \hat{\Delta}^2 &\approx \frac{0.0347^2}{0.4167^2 \cdot 0.5833^2} = 0.0204.\end{aligned}$$

Solution 10.7.4. With $\hat{D}_{GC} = 0.0347$ and the results from above we obtain $\sqrt{\widehat{\text{Var}}_{H_0}(\hat{D}_{GC})} = \sqrt{\frac{1}{2n}(\hat{p}_G \hat{q}_G \hat{p}_C \hat{q}_C)} = \sqrt{\frac{1}{24} 0.5833^2 0.4167^2} = 0.0496$ and $t_S = 0.0347/0.0496 = 0.6994$ as score test statistic. For the Wald test, we get

Table S.12 Genotype frequencies at two single nucleotide polymorphisms (SNPs) for the entire group.

SNP 2	SNP 1			Total
	AA	GA	GG	
TT	1	1	0	2
CT	2	3	4	9
CC	0	3	1	4
Total	3	7	5	15

Table S.13 Observed haplotypes at two single nucleotide polymorphisms (SNPs) in the entire group, excluding doubly heterozygous individuals.

SNP 2	SNP 1		Total
	G	A	
C	9	5	14
T	5	5	10
Total	14	10	24

$$\sqrt{\widehat{\text{Var}}(\widehat{\mathcal{D}}_{GC})} = \sqrt{\frac{1}{2n}(\hat{p}_G \hat{q}_G \hat{p}_C \hat{q}_C + (1 - 2\hat{p}_G)(1 - 2\hat{p}_C)\widehat{\mathcal{D}}_{GC} - \widehat{\mathcal{D}}_{GC}^2)} = 0.0495$$

so that $t_W = 0.7008$.

Solution 10.7.5. The asymptotic 95% confidence interval for $\mathcal{D}_{GC}$ is given by $\widehat{\mathcal{D}}_{GC} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\widehat{\mathcal{D}}_{GC})}$ that is $0.0347 \pm 1.96 \cdot 0.0495 = [-0.0623; 0.1317]$.

Solution 10.8 This problem can be solved by using the approach described in detail in Example 10.2.

Solution 10.8.1. $\hat{\Delta} = 0.7331$.

Solution 10.8.2. With the empirical means $\bar{x} = 0.8831$ and $\bar{y} = 0.5584$ and the empirical marginal variances $s_x^2 = 0.6309$ and $s_y^2 = 0.4604$, one obtains $\hat{\zeta}_{ij}$ and $\hat{\pi}_{ij}$ for $i, j = 0, 1, 2$. These are given by

$$\begin{aligned} [\hat{\zeta}_{ij}]_{i,j=0}^2 &= \begin{pmatrix} 0.0000 & -1.5456 & -4.6837 \\ -0.5910 & -0.2811 & -1.5636 \\ -2.3441 & -0.1786 & 0.3945 \end{pmatrix}, \\ [\hat{\pi}_{ij}]_{i,j=0}^2 &= \begin{pmatrix} 0.3766 & 0.0000 & 0.0000 \\ 0.1299 & 0.2338 & 0.0000 \\ 0.0390 & 0.1169 & 0.1039 \end{pmatrix}. \end{aligned}$$

Subsequently, one can obtain the estimator of the asymptotic variance of $\hat{\Delta}$ as $\widehat{\text{Var}}(\hat{\Delta}) = 0.0033$.

Solution 10.8.3. The z transformed correlation coefficient and its asymptotic variance are $\hat{\gamma} = 0.9354$ and $\widehat{\text{Var}}(\hat{\gamma}) = 0.01530$. The asymptotic 95% confidence interval for γ subsequently is $[0.6929, 1.1778]$. After back transformation, the asymptotic 95% confidence interval for Δ is obtained as $[0.5998, 0.82678]$.

Solution 11.1 They talk very quietly because they are discrete. Indeed, several modifications of test statistic (11.1) and further test statistics utilizing the continuous χ^2 distribution have been made to account for the discreteness of the underlying exact distribution. Alternatively, the exact test of Fisher for comparing two binomial distributions is often applied in this context although it is conservative.

Table S.14 Estimated haplotype and allele frequencies of Problem 10.8.3.

SNP 2	SNP 1		
	A	G	
T	0.2083	0.2083	0.4167
C	0.2083	0.3750	0.5833
	0.4167	0.5833	

Table S.15 Genotype frequencies at the notional single nucleotide polymorphism (SNP) 1 for asthma patients and healthy controls assuming a recessive mode of inheritance.

	GG or GA	AA	Total
Cases	6	2	8
Controls	6	1	7
Total	12	3	15

Solution 11.2 χ_c^2 is given by $\sum_{k=1}^4 \frac{(o_k - e_k)^2}{e_k}$ with the notation of Table 11.3. e_1 was obtained as $(2n_2 + n_1) \frac{2r}{2n}$ and is also identical to $e_1 = \frac{(o_1 + o_2)(o_1 + o_3)}{2n}$. With noting that $2n = \sum_{k=1}^4 o_k$, we obtain

$$\begin{aligned} (o_1 - e_1)^2 &= \left(o_1 - \frac{(o_1 + o_2)(o_1 + o_3)}{2n} \right)^2 = \left(\frac{2o_1n - o_1^2 - o_1o_2 - o_1o_3 - o_2o_3}{2n} \right)^2 \\ &= \left(\frac{o_1(\sum_{k=1}^4 o_k) - o_1^2 - o_1o_2 - o_1o_3 - o_2o_3}{2n} \right)^2 = \frac{(o_1o_4 - o_2o_3)^2}{4n^2}. \end{aligned}$$

Analogously, we can show $(o_2 - e_2)^2 = (o_3 - e_3)^2 = (o_4 - e_4)^2 = \frac{(o_1o_4 - o_2o_3)^2}{4n^2}$ so that χ_c^2 may be rewritten as

$$\chi_c^2 = \frac{(o_1o_4 - o_2o_3)^2}{4n^2} \left(\frac{1}{e_1} + \frac{1}{e_2} + \frac{1}{e_3} + \frac{1}{e_4} \right). \quad (\text{S.8})$$

By using the notation of Table 11.2, the second term of this equation can be expressed as

$$\begin{aligned} \frac{1}{e_1} + \frac{1}{e_2} + \frac{1}{e_3} + \frac{1}{e_4} &= \frac{2n}{2r(2n_2 + n_1)} + \frac{2n}{2r(2n_0 + n_1)} + \frac{2n}{2s(2n_2 + n_1)} + \frac{2n}{2s(2n_0 + n_1)} \\ &= \frac{2n(2s(2n_0 + n_1) + 2s(2n_2 + n_1) + 2r(2n_0 + n_1) + 2r(2n_2 + n_1))}{2r \cdot 2s \cdot (2n_0 + n_1) \cdot (2n_2 + n_1)} \\ &= \frac{2n(2s \cdot 2n + 2r \cdot 2n)}{2r \cdot 2s \cdot (2n_0 + n_1) \cdot (2n_2 + n_1)} = \frac{2n \cdot 4n^2}{2r \cdot 2s \cdot (2n_0 + n_1) \cdot (2n_2 + n_1)}, \end{aligned}$$

which completes the proof after insertion of this result in Eq. (S.8).

Solution 11.3 We assume that the genotype x_j is random, and we consider the score function $S = \sum_{j=1}^n (y_j - \bar{y})x_j$, where $\bar{y}$ is the phenotypic mean in the entire sample. This score function approach can be used for quantitative traits as well. Here, we only consider dichotomous traits. For these, $\bar{y} = r/n$ is the proportion

of cases, and the score function can be re-formulated as $S = \sum_{i=0}^2 r_i x_i - r\bar{x}$. In the case of retrospective sampling, $P(x_j|y_j)$ is the distribution of interest, and in this case the variance of S is $\mathbb{V}ar(S) = \mathbb{V}ar(x_j) \sum_{j=1}^n (y_j - \bar{y})^2$. We assume that $V_x = \mathbb{V}ar(x_j)$ is constant for all subjects so that it can be estimated by maximum likelihood as $\hat{V}_x = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$.

To complete the proof, we note that $\sum_{j=1}^n (x_j - \bar{x})^2 = \sum_{i=0}^2 n_i (x_i - \bar{x})^2$ and that $\sum_{j=1}^n (y_j - \bar{y})^2 = \sum_{j=1}^n y_j^2 - n\bar{y}^2 = r - r^2/n = rs/n$. Finally, we remark that the variance of S is identical for prospective sampling [127].

Solution 11.4 Assuming a recessive mode of inheritance, the genotypes can be summarized as shown in Table S.15. The OR estimate associated with the recessive genotype at SNP 1 is given by $\text{OR} = (2 \cdot 6)/(1 \cdot 6) = 2.0$. The exact p -value for the test $H_0 : \text{OR} = 1$ against $H_1 : \text{OR} \neq 1$ is equal to 1, and an exact 95% confidence interval is $[0.0785; 137.3531]$. Although the approximation of the log OR to the normal distribution is bad because of the low sample size, we illustrate the calculation of the asymptotic statistics. The corresponding χ^2 statistic forming the basis of the test for association is calculated as

$$\chi^2 = 15 \cdot \frac{(2 \cdot 6 - 1 \cdot 6)^2}{8 \cdot 7 \cdot 12 \cdot 3} = 0.2679.$$

Based on the approximated χ^2 distribution with 1 d.f., the associated p -value is $p \approx 0.6047$. At a significance level of $\alpha = 0.05$, the null hypothesis of no association is not rejected. Finally, an approximate 95% confidence interval is given by

$$\text{CI}(\text{OR}) = \left[2.0 \cdot \exp\left(\pm 1.96 \sqrt{\frac{1}{2} + \frac{1}{6} + \frac{1}{6} + \frac{1}{1}}\right) \right] \approx [0.1408; 28.4174].$$

We again stress that the standard rule of thumb for approximating the distribution of the OR to the normal distribution—there are at least 5 observations per cell—is not fulfilled because of the low sample size.

Solution 11.5 The X chromosomal trend test statistic can be derived using the arguments used in the proof of Solution 11.3. We assume that the genotype x_j is random, and we consider the *score function* $S = \sum_{j=1}^n (y_j - \bar{y})x_j$, where $\bar{y}$ is the phenotypic mean in the entire sample. In the case of retrospective sampling, $P(x_j|y_j)$ is the distribution of interest, and in this case the variance of S is $\mathbb{V}ar(S) = \mathbb{V}ar(x_{j,f}) \sum_{j=1}^{n_f} (y_j - \bar{y})^2 + \mathbb{V}ar(x_{j,m}) \sum_{j=n_f+1}^n (y_j - \bar{y})^2$. We assume that $V_{x,f} = \mathbb{V}ar(x_{j,f})$ and $V_{x,m} = \mathbb{V}ar(x_{j,m})$ are constant within females and males, respectively. $V_{x,f} = \mathbb{V}ar(x_{j,f})$ can be estimated by maximum likelihood analogously to V_x in Solution 11.3, i.e., $\hat{V}_{x,f} = \frac{1}{n_f} \sum_{j=1}^{n_f} (x_j - \bar{x})^2$. Males have only a single allele so that the variance can be expressed in terms of the binomial distribution $p(1-p)$. The factor 4 is obtained because of the 0 and 2 scores for males rather than the standard 0 and 1 scores for the binomial distribution. The variance in males therefore is $4p(1-p)$.

Solution 11.6

Solution 11.6.1. Assuming the different allele frequencies of 0.4, 0.2, and 0.05 in the controls, the allele frequencies in the cases and the mean allele frequencies are calculated to be

$$\begin{aligned} p_{A,u} = 0.4 : p_{A,a} &= \frac{0.4 \cdot 2}{1+0.4 \cdot (2-1)} \approx 0.5714 \quad \text{and} \quad \bar{p} = \frac{1}{2}(0.5714+0.4) \approx 0.4857, \\ p_{A,u} = 0.2 : p_{A,a} &= \frac{0.2 \cdot 2}{1+0.2 \cdot (2-1)} \approx 0.3333 \quad \text{and} \quad \bar{p} = \frac{1}{2}(0.3333+0.2) \approx 0.2667, \\ p_{A,u} = 0.05 : p_{A,a} &= \frac{0.05 \cdot 2}{1+0.05 \cdot (2-1)} \approx 0.0952 \quad \text{and} \quad \bar{p} = \frac{1}{2}(0.0952+0.05) \approx 0.0726. \end{aligned}$$

With $z_{0.995} = 2.5758$ and $z_{0.8} = 0.8416$, this renders the respective required sample sizes to be

$$n \geq 2 \cdot \frac{\left(2.5758\sqrt{2 \cdot 0.4857(1-0.4857)} + 0.8416\sqrt{0.5714(1-0.5714)+0.4(1-0.4)}\right)^2}{(0.5714-0.4)^2} \approx 394.1930,$$

thus $n \geq 395$ for $p_{A,u} = 0.4$. Similarly, we obtain $n \geq 511 \geq 510.9865$ for $p_{A,u} = 0.2$ and $n \geq 1535 \geq 1534.4424$ for $p_{A,u} = 0.05$.

Solution 11.6.2. If the allele frequencies at the marker and the disease loci both equal 0.2, $\mathcal{D}_{max} = 0.16$. In the case that allele frequencies are identical and $\mathcal{D} = \mathcal{D}_{max}$, the OR of the marker allele will also equal the OR of the disease allele. However, assuming lower levels of LD with $\mathcal{D} = \frac{2}{3}\mathcal{D}_{max} \approx 0.1067$ or $\mathcal{D} = \frac{1}{2}\mathcal{D}_{max} = 0.08$, the OR of the marker allele reduces from $OR_D = 2.0$ to

$$OR_A \approx 1 + \frac{0.1067(2-1)}{0.2\left((1-0.2) + (0.2(1-0.2) - 0.1067)(2-1)\right)} \approx 1.6252$$

for $\mathcal{D} = 0.1067$, and, for $\mathcal{D} = 0.08$, to

$$OR_A \approx 1 + \frac{0.08(2-1)}{0.2\left((1-0.2) + (0.2(1-0.2) - 0.08)(2-1)\right)} \approx 1.4545.$$

What happens if the $\mathcal{D} = \mathcal{D}_{max}$, but the allele frequencies at the disease and the marker locus differ? Considering allele frequencies of 0.3 and 0.4 at the marker in contrast to the frequency of 0.2 at the disease locus renders the following maximal values for $\mathcal{D}$:

$$\begin{aligned} p_{A,u} = 0.3 : \mathcal{D}_{max} &= \min\{0.2(1-0.3), 0.3(1-0.2)\} = 0.14, \\ p_{A,u} = 0.4 : \mathcal{D}_{max} &= \min\{0.2(1-0.4), 0.4(1-0.2)\} = 0.12. \end{aligned}$$

Hence, assuming $\mathcal{D} = \mathcal{D}_{max}$, the OR of the marker allele is estimated to be

$$OR_A = 1 + \frac{0.14(2-1)}{0.3\left((1-0.3) + (0.2(1-0.3) - 0.14)(2-1)\right)} = 1.7778$$

for $p_{A,u} = 0.3$, and, for $p_{A,u} = 0.4$, to be

$$OR_A = 1 + \frac{0.12(2-1)}{0.4\left((1-0.4) + (0.2(1-0.4) - 0.12)(2-1)\right)} = 1.6.$$

Solution 11.7

Solution 11.7.1. Application of the tool HWG-TEST with 17,000 replicates revealed a p -value of 0.27 in the controls and 0.017 in the cases. Thus, there is evidence for deviation from HWE in the cases at the 5% test level.

Solution 11.7.2. The sum of the 20 χ^2 statistics is $\chi_P^2 \approx 31.6126$. The p -value of the χ^2 statistic with 20 d.f. consequently is approximately 0.0476; thus, there is some evidence for population stratification at the 5% test level.

Solution 11.7.3. The median of the 20 χ^2 statistics is $(0.2750 + 0.3573)/2 = 0.3162$ so that $\hat{\lambda} \approx 0.3162/0.4549 \approx 0.6950$. The bounded variance inflation factor therefore is 1.

Solution 11.7.4. By using the data displayed in Table 11.28, we calculate the following test statistics $\chi_c^2 = 37.1816$, $\chi_A^2 = 37.1816$, $\chi_g^2 = 27.7282$, $\chi_{dom}^2 = 21.6925$, $\chi_{rec}^2 = 20.0638$, $\chi_{het}^2 = 0.3439$, $\chi_{trend}^2 = 27.5419$. χ_{het}^2 yields no significant result ($p = 0.5576$). All other tests give p -values of less than 10^{-5} . χ_c^2 and χ_A^2 even reveal a p -value of about 10^{-9} . No adjustments for population stratification are required because the bounded estimate of the variance inflation factor is 1.

Solution 11.7.5. The main problem is the multiple testing problem that is not taken into account. Tests are carried out one after another, without specifying the consequences of the test procedure. Thus, it is unclear what should have been done if a deviation from HWE had been detected in the controls. Similarly, the trend test has been formulated as the one of interest in the Problems. However, in applications it often remains unclear whether the investigators have chosen the trend test because of a significant deviation from HWE in either study group. Instead of calculating all test statistics and comparing results, one should, beforehand, carefully choose an appropriate test statistic. If deviation from HWE has been detected, χ_c^2 and χ_A^2 must not be used as they can lead to inflated type I error levels. Also, if one has a specific knowledge about the underlying mode of inheritance at the observed SNP, one should rather decide for the most accurate test than trying all. Badly chosen tests will have little power to detect an association, whereas correct selection of the underlying genetic model and corresponding test statistic will maximize power.

In summary, often there seems to be a lack of a statistical analysis plan (SAP), which is standard in controlled clinical trials.

Solution 11.8 In the sample of the replication study, 17 out of 76 developed sepsis after polytrauma, i.e., $\hat{p} = 0.2237$, and the 95% confidence interval (CI) is $CI(p) = [0.1423; 0.3317]$. The risk difference is estimated to be $\widehat{RD} = 0.2752$ (95% CI: $CI(RD) = [0.0230; 0.5042]$).

The three point estimates of the relative risk are $\widehat{RR}_G = 2.0824$, $\widehat{RR}_{G,0.5} = 2.0839$ and $\widehat{RR}_{G,PB} = 1.9652$. The 95% CI is $CI(RR_{G,0.5}) = [1.1360; 3.8227]$, and the 95% credible interval according to Price and Bonett [542] is $CI(RR_{G,PB}) = [1.0449; 3.6963]$. From these we obtain for the attributable risk in variant carriers $\widehat{AR}_G = 0.5198$, $\widehat{AR}_{G,0.5} = 0.5201$, $\widehat{AR}_{G,PB} = 0.4912$, yielding $CI(AR_G) =$

$[0.1197; 0.7384]$ as 95% CI and $CI(AR_{G,PB}) = [0.0429; 0.7295]$ as 95% credible interval.

The attributable risk should not be estimated using the formulae provided in this textbook because the replication study is a cohort study, and the formulae are only applicable to case-control studies.

For the odds ratio we obtain $\widehat{OR} = 3.3000$ as point estimate and $CI(OR) = [1.0786; 10.0960]$ as 95% confidence interval.

Solution 12.1

Solution 12.1.1. There are 2·10 transmissions from 10 trios with 16 transmissions from heterozygous parents. In these, there are 13 instances in which allele T was transmitted while allele C was not transmitted. By comparison, there were three cases in which allele C was transmitted and T was not transmitted. Hence, T_{TDT} is given by $T_{TDT} = (13 - 3)^2 / (13 + 3) = 6.25$. The corresponding p -value from the χ_1^2 distribution is approximately 0.0124.

In four situations, the allele T was not transmitted. Hence, the estimated frequency of the T allele is $\hat{p} = 4 / (2 \cdot 10) = 0.2$. The numbers of trios in which 0, 1 or 2, T alleles were transmitted to the affected offspring are $n_0 = 1$, $n_1 = 4$ and $n_2 = 5$. From this, we estimate the GRRs and the PAR by

$$\hat{\gamma}_1 = \frac{4 \cdot (1-0.2)}{2 \cdot 1 \cdot 0.2} = 8, \hat{\gamma}_2 = \frac{5 \cdot (1-0.2)^2}{1 \cdot (0.2)^2} = 80, \text{ and } \widehat{PAR} = 1 - \frac{4 \cdot 10 \cdot 1}{(2 \cdot 10 - 4)^2} \approx 0.8438.$$

To calculate the 95% confidence intervals for γ_1 and γ_2 , we first compute the abbreviations a and b to be $a = 1 + 0.2 \cdot (2 \cdot 8 - 1) = 4$ and $b = 1 + 0.2 + 0.2^2 \cdot (8 - 1) = 1.48$. Then, the asymptotic 95% confidence intervals for γ_1, γ_2 and PAR are given by

$$\begin{aligned} 8 \cdot \exp\left(\pm 1.96 \sqrt{\frac{4 \cdot 80 \cdot 0.2^2 + (4^2 + 8)(1-0.2)}{2 \cdot 10 \cdot 0.2 \cdot (1-0.2)^2 \cdot 8}}\right) &\approx [0.6903; 92.7068], \\ 80 \cdot \exp\left(\pm 1.96 \sqrt{\frac{80^2 \cdot 0.2^4 + 2 \cdot 1.4800 \cdot 80 \cdot 0.2 \cdot (1-0.2) + 4 \cdot (1-0.2)^3}{10 \cdot 0.2^2 \cdot (1-0.2)^2 \cdot 80}}\right) &\approx [3.7216; 1719.7116], \\ 0.8438 \pm 1.96 \sqrt{\frac{0.2}{10 \cdot (1-0.2)^2} \cdot \frac{2 \cdot (1-0.2)(1+8) + 80 \cdot 0.2}{(80 \cdot 0.2^2 + 4 \cdot (1-0.2))^2}} &\approx [0.5453; 1.1423]. \end{aligned}$$

Solution 12.1.2. If there is missing parental data, the 1-TDT might be employed. In the families with missing paternal genotypes, there are five cases where the child is homozygous TT and the mother is heterozygous TC, and one case in the child has TC and the mother TT. Further, in one instance the child is homozygous with C and the mother heterozygous, and there is no case of the mother being homozygous TT. The 1-TDT statistic is therefore given by $T_1 = ((5+1) - (0+1))^2 / ((5+1) + (0+1)) \approx 3.5714$ with an associated p -value from the χ_1^2 distribution of 0.0588.

Solution 12.2 Abbreviating affected with aff and genotype with g, application of Bayes' theorem to $P(DN|aff)$ and $P(NN|aff)$ renders

$$P(DN|aff) = \frac{P(aff|DN)P(DN)}{\sum P(aff|g_i)} \Leftrightarrow \frac{P(DN|aff) \sum P(aff|g_i)}{P(DN)} = P(aff|DN),$$

$$P(NN|\text{aff}) = \frac{P(\text{aff}|NN)P(NN)}{\sum P(\text{aff}|g_i)} \Leftrightarrow \frac{P(NN|\text{aff}) \sum P(\text{aff}|g_i)}{P(NN)} = P(\text{aff}|NN).$$

Inserting this into Eq. (2.2) for γ_1 gives

$$\frac{P(\text{aff}|DN)}{P(\text{aff}|NN)} = \frac{P(DN|\text{aff}) \sum P(\text{aff}|g_i)}{P(DN)} \cdot \frac{P(NN)}{P(NN|\text{aff}) \sum P(\text{aff}|g_i)} = \frac{P(DN|\text{aff})}{P(DN)} \cdot \frac{P(NN)}{P(NN|\text{aff})}.$$

The same procedure can be used for deriving the alternative formula for γ_2 by using $P(DD|\text{aff})$ instead of $P(DN|\text{aff})$.

Solution 12.3 Let Y_1 and Y_2 denote the number of AA and Aa genotypes in the affected offspring, respectively, so that $Y = 2Y_1 + Y_2$. Utilizing a hypergeometric distribution, we obtain $\mathbb{E}(Y_1) = \sum_{i=1}^n n_{AA,i} \frac{n_{\text{aff},i}}{n_i}$ and $\mathbb{E}(Y_2) = \sum_{i=1}^n n_{Aa,i} \frac{n_{\text{aff},i}}{n_i}$, yielding

$$\mathbb{E}(Y) = \sum_{i=1}^n (2n_{AA,i} + n_{Aa,i}) \frac{n_{\text{aff},i}}{n_i}.$$

The variances of Y_1 and Y_2 and the covariance between Y_1 and Y_2 are given by

$$\mathbb{V}\text{ar}(Y_1) = \sum_{i=1}^n \frac{n_i - n_{AA,i}}{n_i - 1} n_{AA,i} \frac{n_{\text{aff},i}}{n_i} \left(1 - \frac{n_{\text{aff},i}}{n_i}\right) = \sum_{i=1}^n n_{\text{aff},i} n_{\text{unaff},i} \frac{(n_i - n_{AA,i}) n_{AA,i}}{n_i^2 (n_i - 1)},$$

$$\mathbb{V}\text{ar}(Y_2) = \sum_{i=1}^n \frac{n_i - n_{Aa,i}}{n_i - 1} n_{Aa,i} \frac{n_{\text{aff},i}}{n_i} \left(1 - \frac{n_{\text{aff},i}}{n_i}\right) = \sum_{i=1}^n n_{\text{aff},i} n_{\text{unaff},i} \frac{(n_i - n_{Aa,i}) n_{Aa,i}}{n_i^2 (n_i - 1)},$$

$$\mathbb{C}\text{ov}(Y_1, Y_2) = - \sum_{i=1}^n \frac{n_{AA,i} n_{Aa,i} n_{\text{aff},i} n_{\text{unaff},i}}{n_i^2 (n_i - 1)},$$

yielding

$$\begin{aligned} \mathbb{V}\text{ar}(Y) &= 4 \cdot \mathbb{V}\text{ar}(Y_1) + \mathbb{V}\text{ar}(Y_2) + 4 \cdot \mathbb{C}\text{ov}(Y_1, Y_2) \\ &= \sum_{i=1}^n n_{\text{aff},i} n_{\text{unaff},i} \frac{4n_{AA,i}(n_i - n_{AA,i} - n_{Aa,i}) + n_{Aa,i}(n_i - n_{Aa,i})}{n_i^2 (n_i - 1)}. \end{aligned}$$

Solution 13.1 In the first step of the algorithm, unambiguous haplotypes are identified: For this, we note that genotype combinations 1 and 4 are homozygous at

all three loci, so that the haplotypes $h_1 = \begin{smallmatrix} G \\ C \\ A \end{smallmatrix}$ and $h_2 = \begin{smallmatrix} G \\ A \\ G \end{smallmatrix}$ are known and resolved.

Furthermore, genotype combination 2 is heterozygous only at the second SNP so that in addition to the already resolved haplotype h_1 , there must be a haplotype $h_3 = \begin{smallmatrix} G \\ A \\ A \end{smallmatrix}$.

In the second step, we check whether any of the three haplotypes thus resolved is part of one of the three yet unresolved genotype combinations. We find that haplotype h_1 could be part of genotype combinations 3 and 5, and that haplotype h_2 may be part of genotype combination 6. The remaining alleles of genotype combination 3 form the haplotype h_2 . Identifying the other remaining alleles as newly resolved

haplotypes, we yield haplotype $h_4 = \frac{A}{C}$ from both genotype combinations 5 and 6.

As all haplotypes are now resolved, the algorithm is stopped.

Solution 13.2

Solution 13.2.1. Three squared correlations exceed the threshold of $\Delta^2 \geq 0.5$: Δ^2 between T-1486C and G1174A = 0.5519, between T-1486C and G2848A = 0.5595, and between G1174A and G2848A = 0.9859. Hence, these three are grouped into one bin, whereas the other two SNPs stay unbinned. Note that the binned SNPs are not contiguous, reflecting the irregular structure of LD. Owing to higher correlations, SNP G2848A would be selected as tagSNP for the bin.

Solution 13.2.2. As its lowest correlation with the omitted SNPs is $\Delta = 0.7480$, the sample size would need to be increased in the order of $1/0.7480^2$, i.e., by about 80%. However, this only considers the correlation with the omitted SNPs but not with hidden SNPs that were not typed in the first place.

Solution 14.1 The positive predictive value (ppV) is defined as the probability to have the disease ($D = 1$) given that the proband is carrier of the susceptibility variant ($G = 1$). According to Bayes theorem, this can be formulated as follows:

$$ppV = P(D=1|G=1) = \frac{P(G=1|D=1) \cdot P(D=1)}{P(G=1|D=1) \cdot P(D=1) + P(G=1|D=0) \cdot P(D=0)}.$$

Based on the definitions of sensitivity ($sens$) and specificity ($spec$) as in Section 14.6.2 to be $sens = P(G=1|D=1)$ and $spec = 1 - P(G=1|D=0)$ as well as the prevalence ($prev$) being the probability for disease $P(D=1)$, this leads to

$$ppV = \frac{sens \cdot prev}{sens \cdot prev + (1 - spec) \cdot (1 - prev)}.$$

The formulation for the negative predictive value (npV) can be derived analogously.

Solution 14.2 The sensitivity ($sens$) and specificity ($spec$) equal 0.79 and $1 - 0.73 = 0.27$. Thus, a test based on the typing of a single SNP would show a moderate sensitivity but an insufficient specificity.

The positive predictive value is given by $ppV = \frac{0.79 \cdot 0.1}{0.79 \cdot 0.1 + 0.73 \cdot 0.9} = 0.11$. The negative predictive value is $npV = \frac{0.23 \cdot 0.9}{0.23 \cdot 0.9 + 0.21 \cdot 0.1} = 0.91$. Thus, the probability for a myocardial infarction if one carries at least one variant is about 11%, whereas it is about $100 - 91\% = 9\%$ if one does not carry a respective variant. Considering that the base risk for myocardial infarction is the prevalence of 10%, only very little information is gained through this test.

Solution A.1 In the first step we apply the definition of $\text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'})$ and use $\mathbb{E}(\hat{\tau}_j) = \mathbb{E}(\hat{\tau}_{j'}) = \frac{1}{2}$:

$$\text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'}) = \mathbb{E}((\hat{\tau}_j - \mathbb{E}(\hat{\tau}_j))(\hat{\tau}_{j'} - \mathbb{E}(\hat{\tau}_{j'}))) = \mathbb{E}(\hat{\tau}_j \hat{\tau}_{j'}) - \frac{1}{4}$$

With $\mathbb{V}\text{ar}(\hat{\tau}_j) = \mathbb{V}\text{ar}(\hat{\tau}_{j'}) = \frac{1}{8}$ we furthermore obtain

$$\begin{aligned}\mathbb{E}(\hat{\tau}_j, \hat{\tau}_{j'}) &= \sum_{i=0, \frac{1}{2}, 1} \sum_{l=0, \frac{1}{2}, 1} \tau_i \tau_l P(\hat{\tau}_j = i, \hat{\tau}_{j'} = l) \\ &= \sum_{i=0, \frac{1}{2}, 1} \sum_{l=0, \frac{1}{2}, 1} \tau_i \tau_l P(\hat{\tau}_j = i | \hat{\tau}_{j'} = l) P(\hat{\tau}_{j'} = l) .\end{aligned}$$

Here, $P(\hat{\tau}_{j'} = l)$ is the a priori probability for an ASP sharing $2l$ alleles IBD, and $P(\hat{\tau}_j = i | \hat{\tau}_{j'} = l)$ can be determined using Table 8.6. The double sum involves nine terms in total with five of them being 0. Subsequently, we obtain

$$\begin{aligned}\mathbb{E}(\hat{\tau}_j, \hat{\tau}_{j'}) &= \dots = \frac{1}{2} \cdot \frac{1}{2} \cdot P(\hat{\tau}_j = \frac{1}{2} | \hat{\tau}_{j'} = \frac{1}{2}) P(\hat{\tau}_{j'} = \frac{1}{2}) \\ &\quad + \frac{1}{2} \cdot 1 \cdot P(\hat{\tau}_j = \frac{1}{2} | \hat{\tau}_{j'} = 1) P(\hat{\tau}_{j'} = 1) \\ &\quad + 1 \cdot \frac{1}{2} \cdot P(\hat{\tau}_j = 1 | \hat{\tau}_{j'} = \frac{1}{2}) P(\hat{\tau}_{j'} = \frac{1}{2}) \\ &\quad + 1 \cdot 1 \cdot P(\hat{\tau}_j = 1 | \hat{\tau}_{j'} = 1) P(\hat{\tau}_{j'} = 1) \\ &= \frac{1}{4} \cdot (\Psi_{jj'}^2 + (1 - \Psi_{jj'})^2) \cdot \frac{1}{2} + \frac{1}{2} \cdot (2\Psi_{jj'}(1 - \Psi_{jj'})) \cdot \frac{1}{4} \\ &\quad + \frac{1}{2} \cdot (\Psi_{jj'}(1 - \Psi_{jj'})) \cdot \frac{1}{2} + 1 \cdot \Psi_{jj'}^2 \cdot \frac{1}{4} \\ &= \frac{1}{8}(1 + 2\Psi_{jj'}) .\end{aligned}$$

Thus, $\text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'}) = \frac{1}{8}(1 + 2\Psi_{jj'} - 2) = \frac{1}{8}(2\Psi_{jj'} - 1)$. With $\Psi_{jj'} = \theta_{jj'}^2 + (1 - \theta_{jj'})^2$, we finally obtain $\text{Cov}(\hat{\tau}_j, \hat{\tau}_{j'}) = \frac{1}{8}(2\theta_{jj'}^2 + 2 - 4\theta_{jj'} + 2\theta_{jj'}^2 - 1) = \frac{1}{8}(1 - 2\theta_{jj'})^2$.

References

1. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*, 66:279–292.
2. Abecasis, G. R., Cherny, S. S., Cookson, W. O. et al. (2001). MERLIN—a rapid analysis of dense genetic maps using sparse genetic flow trees. *Nat Genet*, 30:97–101.
3. Abel, L. & Müller-Myhsok, B. (1998). Robustness and power of the maximum-likelihood-binomial and maximum-likelihood-score methods, in multipoint linkage analysis of affected-sibship data. *Am J Hum Genet*, 63:638–647.
4. Adams, D. (2003). *The Hitchhiker's Guide to the Galaxy: The Original Radio Scripts*. Pan Books: London.
5. Affymetrix (2007). *SNP Concordance Across Affymetrix Genotyping Arrays*. Affymetrix: Santa Clara (CA).
6. Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics*, 55:597–602.
7. Akey, J., Jin, L. & Xiong, M. (2001). Haplotypes vs. single marker linkage disequilibrium tests: What do we gain? *Eur J Hum Genet*, 9:291–300.
8. Akey, J. M., Zhang, K., Xiong, M. et al. (2003). The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol Biol Evol*, 20:232–242.
9. Alcaïs, A. & Abel, L. (1999). Maximum-likelihood-binomial method for genetic model-free linkage analysis of quantitative traits in sibships. *Genet Epidemiol*, 17:102–117.
10. Alcaïs, A. & Abel, L. (2000). Linkage analysis of quantitative trait loci: Sib pairs or sibships? *Hum Hered*, 50:251–256.
11. Allen, A. S., Rathouz, P. J. & Satten, G. A. (2003). Informative missingness in genetic association studies: Case-parent designs. *Am J Hum Genet*, 72:671–680.
12. Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet*, 60:676–690.

13. Allison, D. B. & Faith, M. S. (1997). Issues in mapping genes for eating disorders. *Psychopharmacol Bull*, 33:359–368.
14. Allison, D. B., Faith, M. S. & Nathan, J. S. (1996). Risch's lambda values for human obesity. *Int J Obes Relat Metab Disord*, 20:990–999.
15. Allison, D. B., Fernández, J. R., Heo, M. et al. (2001). Testing the robustness of the new Haseman–Elston quantitative-trait loci-mapping procedure. *Am J Hum Genet*, 67:249–252.
16. Allison, D. B., Neale, M. C., Kezis, M. I. et al. (1996). Assortative mating for relative weight: Genetic implications. *Behav Genet*, 26:103–111.
17. Allison, D. B., Neale, M. C., Zannolli, R. et al. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am J Hum Genet*, 65:531–544.
18. Almasy, L. & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, 62:1198–211.
19. Altman, D. G. & Royston, P. (2000). What do we mean by validating a prognostic model? *Stat Med*, 19:453–473.
20. Altmüller, J., Palmer, L. J., Fischer, G. et al. (2001). Genomewide scans of complex human diseases: True linkage is hard to find. *Am J Hum Genet*, 69:936–950.
21. Altshuler, D. & Daly, M. (2007). Guilt beyond a reasonable doubt. *Nat Genet*, 39:813–815.
22. Altshuler, D., Hirschhorn, J., Klannemark, M. et al. (2000). The common *PPARG* gamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet*, 26:76–80.
23. Amos, C., Elston, R., Wilson, A. et al. (1989). A more powerful robust sib-pair test of linkage for quantitative traits. *Genet Epidemiol*, 6:435–449.
24. Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet*, 54:535–543.
25. Amos, C. I., Dawson, D. V. & Elston, R. C. (1990). The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. *Am J Hum Genet*, 47:842–853.
26. Amos, C. I. & Elston, R. C. (1989). Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet Epidemiol*, 6:349–360.
27. Amos, C. I., Elston, R. C., Bonney, G. E. et al. (1990). A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. *Am J Hum Genet*, 47:247–254.
28. Amos, C. I. & Williamson, J. A. (1993). Robustness of the maximum-likelihood (LOD) method for detecting linkage. *Am J Hum Genet*, 52:213–214.
29. Antonarakis, S. E. (1998). Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum Mutat*, 11:1–3.
30. Ardlie, K. G., Kruglyak, L. & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*, 3:299–309.
31. Ardlie, K. G., Lunetta, K. L. & Seielstad, M. (2002). Testing for population subdivision and association in four case-control studies. *Am J Hum Genet*, 71:304–311.
32. Arminger, G. (1995). Specification and estimation of mean structures. In: *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C. C. Clogg & M. E. Sobel, ed., pp. 77–183. Plenum Press: New York.
33. Arminger, G. & Schoenberg, R. J. (1989). Pseudo maximum-likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika*, 54:409–425.

34. Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386.
35. Attia, J., Ioannidis, J. P., Thakkinstian, A. et al. (2009). How to use an article about genetic association: B: Are the results of the study valid? *J Am Med Assoc*, 301:191–197.
36. Attia, J., Ioannidis, J. P., Thakkinstian, A. et al. (2009). How to use an article about genetic association: C: What are the results and will they help me in caring for my patients? *J Am Med Assoc*, 301:304–308.
37. Ayres, K. L. & Balding, D. J. (1998). Measuring departures from Hardy–Weinberg: A Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity*, 80:769–777.
38. Bacanu, S. A., Devlin, B. & Roeder, K. (2000). The power of genomic control. *Am J Hum Genet*, 66:1933–1944.
39. Bader, J. S. (2001). The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics*, 2:11–24.
40. Badzioch, M. D., Thomas, D. C. & Jarvik, G. P. (2003). Summary report: Missing data and pedigree and genotyping errors. *Genet Epidemiol*, 25 Suppl 1:S36–S42.
41. Bagos, P. G. (2008). A unification of multivariate methods for meta-analysis of genetic association studies. *Stat Appl Genet Mol Biol*, 7:31.
42. Bagos, P. G. & Nikolopoulos, G. K. (2007). A method for meta-analysis of case-control genetic association studies using logistic regression. *Stat Appl Genet Mol Biol*, 6:17.
43. Bailey-Wilson, J. (2003). Parametric and nonparametric linkage analysis. In: *Nature Encyclopedia of the Human Genome*, Vol. 4, pp. 490–492. Macmillan Publishers Ltd.: London.
44. Barber, M. J., Cordell, H. J., MacGregor, A. J. et al. (2004). Gamma regression improves Haseman–Elston and variance components linkage analysis for sib-pairs. *Genet Epidemiol*, 26:97–107.
45. Barrett, J. C. & Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nat Genet*, 38:659–662.
46. Bartsch, D. K., Kress, R., Sina-Frey, M. et al. (2004). Prevalence of familial pancreatic cancer in germany. *Int J Cancer*, 110:902–906.
47. Basilevsky, A. (1983). *Applied Matrix Algebra in the Statistical Sciences*. North-Holland: New York.
48. Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.
49. Becker, T. & Knapp, M. (2004). A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet*, 75:561–570.
50. Beckmann, L., Ziegler, A., Duggal, P. et al. (2005). Haplotypes and haplotype-tagging single-nucleotide polymorphism: Presentation Group 8 of Genetic Analysis Workshop 14. *Genet Epidemiol*, 29 Suppl 1:S59–S71.
51. Bell, P. A., Chaturvedi, S., Gelfand, C. A. et al. (2002). SNPstream UHT: Ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *BioTechniques*, Suppl:70–72, 74, 76–77.
52. Benichou, J. & Palta, M. (2005). Rates, risks, measures of association and impact. In: *Handbook of Epidemiology*, W. Ahrens & I. Pigeot, ed., pp. 91–156. Springer: Heidelberg.
53. Bennett, D. C. & Lamoreux, M. L. (2003). The color loci of mice – a genetic century. *Pigment Cell Res*, 16:333–344.
54. Bennett, P. (2000). Demystified microsatellites. *Mol Pathol*, 53:177–183.

55. Bennett, R. L., Steinhaus, K. A., Uhrich, S. B. et al. (1995). Recommendations for standardized human pedigree nomenclature. *Am J Hum Genet*, 56:745–752.
56. Berghöfer, B., Frommer, T., König, I. R. et al. (2005). Common human toll-like receptor 9 (TLR9) polymorphisms and haplotypes: Association with atopy and functional relevance. *Clin Exp Allergy*, 35:1147–1154.
57. Bezdek, J. C. & Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE T Syst Man Cy B*, 28:301–315.
58. Bickeböller, H. & Clerget-Darpoux, F. (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol*, 12:865–870.
59. Biernacka, J. M., Sun, L. & Bull, S. B. (2004). Simultaneous localization of two linked disease susceptibility genes. *Genet Epidemiol*, 28:33–47.
60. Bird, T. D., Sumi, S. M., Nemens, E. J. et al. (1989). Phenotypic heterogeneity in familial Alzheimer's disease: A study of 24 kindreds. *Ann Neurol*, 25:12–25.
61. Bishop, D. T. & Williamson, J. A. (1990). The power of identity-by-state methods for linkage analysis. *Am J Hum Genet*, 46:254–265.
62. Blackwelder, W. (1977). *Statistical Methods for the Detection of Genetic Linkage from Sibship Data*. PhD thesis, Institute of Statistics Mimeo Series No. 114. University of North Carolina, Chapel Hill, NC.
63. Blackwelder, W. C. & Elston, R. C. (1982). Power and robustness of sib-pair linkage tests and extension to larger sibships. *Commun Stat Theor Meth*, 11:449–484.
64. Blackwelder, W. C. & Elston, R. C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol*, 2:85–97.
65. Blackwell, P. G. (2002). Ornstein-Uhlenbeck process. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 585–588. John Wiley & Sons: New York.
66. Bland, J. M. & Altman, D. G. (1995). Comparing methods of measurement: Why plotting difference against standard method is misleading. *Lancet*, 346:1085–1087.
67. Blangero, J., Williams, J. T. & Almasy, L. (2000). Quantitative trait locus mapping using human pedigrees. *Hum Biol*, 72:35–62.
68. Blettner, M., Sauerbrei, W., Schlehofer, B. et al. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol*, 28:1–9.
69. Boehnke, M. & Langefeld, C. D. (1998). Genetic association mapping based on discordant sib pairs: The discordant-alleles test. *Am J Hum Genet*, 62:950–961.
70. Boes, T. & Neuhäuser, M. (2005). Normalization for Affymetrix GeneChips. *Methods Inf Med*, 44:414–417.
71. Bolstad, B. M., Irizarry, R. A., Astrand, M. et al. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193.
72. Bonin, A., Bellemain, P., Eidesen, B. et al. (2004). How to track and assess genotyping errors in population genetics studies. *Mol Ecol*, 13:3261–3273.
73. Bonney, G. E. (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics*, 42:611–625.
74. Bonney, G. E., Lathrop, G. M. & Lalouel, J. M. (1988). Combined linkage and segregation analysis using regressive models. *Am J Hum Genet*, 43:29–37.
75. Boomsma, D., Busjahn, A. & Peltonen, L. (2002). Classical twin studies and beyond. *Nat Rev Genet*, 3:872–882.

76. Borecki, I. B., Bonney, G., Rice, T. et al. (1993). Influence of genotype-dependent effects of covariates on the outcome of segregation analysis of the body mass index. *Am J Hum Genet*, 53:676–687.
77. Botstein, D., White, R. L., Skolnick, M. et al. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32:314–331.
78. Botto, L. D. & Yang, Q. (2001). *5,10-Methylenetetrahydrofolate reductase* gene variants and congenital anomalies: A HuGE review. *Am J Epidemiol*, 151:862–877.
79. Bouchard, C. & Perusse, L. (1993). Genetic aspects of obesity. *Ann N Y Acad Sci*, 699:26–35.
80. Bouchard, T. J., J., Lykken, D. T., McGue, M. et al. (1990). Sources of human psychological differences: The Minnesota Study of Twins Reared Apart. *Science*, 250:223–228.
81. Bourgain, C., Abney, M., Schneider, D. et al. (2004). Testing for Hardy–Weinberg equilibrium in samples with related individuals. *Genetics*, 168:2349–2361.
82. Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *J R Statist Soc B*, 26:211–252.
83. Boyles, A. L., Scott, W. K., Martin, E. R. et al. (2005). Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Hum Hered*, 59:220–227.
84. Breiman, L. (2001). Random forests. *Mach Learn*, 45:5–32.
85. Breslow, N. & Day, N. (1980). *Statistical Methods in Cancer Research. Vol. 1, The Analysis of Case-Control Studies*, Vol. IARC Scientific Publications No. 32. IARC: Lyon.
86. Breslow, N. & Day, N. (1987). *Statistical Methods in Cancer Research. Vol. 2, The Design and Analysis of Cohort Studies*, Vol. IARC Scientific Publications No. 82. IARC: Lyon.
87. Bretz, F., Landgrebe, J. & Brunner, E. (2005). Multiplicity issues in microarray experiments. *Methods Inf Med*, 44:431–437.
88. Broman, K. W. (2005). The genomes of recombinant inbred lines. *Genetics*, 169:1133–1146.
89. Brookes, A. J. (1999). The essence of SNPs. *Gene*, 234:177–186.
90. Brookfield, J. F. Y. (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol Ecol*, 5:453–455.
91. Brown, T. P., Rumsby, P. C., Capleton, A. C. et al. (2006). Pesticides and Parkinson's disease—is there a link? *Environ Health Perspect*, 114:156–164.
92. Browne, M. W. & Arminger, G. (1995). Specification and estimation of mean- and covariance-structure models. In: *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C. C. Clogg & M. E. Sobel, ed., pp. 185–249. Plenum Press: New York.
93. Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet*, 124:439–450.
94. Browning, S. R. & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81:1084–1097.
95. Brzustowicz, L. M., Mérette, C., Xie, X. et al. (1993). Molecular and statistical approaches to the detection and correction of errors in genotype databases. *Am J Hum Genet*, 53:1137–1145.
96. Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am J Hum Genet*, 49:985–994.
97. Butler, H. & Ragoussis, J. (2008). BeadArray-based genotyping. *Methods Mol Biol*, 439:53–74.

98. Butler, K., Field, C., Herbinger, C. M. et al. (2004). Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Mol Ecol*, 13:1589–1600.
99. Caesar, C. I. T. (1986). *De Bello Gallico I*. Bristol Classical Press: Bristol.
100. Calafell, F., Shuster, A., Speed, W. C. et al. (1998). Short tandem repeat polymorphism evolution in humans. *Eur J Hum Genet*, 6:38–49.
101. Calle, M. L., Urrea, V., Vellalta, G. et al. (2008). Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Stat Med*, 27:6532–6546.
102. Camp, N. J. & Bansal, A. (2003). Complex multifactorial genetic diseases. In: *Nature Encyclopedia of the Human Genome*, Vol. 1, pp. 930–935. Macmillan Publishers Ltd.: London.
103. Cardon, L. R. & Fulker, D. W. (1994). The power of interval mapping of quantitative trait loci, using selected sib pairs. *Am J Hum Genet*, 55:825–833.
104. Carey, G. & Williamson, J. (1991). Linkage analysis of quantitative traits: Increased power by using selected samples. *Am J Hum Genet*, 49:786–796.
105. Cargill, M., Altshuler, D., Ireland, J. et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*, 22:231–238.
106. Carlson, C. S., Eberle, M. A., Rieder, M. J. et al. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74:106–120.
107. Carter, B. S., Beaty, T. H., Steinberg, G. D. et al. (1992). Mendelian inheritance of familial prostate cancer. *PNAS USA*, 89:3367–3371.
108. Cavalli-Sforza, L. L. & Bodmer, W. F. (1971). *The Genetics of Populations*. Freeman: San Francisco.
109. Cerda-Flores, R. M., Barton, S. A., Marty-Gonzalez, L. F. et al. (1999). Estimation of nonpaternity in the Mexican population of Nuevo Leon: A validation study with blood group markers. *Am J Phys Anthropol*, 109:281–293.
110. Cervino, A. C. & Hill, A. V. (2000). Comparison of tests for association and linkage in incomplete families. *Am J Hum Genet*, 67:120–132.
111. Chakraborty, R. (2002). Genetic distance. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 334–336. John Wiley & Sons: New York.
112. Chakraborty, R., de Andrade, M., Daiger, S. P. et al. (1992). Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann Hum Genet*, 56:45–57.
113. Chakravarti, A. & Buetow, K. H. (1985). A strategy for using multiple linked markers for genetic counseling. *Am J Hum Genet*, 37:984–997.
114. Chakravarti, A. (2002). Linkage disequilibrium. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 472–475. John Wiley & Sons: New York.
115. Chanock, S. J., Manolio, T., Boehnke, M. et al. (2007). Replicating genotype-phenotype associations. *Nature*, 447:655–660.
116. Chapman, J. M., Cooper, J. D., Todd, J. A. et al. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Hum Hered*, 56:18–31.
117. Chen, W. J. (1992). Estimating age at onset distributions: A review of methods and issues. *Psychiatr Genet*, 2:219–238.

118. Chen, W. M., Broman, K. W. & Liang, K.-Y. (2004). Quantitative trait linkage analysis by generalized estimating equations: Unification of variance components and Haseman–Elston regression. *Genet Epidemiol*, 26:265–272.
119. Chen, Y.-H. (2004). New approach to association testing in case-parent designs under informative parental missingness. *Genet Epidemiol*, 27:131–140.
120. Chiang, C. L. (1998). Stochastic processes. In: *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, ed., Vol. 6, pp. 4318–4345. John Wiley & Sons: New York.
121. Chow, J. C., Yen, Z., Ziesche, S. M. et al. (2005). Silencing of the mammalian X chromosome. *Annu Rev Genomics Hum Genet*, 6:69–92.
122. Cienko, C. (1998). Asymptotic distribution of the maximum likelihood ratio test for gene detection. *Statistics*, 31:261–285.
123. Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol*, 7:111–122.
124. Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genet Epidemiol*, 27:321–333.
125. Claus, E. B., Risch, N. & Thompson, W. D. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am J Hum Genet*, 48:232–242.
126. Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet*, 65:1170–1177.
127. Clayton, D. (2008). Testing for association on the X chromosome. *Biostatistics*, 9:593–600.
128. Clayton, D., Chapman, J. & Cooper, J. (2004). Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol*, 27:415–428.
129. Clayton, D. & Jones, H. (1999). Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet*, 65:1161–1169.
130. Clayton, D. G. (2009). Prediction and interaction in complex disease genetics: Experience in type 1 diabetes. *PLoS Genet*, 5:e1000540.
131. Clayton, D. G., Walker, N. M., Smyth, D. J. et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, 37:1243–1246.
132. Clerget-Darpoux, F. & Elston, R. C. (2007). Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum Hered*, 64:91–96.
133. Cleves, M. A., Olson, J. M. & Jacobs, K. B. (1997). Exact transmission-disequilibrium tests with multiallelic markers. *Genet Epidemiol*, 14:337–347.
134. Cloninger, C. R. (1987). Neurogenetic adaptive mechanisms in alcoholism. *Science*, 236:410–416.
135. Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10:417–451.
136. Collet, J., Hulot, J. S., Pena, A. et al. (2009). Cytochrome P450 2C19 polymorphism in young patients treated with clopidogrel after myocardial infarction: A cohort study. *Lancet*, 373:309–317.
137. Collins, A., Ennis, S., Taillon-Miller, P. et al. (2001). Allelic association with SNPs: Metrics, populations, and the linkage disequilibrium map. *Hum Mutat*, 17:255–262.
138. Collins, A., Lau, W. & Vega, F. M. D. L. (2004). Mapping genes for common diseases: The case for genetic (LD) maps. *Hum Hered*, 58:2–9.
139. Collins, A. & Morton, N. E. (1995). Nonparametric tests for linkage with dependent sib pairs. *Hum Hered*, 45:311–318.

140. Commenges, D. (1994). Robust genetic linkage analysis based on a score test of homogeneity: The weighted pairwise correlation statistic. *Genet Epidemiol*, 11:189–200.
141. Conti, D. V. & Gauderman, W. J. (2004). SNPs, haplotypes, and model selection in a candidate gene region: The SIMPle analysis for multilocus data. *Genet Epidemiol*, 27:429–441.
142. Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115:928–935.
143. Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, 11:2463–2468.
144. Cordell, H. J. (2004). Bias toward the null hypothesis in model-free linkage analysis is highly dependent on the test statistic used. *Am J Hum Genet*, 74:1294–1302.
145. Cordell, H. J. (2004). Properties of case/pseudocontrol analysis for genetic association studies: Effects of recombination, ascertainment, and multiple affected offspring. *Genet Epidemiol*, 26:186–205.
146. Cordell, H. J. (2009). Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10:392–404.
147. Cordell, H. J. & Clayton, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes. *Am J Hum Genet*, 70:124–141.
148. Cordell, H. J., Todd, J. A., Bennett, S. T. et al. (1995). Two-locus maximum LOD score analysis of a multifactorial trait: Joint consideration of *IDDM2* and *IDDM4* with *IDDM1* in type I diabetes. *Am J Hum Genet*, 57:920–934.
149. Cupples, L. A., Risch, N., Farrer, L. A. et al. (1991). Estimation of morbid risk and age at onset with missing information. *Am J Hum Genet*, 49:76–87.
150. Curtis, D. (1997). Use of siblings as controls in case-control association studies. *Ann Hum Genet*, 61:319–333.
151. Curtis, D. & Sham, P. C. (1995). A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet*, 56:811–812.
152. Daly, M. J., Rioux, J. D., Schaffner, S. F. et al. (2001). High-resolution haplotype structure in the human genome. *Nat Genet*, 29:229–232.
153. Darvasi, A. & Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait. *Theor Appl Genet*, 85:353–359.
154. Davis, S. & Weeks, D. E. (1997). Comparison of nonparametric statistics for detection of linkage in nuclear families: Single-marker evaluation. *Am J Hum Genet*, 61:1431–1444.
155. Daw, E. W., Thompson, E. A. & Wijsman, E. M. (2000). Bias in multipoint linkage analysis arising from map misspecification. *Genet Epidemiol*, 19:366–380.
156. Dawson, E., Abecasis, G. R., Bumpstead, S. et al. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418:544–548.
157. de Bakker, P. I., Ferreira, M. A., Jia, X. et al. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*, 17:R122–128.
158. Deffenbacher, K., Kenyon, J., Hoover, D. et al. (2004). Refinement of the 6p21.3 quantitative trait locus influencing dyslexia: Linkage and association analyses. *Hum Genet*, 115:128–138.
159. den Dunnen, J. T. & Antonarakis, S. E. (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum Mutat*, 15:7–12.
160. DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Control Clin Trials*, 7:177–188.

161. Devlin, B. & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29:311–322.
162. Devlin, B. & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55:997–1004.
163. Devlin, B., Roeder, K. & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*, 60:155–166.
164. Dixon, S. J., Heinrich, N., Holmboe, M. et al. (2009). Use of cluster separation indices and the influence of outliers: application of two new separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles. *J Chemometr*, 23:19–31.
165. Does, A., Johnson, N. A. & Thiel, T. (2004). *Rediscovering Biology*. Annenberg Media: <http://www.learner.org/channel/courses/biology/>.
166. Douglas, J. A., Boehnke, M. & Lange, K. (2000). A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am J Hum Genet*, 66:1287–1297.
167. Douglas, J. A., Skol, A. D. & Boehnke, M. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am J Hum Genet*, 70:487–495.
168. Dracopoli, N. C., O'Connell, P., Elsner, T. I. et al. (1991). The CEPH consortium linkage map of human chromosome 1. *Genomics*, 9:686–700.
169. Draper, N. R. & Smith, H. (1998). *Applied Regression Analysis*, (3 ed.). John Wiley & Sons: New York.
170. Drigalenko, E. (1998). How sib pairs reveal linkage. *Am J Hum Genet*, 63:1242–1245.
171. Drożdżik, M., Białecka, M., Myśliwiec, K. et al. (2003). Polymorphism in the P-glycoprotein drug transporter *MDR1* gene: a possible link between environmental and genetic factors in Parkinson's disease. *Pharmacogenetics*, 13:259–263.
172. Dubay, C., Vincent, M., Samani, N. J. et al. (1993). Genetic determinants of diastolic and pulse pressure map to different loci in Lyon hypertensive rats. *Nat Genet*, 3:354–357.
173. Dudbridge, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered*, 66:87–98.
174. Dudbridge, F. & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32:227–234.
175. Duggal, P., Gillanders, E. M., Holmes, T. N. et al. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, 9:516.
176. Eaves, L. J., Fulker, D. W. & Heath, A. C. (1989). The effects of social homogamy and cultural inheritance on the covariances of twins and their parents: A LISREL model. *Behav Genet*, 19:113–122.
177. Edwards, J. H. (1971). The analysis of X-linkage. *Ann Hum Genet*, 34:229–250.
178. Edwards, J. H. (1998). Penrose and sib-pairs. *Ann Hum Genet*, 62:365–377.
179. Elston, R. C. (1998). Linkage and association. *Genet Epidemiol*, 15:565–576.
180. Elston, R. C. (1998). Statistical Genetics '98: Methods of linkage analysis – and the assumptions underlying them. *Am J Hum Genet*, 63:931–934.
181. Elston, R. C., Buxbaum, S., Jacobs, K. B. et al. (2000). Haseman and Elston revisited. *Genet Epidemiol*, 19:1–17.
182. Elston, R. C. & Forthofer, R. (1977). Testing for Hardy–Weinberg equilibrium in small samples. *Biometrics*, 33:536–542.

183. Elston, R. C. & Keats, B. J. B. (1985). Genetic Analysis Workshop III: Sib pair analyses to determine linkage groups and to order loci. *Genet Epidemiol*, 2:211–213.
184. Elston, R. C., Kringlen, E. & Namboodiri, K. K. (1973). Possible linkage relationships between certain blood groups and schizophrenia or other psychoses. *Behav Genet*, 3:101–106.
185. Elston, R. C., Lin, D. & Zheng, G. (2007). Multistage sampling for genetic studies. *Annu Rev Genomics Hum Genet*, 8:327–342.
186. Elston, R. C., Song, D. & Iyengar, S. K. (2005). Mathematical assumptions versus biological reality: Myths in affected sib pair linkage analysis. *Am J Hum Genet*, 76:152–156.
187. Elston, R. C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Hum Hered*, 21:523–542.
188. Epstein, M. P., Allen, A. S. & Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet*, 80:921–930.
189. Epstein, M. P. & Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet*, 73:1316–1329.
190. Epstein, M. P., Veal, C. D., Trembath, R. C. et al. (2005). Genetic association analysis using data from triads and unrelated subjects. *Am J Hum Genet*, 76:592–608.
191. Erdmann, J., Großhennig, A., Braund, P. S. et al. (2009). Novel susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet*, 41:280–282.
192. Etzel, C. J., Shete, S., Beasley, T. M. et al. (2003). Effect of Box–Cox transformation on power of Haseman–Elston and maximum-likelihood variance components tests to detect quantitative trait loci. *Hum Hered*, 55:108–116.
193. Evans, J. P., Skrzynia, C. & Burke, W. (2001). The complexities of predictive genetic testing. *Brit Med J*, 322:1052–1056.
194. Evans, S. N., McPeck, M. S. & Speed, T. P. (1993). A characterisation of crossover models that possess map functions. *Theor Popul Biol*, 43:80–90.
195. Ewens, W. J. & Grant, G. R. (2001). *Statistical Methods in Bioinformatics: An Introduction*. Springer: New York.
196. Ewens, W. J., Li, M. & Spielman, R. S. (2008). A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. *PLoS Genet*, 4:e1000180.
197. Excoffier, L. & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12:921–927.
198. Falconer, D. S. & Mackay, T. F. C. (1995). *Introduction to Quantitative Genetics*, (4 ed.). Addison Wesley Longman: New York.
199. Falk, C. T. & Rubinstein, P. (1987). Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet*, 51:227–233.
200. Fallin, D. & Schork, N. J. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet*, 67:947–959.
201. Faraway, J. J. (1993). Improved sib-pair linkage test for disease susceptibility loci. *Genet Epidemiol*, 10:225–233.
202. Farbrother, J. E., Kirov, G., Owen, M. J. et al. (2004). Family aggregation of high myopia: Estimation of the sibling recurrence risk ratio. *Invest Ophthalmol Vis Sci*, 45:2873–2878.
203. Fardo, D., Celedon, J. C., Raby, B. A. et al. (2007). On dichotomizing phenotypes in family-based association tests: Quantitative phenotypes are not always the optimal choice. *Genet Epidemiol*, 31:376–382.

204. Farrall, M. (1997). Affected sibpair linkage tests for multiple linked susceptibility genes. *Genet Epidemiol*, 14:103–115.
205. Feder, J. N., Gnirke, A., Thomas, W. et al. (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet*, 13:399–408.
206. Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *J Appl Prob*, 30:766–779.
207. Feingold, E. (2001). Methods for linkage analysis of quantitative trait loci in humans. *Theor Popul Biol*, 60:167–180.
208. Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st century. *Am J Hum Genet*, 71:217–222.
209. Feingold, E., Brown, P. O. & Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am J Hum Genet*, 53:234–251.
210. Felsenstein, J. (1979). A mathematical tractable family of genetic mapping functions with different amounts of interference. *Genetics*, 91:769–775.
211. Fernando, P., Evans, B. J., Morales, J. C. et al. (2001). Electrophoresis artefacts – a previously unrecognized cause of error in microsatellite analysis. *Mol Ecol Notes*, 1:325–328.
212. Fernando, R. L., Stricker, C. & Elston, R. C. (1993). An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theor Appl Genet*, 87:89–93.
213. Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb*, 52:399–433.
214. Fisher, S. A., Lewis, C. M. & Wise, L. H. (2001). Detecting population outliers and null alleles in linkage data: Application to GAW12 asthma studies. *Genet Epidemiol*, 21 Suppl 1:S18–S23.
215. Fisher, S. E. & DeFries, J. C. (2002). Developmental dyslexia: Genetic dissection of a complex cognitive trait. *Nat Rev Neurosci*, 3:767–780.
216. Fitze, G., Cramer, J., Ziegler, A. et al. (2002). Association between c135G/A genotype and *RET* proto-oncogene germline mutations and phenotype of Hirschsprung's disease. *Lancet*, 359:1200–1205.
217. Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons: New York.
218. Forrest, W. (2001). Weighting improves the 'New Haseman–Elston' method. *Hum Hered*, 52:47–54.
219. Forton, J., Kwiatkowski, D., Rockett, K. et al. (2005). Accuracy of haplotype reconstruction from haplotype-tagging single-nucleotide polymorphisms. *Am J Hum Genet*, 76:438–448.
220. Franjkovic, I., Gessner, A., König, I. R. et al. (2005). Effects of common atopy-associated amino acid substitutions in the IL-4 receptor alpha chain on IL-4 induced phenotypes. *Immunogenetics*, 56:808–817.
221. Franke, D., Philippi, A., Tores, F. et al. (2005). On confidence intervals for genotype relative risks and attributable risks from case parent trio design for candidate-gene studies. *Hum Hered*, 60:81–88.
222. Franke, D. & Ziegler, A. (2005). Weighting affected sib pairs by marker informativity. *Am J Hum Genet*, 77:230–241.
223. Frankel, W. N. & Schork, N. J. (1996). Who's afraid of epistasis? *Nat Genet*, 14:371–373.

224. Freedman, M. L., Reich, D., Penney, K. L. et al. (2004). Assessing the impact of population stratification on genetic association studies. *Nat Genet*, 36:388–393.
225. Freidlin, B., Podgor, M. J. & Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics*, 55:883–886.
226. Freidlin, B., Zheng, G., Li, Z. et al. (2002). Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Hum Hered*, 53:146–152.
227. Freimer, N. B., Sandkuijl, L. A. & Blower, S. M. (1993). Incorrect specification of marker allele frequencies: Effects on linkage analysis. *Am J Hum Genet*, 52:1102–1110.
228. Fu, W., Wang, Y., Wang, Y. et al. (2009). Missing call bias in high-throughput genotyping. *BMC Genomics*, 10:106.
229. Fulker, D. W. & Cardon, L. R. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *Am J Hum Genet*, 54:1092–1103.
230. Fulker, D. W. & Cherny, S. S. (1996). An improved multipoint sib-pair analysis of quantitative traits. *Behav Genet*, 26:527–532.
231. Fulker, D. W., Cherny, S. S. & Cardon, L. R. (1995). Multipoint interval mapping of quantitative trait loci, using sibpairs. *Am J Hum Genet*, 56:1224–1233.
232. Fulker, D. W., Cherny, S. S., Sham, P. C. et al. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet*, 64:259–267.
233. Gabriel, S. B., Schaffner, S. F., Nguyen, H. et al. (2002). The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229.
234. Gao, X., Becker, L. C., Becker, D. M. et al. (2009). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet Epidemiol*, 34:100–105.
235. Gao, X., Starmer, J. & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*, 32:361–369.
236. Gauderman, W. J. (2003). Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genet Epidemiol*, 25:327–338.
237. Gauderman, W. J. & Morrison, J. L. (2000). Evidence for age-specific genetic relative risks in lung cancer. *Am J Epidemiol*, 151:41–49.
238. Gaudet, M., Fara, A. G., Beritognolo, I. et al. (2009). Allele-specific PCR in SNP genotyping. *Methods Mol Biol*, 578:415–424.
239. Ge, Y., Dudoit, S. & Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12:1–77.
240. Geller, F. & Ziegler, A. (2000). Studiendesigns zur Rekrutierung von Geschwisterpaaren für die genetische Kartierung quantitativer Phänotypen. *Med Genet*, 12:423–427.
241. Geller, F. & Ziegler, A. (2002). Detection rates for genotyping errors in SNPs using the trio design. *Hum Hered*, 54:111–117.
242. Génin, E. & Clerget-Darpoux, F. (1996). Consanguinity and the sib-pair method: An approach using identity by descent between and within individuals. *Am J Hum Genet*, 59:1149–1162.
243. George, A. W. & Thompson, E. A. (2003). Discovering disease genes: Multipoint linkage analysis via a new Markov chain Monte Carlo approach. *Stat Sci*, 18:515–531.
244. Gibson, C. J. & Gruen, J. R. (2008). The human lexinome: Genes of language and reading. *J Commun Disord*, 41:409–420.
245. Glaser, B. & Holmans, P. (2009). Comparison of methods for combining case-control and family-based association studies. *Hum Hered*, 68:106–116.
246. Glidden, D. V., Liang, K.-Y., Chiu, Y.-F. et al. (2003). Multipoint affected sibpair linkage methods for localizing susceptibility genes of complex diseases. *Genet Epidemiol*, 24:107–117.

247. Go, R. C., Elston, R. C. & Kaplan, E. B. (1978). Efficiency and robustness of pedigree segregation analysis. *Am J Hum Genet*, 30:28–37.
248. Gochhait, S., Bukhari, S. I., Bairwa, N. et al. (2007). Implication of *BRCA2*-26G>A 5' untranslated region polymorphism in susceptibility to sporadic breast cancer and its modulation by *p53* codon 72 Arg>Pro polymorphism. *Breast Cancer Res*, 9:R71.
249. Goddard, K. A. & Wijsman, E. M. (2002). Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. *Genet Epidemiol*, 22:205–220.
250. Goddard, K. A., Ziegler, A. & Wellek, S. (2009). Adapting the logical basis of tests for Hardy–Weinberg equilibrium to the real needs of association studies in Human and Medical Genetics. *Genet Epidemiol*, 33:569–580.
251. Goddard, K. A. B., Witte, J. S., Suarez, B. K. et al. (2001). Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am J Hum Genet*, 68:1197–1206.
252. Goldin, L. R. & Weeks, D. E. (1993). Two-locus models of disease: Comparison of likelihood and nonparametric linkage methods. *Am J Hum Genet*, 53:908–915.
253. Goldstein, D. R., Dudoit, S. & Speed, T. P. (2001). Power and robustness of a score test for linkage analysis of quantitative traits using identity by descent data on sib pairs. *Genet Epidemiol*, 20:415–431.
254. Gonick, L. & Wheelis, M. (1991). *The Cartoon Guide to Genetics*, (revised ed.). HarperCollins: New York.
255. Gonzalez, J. R., Carrasco, J. L., Dudbridge, F. et al. (2008). Maximizing association statistics over genetic models. *Genet Epidemiol*, 32:246–254.
256. Gordon, D. & Devoto, M. (2008). Advances in family-based association analysis. Introduction. *Hum Hered*, 66:65–66.
257. Göring, H. H. H. & Terwilliger, J. D. (2000). Linkage analysis in the presence of errors I: Complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet*, 66:1095–1106; Erratum in: *Am J Hum Genet* 2000; 66:1472.
258. Gorroochurn, P., Heiman, G. A., Hodge, S. E. et al. (2006). Centralizing the non-central chi-square: A new method to correct for population stratification in genetic case-control association studies. *Genet Epidemiol*, 30:277–289.
259. Greenberg, D. A. & Abreu, P. C. (2001). Determining trait locus position from multipoint analysis: Accuracy and power of three different test statistics. *Genet Epidemiol*, 21:299–314.
260. Greenwood, C. M. T. & Bull, S. B. (1997). Incorporation of covariates into genome scanning using sib pair analysis in bipolar affective disorder. *Genet Epidemiol*, 14:635–640.
261. Greenwood, C. M. T. & Bull, S. B. (1999). Analysis of affected sib-pairs, with covariates – with and without constraints. *Am J Hum Genet*, 64:871–885.
262. Gregersen, J. W., Kranc, K. R., Ke, X. et al. (2006). Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*, 443:574–577.
263. Gregorius, H. R. (1980). The probability of losing an allele when diploid genotypes are sampled. *Biometrics*, 36:643–652.
264. Grimes, D. A. & Schulz, K. F. (2002). Cohort studies: Marching towards outcomes. *Lancet*, 359:341–345.
265. Guedj, M., Nuel, G. & Prum, B. (2008). A note on allelic tests in case-control association studies. *Ann Hum Genet*, 72:407–409.
266. Guo, C. Y., Lunetta, K. L., DeStefano, A. L. et al. (2007). Informative-transmission disequilibrium test (i-TDT): Combined linkage and association mapping that includes unaffected offspring as well as affected offspring. *Genet Epidemiol*, 31:115–133.

267. Guo, S. W. (1999). The behaviors of some heritability estimators in the complete absence of genetic factors. *Hum Hered*, 49:215–228.
268. Guo, S. W. & Thompson, E. A. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, 48:361–372.
269. Guo, X. & Elston, R. C. (1999). Linkage information content of polymorphic genetic markers. *Hum Hered*, 49:112–118.
270. Guo, X. & Elston, R. C. (2000). Two-stage global search designs for linkage analysis I: Use of the mean statistic for affected sib pairs. *Genet Epidemiol*, 18:97–110.
271. Guo, X., Olson, J. M., Elston, R. C. et al. (2002). The linkage information content value of polymorphism genetic markers in model-free linkage analysis. *Hum Hered*, 53:45–48.
272. Hädicke, O., Pahlke, F. & Ziegler, A. (2008). A general approach for sample size and power calculations based on the Haseman–Elston method. *Biom J*, 50:257–269.
273. Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, 8:299–309.
274. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002). Cluster validity methods: Part I. *Sigmod Rec*, 31:40–45.
275. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002). Clustering validity checking methods: Part II. *Sigmod Rec*, 31:19–27.
276. Hall, J. M., Lee, M. K., Newman, B. et al. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250:1684–1689.
277. Halldorsson, B. V., Istrail, S. & De La Vega, F. M. (2004). Optimal selection of SNP markers for disease association studies. *Hum Hered*, 58:190–202.
278. Hallgren, B. (1950). Specific dyslexia ('congenital word blindness'): A clinical and genetic study. *Acta Psychiatr Neurol Scand*, 65 (Suppl):1–287.
279. Halperin, E. & Stephan, D. A. (2009). SNP imputation in association studies. *Nat Biotechnol*, 27:349–351.
280. Halpern, J. & Whittemore, A. S. (1999). Multipoint linkage analysis: A cautionary note. *Hum Hered*, 49:194–196.
281. Hamer, D. & Sirota, L. (2000). Beware the chopsticks gene. *Mol Psychiatry*, 5:11–13.
282. Hampe, J., Franke, A., Rosenstiel, P. et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nat Genet*, 39:207–211.
283. Handl, J. & Knowles, J. (2005). Exploiting the trade-off—the benefits of multiple objectives in data clustering. In: *Evolutionary Multi-Criterion Optimization*, C. A. C. Coello, A. H. Aguirre & E. Zitzler, ed., pp. 547–560. Springer: Heidelberg.
284. Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21:3201–3212.
285. Hao, K., Li, C., Rosenow, C. et al. (2004). Detect and adjust for population stratification in population-based association study using genomic control markers: An application of Affymetrix Genechip Human Mapping 10K array. *Eur J Hum Genet*, 12:1001–1006.
286. Hao, K., Li, C., Rosenow, C. et al. (2004). Estimation of genotype error rate using samples with pedigree information—an application on the GeneChip Mapping 10K array. *Genomics*, 84:623–630.
287. Haseman, J. K. & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet*, 2:3–19.
288. Hawley, M. E. & Kidd, K. K. (1995). HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered*, 86:409–411.

289. Hebebrand, J., Sommerlad, C., Geller, F. et al. (2001). The genetics of obesity: Practical implications. *Int J Obes Relat Metab Disord*, 25 Suppl 1:S10–18.
290. Hecker, M., Bohnert, A., König, I. R. et al. (2003). Novel genetic variation of human interleukin-21 receptor is associated with elevated IgE levels in females. *Genes Immun*, 4:228–233.
291. Hedrick, P. W. (2005). *Genetics of Populations*, (3 ed.). Jones and Bartlett: Sudbury (MA).
292. Heidinger, K., König, I. R., Bohnert, A. et al. (2005). Polymorphisms of the human surfactant protein-d (SFTPD) gene: Strong evidence that serum levels of surfactant protein-D (SP-D) are genetically influenced. *Immunogenetics*, 57:1–7.
293. Hetrick, K., Marosy, B., Zilka, M. et al. (2009). Genome-wide association studies (GWAS): Performance comparison between Illumina 1m BeadChip and Affymetrix Genome-Wide SNP Array 6.0. p. Abstract 1764. American Society of Human Genetics, 59th Annual Meeting, <http://www.ashg.org/2009meeting/abstracts/fulltext/fl11014.htm>.
294. Higgins, J. P. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat Med*, 21:1539–1558.
295. Hills, M. (2002). Likelihood. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 435–440. John Wiley & Sons: New York.
296. Hindorff, L. A., Sethupathy, P., Junkins, H. A. et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106:9362–9367.
297. Hirschhorn, J. N., Lohmueller, K., Byrne, E. et al. (2002). A comprehensive review of genetic association studies. *Genet Med*, 4:45–61.
298. Hodge, S. E. (1984). The information contained in multiple sibling pairs. *Genet Epidemiol*, 1:109–122.
299. Hodge, S. E., Flodman, P. L., Duryea, M. F. et al. (1998). Estimating recombination fraction separately for females and males: A counterintuitive result. *Hum Hered*, 48:42–48.
300. Hoggart, C. J., Parra, E. J., Shriver, M. D. et al. (2003). Control of confounding of genetic associations in stratified populations. *Am J Hum Genet*, 72:1492–1504.
301. Hoh, J., Wile, A. & Ott, J. (2001). Trimming, weighting, and grouping snps in human case-control association studies. *Genome Res*, 11:2115–2119.
302. Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet*, 52:362–374.
303. Holmans, P. (1998). Affected sib-pair methods for detecting linkage to dichotomous traits: A review. *Hum Biol*, 70:1025–1040.
304. Holmans, P. (2001). Nonparametric linkage. In: *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop & C. Cannings, ed., pp. 487–504. John Wiley & Sons: New York.
305. Holmans, P. & Clayton, D. (1995). Efficiency of typing unaffected relatives in an affected sib-pair linkage study with single locus and multiple tightly-linked markers. *Am J Hum Genet*, 57:1221–1232.
306. Holtzman, N. A. & Marteau, T. M. (2000). Will genetics revolutionize medicine? *New Engl J Med*, 343:141–144.
307. Honey, K. (2008). GINA: Making it safe to know what's in your genes. *J Clin Invest*, 118:2369.
308. Hopper, J. L. & Visscher, P. M. (2002). Variance component analysis. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 778–788. John Wiley & Sons: New York.

309. Horaitis, O. & Cotton, R. G. (2004). The challenge of documenting mutation across the genome: The human genome variation society approach. *Hum Mutat*, 23:447–452.
310. Horne, B. D. & Camp, N. J. (2004). Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet Epidemiol*, 26:11–21.
311. Horvath, S. & Laird, N. M. (1998). A discordant-sibship test for disequilibrium and linkage: No need for parental data. *Am J Hum Genet*, 63:1886–1897.
312. Horvath, S., Laird, N. M. & M., K. (2000). The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. *Am J Hum Genet*, 66:1161–1167.
313. Horvath, S., Xu, X. & Laird, N. M. (2001). The family based association test method: Strategies for studying general genotype-phenotype associations. *Eur J Hum Genet*, 9:301–306.
314. Hothorn, L. A. (2009). Estimating simultaneous confidence intervals for odds ratios in GWA with biallelic marker and unknown mode of inheritance. West North American Region: Portland, http://www.mth.pdx.edu/events/wnar/abstracts/WNAR_abstracts.asp?id=50.
315. Hothorn, L. A. & Hothorn, T. (2009). Order-restricted scores test for the evaluation of population-based case-control studies when the genetic model is unknown. *Biom J*, 51:659–669.
316. Howie, B. N., Donnelly, P. & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5:e1000529.
317. Hu, X., Schrodi, S. J., Ross, D. A. et al. (2004). Selecting tagging SNPs for association studies using power calculations from genotype data. *Hum Hered*, 57:156–170.
318. Hua, J., Craig, D. W., Brun, M. et al. (2007). SNiPer-HD: Improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics*, 23:57–63.
319. Huang, J. & Vieland, V. (1997). A new statistical test for age-of-onset anticipation: Application to bipolar disorder. *Genet Epidemiol*, 14:1091–1096.
320. Huang, J. & Vieland, V. (2001). Comparison of ‘model-free’ and ‘model-based’ linkage statistics in the presence of locus heterogeneity: Single data set and multiple data set applications. *Hum Hered*, 51:217–225.
321. Huang, J. & Vieland, V. (2001). The null distribution of the heterogeneity LOD score does depend on the assumed genetic model for the trait. *Hum Hered*, 52:217–222.
322. Huang, L., Li, Y., Singleton, A. B. et al. (2009). Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet*, 84:235–250.
323. Huang, Q., Shete, S. & Amos, C. I. (2004). Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet*, 75:1106–1112.
324. Humphries, S. E., Cooper, J. A., Talmud, P. J. et al. (2007). Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men. *Clin Chem*, 53:8–16.
325. Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nat Rev Genet*, 6:287–298.
326. Idury, R. M. & Elston, R. C. (1997). A faster and more general hidden Markov model algorithm for multipoint likelihood calculation. *Hum Hered*, 47:197–202.
327. Igl, B.-W., König, I. R. & Ziegler, A. (2009). What do we mean by “replication” and “validation” in genome-wide association studies? *Hum Hered*, 67:66–68.

328. Iles, M. M. (2005). The effect of SNP marker density on the efficacy of haplotype tagging SNPs – A warning. *Ann Hum Genet*, 69:209–215.
329. Iles, M. M. (2008). What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet*, 4:e33.
330. Infante-Rivard, C., Mirea, L. & Bull, S. B. (2009). Combining case-control and case-trio data from the same population in genetic association analyses: Overview of approaches and illustration with a candidate gene study. *Am J Epidemiol*, 170:657–664.
331. International Conference on Harmonisation E9 Expert Working Group, I. C. H. (1999). Statistical principles for clinical trials (ICH E9). *Stat Med*, 18:1905–1942.
332. International Organization for Standardization, I. S. O. (1994). *ISO 5725-1:1994. Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions*. International Organization for Standardization: Geneva, Switzerland.
333. International Organization for Standardization, I. S. O. (2006). *ISO 3534-1:2006. Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization: Geneva, Switzerland.
334. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. et al. (2001). Replication validity of genetic association studies. *Nat Genet*, 29:306–309.
335. Ittrich, C. (2005). Normalization of two-channel microarray data. *Methods Inf Med*, 44:418–422.
336. Jacobsen, N., Bentzen, J., Meldgaard, M. et al. (2002). LNA-enhanced detection of single nucleotide polymorphisms in the apolipoprotein E. *Nucleic Acids Res*, 30:E100.
337. Jacquard, A. (1983). Heritability: One word, three concepts. *Biometrics*, 39:465–477.
338. James, J. W. (1971). Frequency in relatives for an all-or-none trait. *Ann Hum Genet*, 35:47–48.
339. Janssens, A. C., Aulchenko, Y. S., Elefante, S. et al. (2006). Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med*, 8:395–400.
340. Janssens, A. C. & van Duijn, C. M. (2008). Genome-based prediction of common diseases: Advances and prospects. *Hum Mol Genet*, 17:R166–173.
341. Jeffreys, A. J., Kauppi, L. & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29:217–222.
342. Jiang, R., Dong, J., Wang, D. et al. (2001). Fine-scale mapping using Hardy–Weinberg disequilibrium. *Ann Hum Genet*, 65:207–219.
343. Johnson, A. D. & O'Donnell, C. J. (2009). An open access database of genome-wide association results. *BMC Med Genet*, 10:6.
344. Kamin, L. J. & Goldberger, A. S. (2002). Twin studies in behavioral research: a skeptical view. *Theor Popul Biol*, 61:83–95.
345. Kaplan, N. L., Martin, E. R. & Weir, B. S. (1997). Power studies for the transmission/disequilibrium tests with multiple alleles. *Am J Hum Genet*, 60:691–702.
346. Karlin, S. & Liberman, U. (1978). Classifications and comparisons of multilocus recombination distributions. *PNAS USA*, 75:6332–6336.
347. Karlin, S. & Liberman, U. (1979). Representation of nonepistatic selection models and analysis of multilocus Hardy–Weinberg equilibrium configurations. *J Math Biol*, 7:353–374.
348. Kathiresan, S., Voight, B. F., Purcell, S. et al. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, 41:334–341.

349. Ke, X., Miretti, M. M., Broxholme, J. et al. (2005). A comparison of tagging methods and their tagging space. *Hum Mol Genet*, 14:2757–2767.
350. Kelly, E. D., Sievers, F. & McManus, R. (2004). Haplotype frequency estimation error analysis in the presence of missing genotype data. *BMC Bioinformatics*, 5:188.
351. Khoury, M. J. (1994). Case-parental control method in the search for disease-susceptibility genes. *Am J Hum Genet*, 55:414–415.
352. Khoury, M. J., Adams, M. J. & Flanders, W. D. (1988). An epidemiologic approach to ecogenetics. *Am J Hum Genet*, 42:89–95.
353. Khoury, M. J., Beaty, T. H. & Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press: New York.
354. Khoury, M. J., Flanders, W. D., Lipton, R. B. et al. (1991). Commentary: The affected sib-pair method in the context of an epidemiologic study design. *Genet Epidemiol*, 8:277–282.
355. Kim, M. & Ramakrishna, R. S. (2005). New indices for cluster validity assessment. *Pattern Recognit Lett*, 26:2353–2363.
356. Kirk, K. M. & Cardon, L. R. (2002). The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet*, 10:616–622.
357. Kistner, E. O. & Weinberg, C. R. (2005). A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents. *Genet Epidemiol*, 29:155–165.
358. Kleinbaum, D. G. (1994). *Logistic Regression: A Self-Learning Text*. Springer: New York.
359. Knapp, M. (1997). The affected sib pair method for linkage analysis. In: *Genetic Mapping of Disease Genes*, I. H. Pawlowitzki, J. H. Edwards & E. A. Thompson, ed., pp. 147–157. Academic Press: London.
360. Knapp, M. (1999). A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet*, 64:1177–1185.
361. Knapp, M. (1999). The transmission/disequilibrium test and parental-genotype reconstruction: The reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet*, 64:861–870.
362. Knapp, M. & Becker, T. (2003). Family-based association analysis with tightly linked markers. *Hum Hered*, 56:2–9.
363. Knapp, M., Seuchter, S. A. & Baur, M. P. (1993). The effect of misspecifying allele frequencies in incompletely typed families. *Genet Epidemiol*, 10:413–418.
364. Knapp, M., Seuchter, S. A. & Baur, M. P. (1993). The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *Am J Hum Genet*, 52:1085–1093.
365. Knapp, M., Seuchter, S. A. & Baur, M. P. (1994). Linkage analysis in nuclear families I: Optimality criteria for affected sib-pair tests. *Hum Hered*, 44:37–43.
366. Knapp, M., Seuchter, S. A. & Baur, M. P. (1994). Linkage analysis in nuclear families II: Relationship between affected sib-pair tests and LOD score analysis. *Hum Hered*, 44:44–51.
367. Knapp, M., Wassmer, G. & Baur, M. P. (1995). The relative efficiency of the Hardy–Weinberg equilibrium-likelihood and the conditional on parental genotype-likelihood methods for candidate. *Am J Hum Genet*, 57:1476–1485.
368. Knowler, W. C., Williams, R. C., Pettitt, D. J. et al. (1988). $Gm^{3;5,13,14}$ and type 2 diabetes mellitus: An association in American Indians with genetic admixture. *Am J Hum Genet*, 43:520–526.

369. Kong, A. & Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet*, 61:1179–1188.
370. König, I. R., Malley, J. D., Pajevic, S. et al. (2008). Patient-centered yes/no prognosis using learning machines. *Int J Data Min Bioinform*, 2:289–341.
371. König, I. R., Schäfer, H., Ziegler, A. et al. (2003). Reducing sample sizes in genome scans: Group sequential study designs with futility stops. *Genet Epidemiol*, 25:339–349.
372. Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Ann Eugen*, 12:172–175.
373. Kovács, F., Legány, C. & Babos, A. (2005). Cluster validity measurement techniques. In: *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, pp. 540–550. Budapest Tech: Budapest.
374. Kraft, P. (2008). Curses – winner’s and otherwise – in genetic epidemiology. *Epidemiol*, 19:649–651.
375. Kraft, P. & Spiegelman, D. (2007). Statistical issues in epidemiological studies of gene-environment interaction. In: *Gene-Environment Interactions: Role in the Modulation of Pulmonary and Autoimmune Disease Risks*, D. Christiani & P. Fraser, ed. Henry Stewart Talks Ltd.: London.
376. Kraft, P., Yen, Y. C., Stram, D. O. et al. (2007). Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*, 63:111–119.
377. Kruglyak, L., Daly, M. J. & Lander, E. S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am J Hum Genet*, 56:519–527.
378. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. et al. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am J Hum Genet*, 58:1347–1363.
379. Kruglyak, L. & Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet*, 57:439–454.
380. Kruglyak, L. & Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics*, 139:1421–1428.
381. Kruglyak, L. & Lander, E. S. (1998). Faster multipoint linkage analysis using Fourier transforms. *J Comput Biol*, 5:1–7.
382. Kuo, T. Y., Lau, W. & Collins, A. R. (2007). LDMAP: The construction of high-resolution linkage disequilibrium maps of the human genome. *Methods Mol Biol*, 376:47–57.
383. Kwok, P.-Y. (2003). *Single Nucleotide Polymorphisms*, Vol. 212. Humana Press: New York.
384. Laakso, M. & Pyörälä, K. (1985). Age of onset and type of diabetes. *Diabetes Care*, 8:114–117.
385. Laan, M. & Pääbo, S. (1997). Demographic history and linkage disequilibrium in human populations. *Nat Genet*, 17:435–438.
386. Lachin, J. M. (1998). Sample size determination. In: *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, ed., Vol. 5, pp. 3892–3903. John Wiley & Sons: New York.
387. Laird, N. M., Horvath, S. & Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genet Epidemiol*, 19:S36–S42.
388. Lake, S. L., Blacker, D. & Laird, N. M. (2000). Family-based tests of association in the presence of linkage. *Am J Hum Genet*, 67:1515–1525.
389. Lake, S. L., Lyon, H., Tantisira, K. et al. (2003). Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered*, 55:56–65.

390. Lalouel, J. M. & Morton, N. E. (1981). Complex segregation analysis with pointers. *Hum Hered*, 31:312–321.
391. Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121:185–199.
392. Lander, E. S. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *PNAS USA*, 84:2363–2367.
393. Lander, E. S. & Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat Genet*, 11:241–247.
394. Lander, E. S. & Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265:2037–2048.
395. Lange, C., Blacker, D. & Laird, N. M. (2004). Family-based association tests for survival and times-to-onset analysis. *Stat Med*, 23:179–189.
396. Lange, C., Silverman, E. K., Xu, X. et al. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*, 4:195–206.
397. Lange, K. (1986). The affected sib-pair method using identity by state relations. *Am J Hum Genet*, 39:148–150.
398. Lange, K. & Goradia, T. M. (1987). An algorithm for automatic genotype elimination. *Am J Hum Genet*, 40:250–256.
399. Lango, H., Palmer, C. N., Morris, A. D. et al. (2008). Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes*, 57:3129–3135.
400. Lao, O., Lu, T. T., Nothnagel, M. et al. (2008). Correlation between genetic and geographic structure in Europe. *Curr Biol*, 18:1241–1248.
401. Leal, S. M. (2002). X-linkage. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 791–796. John Wiley & Sons: New York.
402. Lee, H., Stram, D. O. & Thomas, D. C. (1993). A generalized estimating equations approach to fitting major gene models in segregation analysis of continuous phenotypes. *Genet Epidemiol*, 10:61–74.
403. Lee, W.-C. (2003). Searching for disease-susceptibility loci by testing for Hardy–Weinberg disequilibrium in a gene bank of affected individuals. *Am J Epidemiol*, 158:397–400.
404. Lenz, W. (1959). *Ursachen des gesteigerten Wachstums der heutigen Jugend. Wissenschaftliche Veröffentlichung der Deutschen Gesellschaft für Ernährung*, Vol. 4. Steinkopff: Darmstadt.
405. Leung, H. M. & Kupper, L. L. (1981). Comparisons of confidence intervals for attributable risk. *Biometrics*, 37:293–302.
406. Leutenegger, A. L., Genin, E., Thompson, E. A. et al. (2002). Impact of parental relationships in maximum LOD score affected sib-pair method. *Genet Epidemiol*, 23:413–425.
407. Li, B. & Leal, S. M. (2008). Ignoring intermarker linkage disequilibrium induces false-positive evidence of linkage for consanguineous pedigrees when genotype data is missing for any pedigree member. *Hum Hered*, 65:199–208.
408. Li, M., Li, C. & Guan, W. (2008). Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet*, 16:635–643.
409. Li, Q. & Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol*, 32:215–226.

410. Li, Y. & Abecasis, G. (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet*, 579:2290.
411. Li, Y. C., Korol, A. B., Fahima, T. et al. (2002). Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Mol Ecol*, 11:2453–2465.
412. Li, Y. J., Scott, W. K., Hedges, D. J. et al. (2002). Age at onset in two common neurodegenerative diseases is genetically controlled. *Am J Hum Genet*, 70:985–993.
413. Liang, K. Y. & Beaty, T. H. (1991). Measuring familial aggregation by using odds-ratio regression models. *Genet Epidemiol*, 8:361–370.
414. Liang, K. Y. & Beaty, T. H. (2000). Statistical designs for familial aggregation. *Stat Methods Med Res*, 9:543–562.
415. Liang, K.-Y., Chiu, Y.-F. & Beaty, T. H. (2001). A robust identity-by-descent procedure using affected sib pairs: Multipoint mapping for complex diseases. *Hum Hered*, 51:64–78.
416. Liang, K.-Y., Chiu, Y.-F., Beaty, T. H. et al. (2001). Multipoint analysis using affected sib pairs: Incorporating linkage evidence from unlinked regions. *Genet Epidemiol*, 21:105–122.
417. Liberman, U. & Karlin, S. (1984). Theoretical models of genetic map functions. *Theor Popul Biol*, 25:331–346.
418. Lin, D. Y. (2006). Evaluating statistical significance in two-stage genomewide association studies. *Am J Hum Genet*, 78:505–509.
419. Lin, D. Y., Hu, Y. & Huang, B. E. (2008). Simple and efficient analysis of disease association with missing genotype data. *Am J Hum Genet*, 82:444–452.
420. Lindley, D. V. (1988). Statistical inference concerning Hardy–Weinberg equilibrium. In: *Bayesian Statistics*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith, ed., pp. 307–326. Oxford University Press: New York.
421. Lindström, S., Yen, Y.-C., Spiegelman, D. et al. (2009). The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Hum Hered*, 68:171–181.
422. Ling, H., Hetrick, K., Bailey-Wilson, J. E. et al. (2009). Application of gender-specific SNP filters in GWAS data. *BMC Proc*, 3 Suppl 7:S57.
423. Liu, F., Elefante, S., van Duijn, C. M. et al. (2006). Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. *Ann Hum Genet*, 70:965–970.
424. Liu, N., Zhang, K. & Zhao, H. (2008). Haplotype-association analysis. *Adv Genet*, 60:335–405.
425. Liu, Y., Titchler, D. & Bull, S. B. (2002). A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. *Genet Epidemiol*, 22:26–40.
426. Liu, Z. & Lin, S. (2005). Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genet Epidemiol*, 29:353–364.
427. Livak, K. J., Flood, S. J., Marmaro, J. et al. (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl*, 4:357–362.
428. Long, J. C., Williams, R. C. & Urbanek, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*, 56:799–810.
429. Louis, E. J. & Dempster, E. R. (1987). An exact test for Hardy–Weinberg and multiple alleles. *Biometrics*, 43:805–811.
430. Lovmar, L., Ahlford, A., Jonsson, M. et al. (2005). Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics*, 6:35.
431. Lui, K. J. (2001). Confidence intervals of the attributable risk under cross-sectional sampling with confounders. *Biom J*, 43:767–779.

432. Lui, K. J. (2004). *Statistical Estimation of Epidemiological Risk*. John Wiley & Sons: New York.
433. Lunetta, K. L. & Rogus, J. J. (1998). Strategy for mapping minor histocompatibility genes involved in graft-versus-host resistance: A novel application of discordant sib pair methodology. *Genet Epidemiol*, 15:595–607.
434. Malécot, G. (1969). *The Mathematics of Heredity*. Freeman: San Francisco (CA).
435. Manaster, C. J., Nanthakumar, E. & Morin, P. A. (1999). Detecting null alleles with vasarely charts. In: *IEEE Visualization*, D. S. Ebert, M. Gross & B. Hamann, ed., pp. 463–466. IEEE Computer Society: San Francisco.
436. Mandal, D. M., Wilson, A. F., Keats, B. J. B. et al. (1998). Factors affecting inflation of type I error of model-based linkage under random ascertainment. *Am J Hum Genet*, 63:A298.
437. Mangs, A. H. & Morris, B. J. (2007). The human pseudoautosomal region (PAR): Origin, function and future. *Curr Genomics*, 8:129–136.
438. Maniatis, N., Collins, A., Gibson, J. et al. (2004). Positional cloning by linkage disequilibrium. *Am J Hum Genet*, 74:846–855.
439. Maniatis, N., Collins, A., Xu, C.-F. et al. (2002). The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *PNAS USA*, 99:2228–2233.
440. Marchini, J., Cardon, L. R., Phillips, M. S. et al. (2004). The effects of human population structure on large genetic association studies. *Nat Genet*, 36:512–517.
441. Marchini, J., Donnelly, P. & Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37:413–417.
442. Marchini, J., Howie, B., Myers, S. et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39:906–913.
443. Maresso, K. & Broeckel, U. (2008). Genotyping platforms for mass-throughput genotyping with SNPs, including human genome-wide scans. *Adv Genet*, 60:107–139.
444. Markianos, K., Daly, M. J. & Kruglyak, L. (2001). Efficient multipoint linkage analysis through reduction of inheritance space. *Am J Hum Genet*, 68:963–977.
445. Marmorstein, L. Y., Ouchi, T. & Aaronson, S. A. (1998). The *BRCA2* gene product functionally interacts with *p53* and *RAD51*. *PNAS USA*, 95:13869–13874.
446. Marteau, T. M. & Croyle, R. T. (1998). The new genetics. Psychological responses to genetic testing. *Brit Med J*, 316:693–696.
447. Martin, E. R., Bass, M. P. & Kaplan, N. L. (2001). Correcting for a potential bias in the pedigree disequilibrium test. *Am J Hum Genet*, 68:1065–1067.
448. Martin, E. R., Monks, S. A., Warren, L. L. et al. (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Am J Hum Genet*, 67:146–154.
449. Martinez, M., Khlat, M., Leboyer, M. et al. (1989). Performance of linkage analysis under misclassification error when the genetic model is unknown. *Genet Epidemiol*, 6:253–258.
450. Matisse, T. C. & Weeks, D. E. (1993). Detecting heterogeneity with the affected-pedigree-member (APM) method. *Genet Epidemiol*, 10:401–406.
451. McGall, G. H. & Christians, F. C. (2002). High-density genechip oligonucleotide probe arrays. *Adv Biochem Eng Biotechnol*, 77:21–42.
452. McKinney, B. A., Reif, D. M., Ritchie, M. D. et al. (2006). Machine learning for detecting gene-gene interactions: A review. *Appl Bioinformatics*, 5:77–88.
453. McKinney, B. A., Reif, D. M., White, B. C. et al. (2007). Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics*, 23:2113–2120.
454. McKusick, V. A. (1998). *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*, (12 ed.). Johns Hopkins University Press: Baltimore (MD).

455. McPeck, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol*, 16:225–249.
456. McPeck, M. S. & Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet*, 66:1076–1094.
457. Mein, C. A., Esposito, L., Dunn, M. G. et al. (1998). A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nat Genet*, 19:297–300.
458. Meldrum, D. (2000). Automation for genomics, part one: Preparation for sequencing. *Genome Res*, 10:1081–1092.
459. Meldrum, D. (2000). Automation for genomics, part two: Sequencers, microarrays, and future trends. *Genome Res*, 10:1288–1303.
460. Mendell, N. R. & Simon, G. A. (1984). A general expression for the variance-covariance matrix of estimates of gene frequency: The effects of departures from Hardy–Weinberg equilibrium. *Ann Hum Genet*, 48:283–286.
461. Meng, Z., Zaykin, D. V., Xu, C. F. et al. (2003). Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet*, 73:115–130.
462. Menges, T., König, I. R., Hossain, H. et al. (2008). Sepsis syndrome and death in trauma patients is associated with variation in the gene encoding *TNF R1*. *Crit Care Med*, 36:1456–1462.
463. Mercaldo, N. D., Lau, K. F. & Zhou, X. H. (2007). Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med*, 26:2170–2183.
464. Meyer, M. R., Tschanz, J. T., Norton, M. C. et al. (1998). APOE genotype predicts when—not whether—one is predisposed to develop Alzheimer disease. *Nat Genet*, 19:321–322.
465. Miller, C. R., Joyce, P. & Waits, L. P. (2002). Assessing allelic dropout and genotyping reliability using maximum likelihood. *Genetics*, 160:357–366.
466. Minelli, C., Thompson, J. R., Abrams, K. R. et al. (2005). The choice of a genetic model in the meta-analysis of molecular association studies. *Int J Epidemiol*, 34:1319–1328.
467. Model, F., König, T., Piepenbrock, C. et al. (2002). Statistical process control for large scale microarray experiments. *Bioinformatics*, 18:S155–S163.
468. Mohamed, S. A., Aherrahrou, Z., Liptau, H. et al. (2006). Novel missense mutations (p.T596M and p.P1797H) in *NOTCH1* in patients with bicuspid aortic valve. *Biochem Biophys Res Commun*, 345:1460–1465.
469. Moll, P. P., Burns, T. L. & Lauer, R. M. (1991). The genetic and environmental sources of body mass index variability: The Muscatine Ponderosity Family Study. *Am J Hum Genet*, 49:1243–1255.
470. Monks, S. A., Kaplan, N. L. & Weir, B. S. (1998). A comparative study of sibship tests of linkage and/or association. *Am J Hum Genet*, 63:1507–1516.
471. Monroe, K. R., Yu, M. C., Kolonel, L. N. et al. (1995). Evidence of an X-linked or recessive genetic component to prostate cancer risk. *Nat Med*, 1:827–829.
472. Moore, J. H., Gilbert, J. C., Tsai, C. T. et al. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol*, 241:252–261.
473. Moore, J. H. & White, B. C. (2007). Tuning ReliefF for genomewide genetic analysis. *Lect Notes Comput Sci*, 4447:166–175.
474. Morgan, T. H. (1911). Random segregation versus coupling in Mendelian inheritance. *Science*, 34:384.
475. Morgan, T. H. (1928). *The Theory of Genes*. Yale University Press: New Haven (CT).

476. Morris, A. P., Curnow, R. N. & Whittaker, J. C. (1997). Randomization tests of disease-marker associations. *Ann Hum Genet*, 61:49–60.
477. Morris, R. W. & Kaplan, N. L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol*, 23:221–233.
478. Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet*, 7:277–318.
479. Morton, N. E. (1998). Significance levels in complex inheritance. *Am J Hum Genet*, 62:690–697.
480. Morton, N. E. & Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *PNAS USA*, 95:11389–11393.
481. Morton, N. E. & MacLean, C. J. (1974). Analysis of family resemblance III: Complex segregation analysis of quantitative traits. *Am J Hum Genet*, 26:489–503.
482. Morton, N. E., Zhang, W., Taillon-Miller, P. et al. (2001). The optimal measure of allelic association. *PNAS USA*, 98:5217–5221.
483. Motsinger-Reif, A. A. (2008). The effect of alternative permutation testing strategies on the performance of multifactor dimensionality reduction. *BMC Res Notes*, 1:139.
484. Motsinger-Reif, A. A., Reif, D. M., Fanelli, T. J. et al. (2008). A comparison of analytical methods for genetic association studies. *Genet Epidemiol*, 32:767–778.
485. Mueller, J. C., Lohmussaar, E., Magi, R. et al. (2005). Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet*, 76:387–398.
486. Mugavin, M. E. (2008). Multidimensional scaling: A brief overview. *Nurs Res*, 57:64–68.
487. Müller, H. H., Pahl, R. & Schäfer, H. (2007). Including sampling and phenotyping costs into the optimization of two stage designs for genome wide association studies. *Genet Epidemiol*, 31:844–852.
488. Murff, H. J., Spigel, D. R. & Syngal, S. (2004). Does this patient have a family history of cancer? An evidence-based analysis of the accuracy of family cancer history. *J Am Med Assoc*, 292:1480–1489.
489. Murray, S. S., Oliphant, A., Shen, R. et al. (2004). A highly informative SNP linkage panel for human genetic studies. *Nat Methods*, 1:113–117.
490. Myers, S., Bottolo, L., Freeman, C. et al. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310:321–324.
491. Nagelkerke, N. J., Hoebee, B., Teunis, P. et al. (2004). Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet*, 12:964–970.
492. Namkung, J., Kim, K., Yi, S. et al. (2009). New evaluation measures for multifactor dimensionality reduction classifiers in gene-gene interaction analysis. *Bioinformatics*, 25:338–345.
493. Neale, M. C. & Cardon, L. R. (1992). *Methodology of Genetic Studies of Twins and Families*. Kluwer: Dordrecht.
494. Neale, M. C. & McArdle, J. J. (1990). The analysis of assortative mating: A LISREL model. *Behav Genet*, 20:287–296.
495. Neale, M. C., Neale, B. M. & Sullivan, P. F. (2002). Nonpaternity in linkage studies of extremely discordant sib pairs. *Am J Hum Genet*, 70:526–529.
496. Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Stat Med*, 17:873–890.
497. Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat Med*, 17:857–872.

498. Newman, S. C. (2001). *Biostatistical Methods in Epidemiology*. John Wiley & Sons: New York.
499. Nica, A. C. & Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Hum Mol Genet*, 17:R129–R134.
500. Nielsen, D. M., Ehm, M. G., Zaykin, D. V. et al. (2004). Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics*, 168:1029–1040.
501. Nishida, N., Koike, A., Tajima, A. et al. (2008). Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. *BMC Genomics*, 9:431.
502. Niu, T. (2004). Algorithms for inferring haplotypes. *Genet Epidemiol*, 27:334–347.
503. Niu, T., Qin, Z. S., Xu, X. et al. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*, 70:157–169.
504. North, B. V., Curtis, D. & Sham, P. C. (2002). A note on calculation of empirical *P* values from Monte Carlo procedure. *Am J Hum Genet*, 71:439–441.
505. North, B. V., Curtis, D. & Sham, P. C. (2003). A note on calculation of empirical *P* values from Monte Carlo procedure. *Am J Hum Genet*, 72:498–499.
506. Nothnagel, M., Ellinghaus, D., Schreiber, S. et al. (2009). A comprehensive evaluation of SNP genotype imputation. *Hum Genet*, 125:163–171.
507. Nothnagel, M., Fürst, R. & Rohde, K. (2002). Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered*, 54:186–198.
508. Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*, 74:765–769.
509. Nyrén, P., Pettersson, B. & Uhlén, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem*, 208:171–175.
510. O'Connell, J. R. & Weeks, D. E. (1995). The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recording and fuzzy inheritance. *Nat Genet*, 11:402–408.
511. O'Connell, J. R. & Weeks, D. E. (1998). PEDCHECK: A program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet*, 63:259–266.
512. Oliphant, A., Barker, D. L., Stuelpnagel, J. R. et al. (2002). BeadArray technology: Enabling an accurate, cost-effective approach to highthroughput genotyping. *Biotechniques*, Suppl 32:56–58, 60–61.
513. Olson, J. M. (1995). Multipoint linkage analysis using sib pairs: An interval mapping approach for dichotomous outcomes. *Am J Hum Genet*, 56:788–798.
514. Olson, J. M. (1999). A general conditional-logistic regression model for affected relative pair linkage studies. *Am J Hum Genet*, 65:1760–1769.
515. Ott, J. (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. *Am J Hum Genet*, 51:283–290.
516. Ott, J. (1999). *Analysis of Human Genetic Linkage*, (3 ed.). Johns Hopkins University Press: Baltimore (MD).
517. Pahl, R., Schäfer, H. & Müller, H. H. (2009). Optimal multistage designs – a general framework for efficient genome-wide association studies. *Biostatistics*, 10:297–309.
518. Palmer, L. J., Jacobs, K. B. & Elston, R. C. (2000). Haseman and Elston revisited: The effects of ascertainment and residual familial correlations on power to detect linkage. *Genet Epidemiol*, 19:456–460.

519. Pardo-Manuel de Villena, F. & Sapienza, C. (2001). Nonrandom segregation during meiosis: The unfairness of females. *Mamm Genome*, 12:331–339.
520. Patil, N., Berno, A. J., Hinds, D. A. et al. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294:1719–1723.
521. Patterson, N., Price, A. L. & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2:e190.
522. Pearson, T. A. & Manolio, T. A. (2008). How to interpret a genome-wide association study. *J Am Med Assoc*, 299:1335–1344.
523. Pe'er, I., Yelensky, R., Altshuler, D. et al. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol*, 32:381–385.
524. Pei, Y. F., Li, J., Zhang, L. et al. (2008). Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE*, 3:e3551.
525. Penrose, L. S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann Eugen*, 6:133–138.
526. Penrose, L. S. (1938). Genetic linkage in graded human characters. *Ann Eugen*, 8:233–237.
527. Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford.
528. Pepe, M. S., Feng, Z., Huang, Y. et al. (2006). Integrating the predictiveness of a marker with its performance as a classifier. Technical report, University of Washington.
529. Pepe, M. S., Janes, H., Longton, G. et al. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*, 159:882–890.
530. Pereira, C. & Rogatko, A. (1984). The Hardy–Weinberg equilibrium under a Bayesian perspective. *Rev Bras Genet*, 4:689–707.
531. Pettigrew, H. M., Gart, J. J. & Thomas, D. G. (1986). The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika*, 73:425–435.
532. Pfeiffer, R. M. & Gail, M. H. (2003). Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genet Epidemiol*, 25:136–148.
533. Pfeiffer, R. M., Pee, D. & Landi, M. T. (2008). On combining family and case-control studies. *Genet Epidemiol*, 32:638–646.
534. Phillips, C. (2007). Online resources for SNP analysis: A review and route map. *Mol Biotechnol*, 35:65–97.
535. Phillips, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, 9:855–867.
536. Plagnol, V., Cooper, J. D., Todd, J. A. et al. (2007). A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet*, 3:e74.
537. Ploughman, L. M. & Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *Am J Hum Genet*, 44:543–551.
538. Podgoreanu, M. V., White, W. D., Morris, R. W. et al. (2006). Inflammatory gene polymorphisms and risk of postoperative myocardial infarction after cardiac surgery. *Circulation*, 114:1275–281.
539. Pompanon, F., Bonin, A., Bellemain, E. et al. (2005). Genotyping errors: Causes, consequences and solutions. *Nat Rev Genet*, 6:847–859.
540. Poznik, G. D., Adamska, K., Xu, X. et al. (2006). A novel framework for sib pair linkage analysis. *Am J Hum Genet*, 78:222–230.

541. Price, A. L., Patterson, N. J., Plenge, R. M. et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38:904–909.
542. Price, R. M. & Bonett, D. G. (2008). Confidence intervals for a ratio of two independent binomial proportions. *Stat Med*, 27:5497–5508.
543. Pritchard, J. K. & Cox, N. J. (2002). The allelic architecture of human disease genes: Common disease-common variant. . . or not? *Hum Mol Genet*, 11:2417–2423.
544. Pritchard, J. K. & Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theor Popul Biol*, 60:227–237.
545. Pritchard, J. K. & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, 65:220–228.
546. Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.
547. Pritchard, J. K., Stephens, M., Rosenberg, N. A. et al. (2000). Association mapping in structured populations. *Am J Hum Genet*, 67:170–181.
548. Purcell, S., Neale, B., Todd-Brown, K. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81:559–575.
549. Purcell, S. & Sham, P. (2004). Properties of structured association approaches to detecting population stratification. *Hum Hered*, 58:93–107.
550. Putter, H., Sandkuijl, L. A. & van Houwelingen, J. C. (2002). Score test for detecting linkage to quantitative traits. *Genet Epidemiol*, 22:345–355.
551. Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum Hered*, 47:342–350.
552. Rabinowitz, D. & Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered*, 50:211–223.
553. Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet*, 10:117–133.
554. Ragoussis, J., Elvidge, G. P., Kaur, K. et al. (2006). Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genet*, 2:e100.
555. Randhawa, J. S. & Easton, A. J. (1999). Demystified. DNA nucleotide sequencing. *Mol Pathol*, 52:117–124.
556. Reich, D. E., Cargill, M., Bolk, S. et al. (2001). Linkage disequilibrium in the human genome. *Nature*, 411:199–204.
557. Reich, K., Mössner, R., König, I. R. et al. (2002). Promoter polymorphisms of the genes encoding tumor necrosis factor- α and interleukin-1 β are associated with different subtypes of psoriasis characterized by early and late disease onset. *J Invest Dermatol*, 118:155–63.
558. Reik, W. & Walter, J. (2001). Genomic imprinting: Parental influence on the genome. *Nat Rev Genet*, 2:21–32.
559. Rice, T. K. (2008). Familial resemblance and heritability. *Adv Genet*, 60:35–49.
560. Risch, N. (1987). Assessing the role of HLA-linked and unlinked determinants of disease. *Am J Hum Genet*, 40:1–14.
561. Risch, N. (1990). Linkage strategies for genetically complex traits. I: Multilocus models. *Am J Hum Genet*, 46:222–228.
562. Risch, N. (1990). Linkage strategies for genetically complex traits. II: Power of affected relative pairs. *Am J Hum Genet*, 46:229–241.
563. Risch, N. (1990). Linkage strategies for genetically complex traits. III: The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet*, 46:242–253.

564. Risch, N. & Giuffra, L. (1991). Model misspecification and multipoint linkage analysis. *Hum Hered*, 42:77–92.
565. Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273:1516–1517.
566. Risch, N. & Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, 268:1584–1589.
567. Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, 405:847–856.
568. Robertson, A. (1973). Linkage between marker loci and those affecting a quantitative trait. *Behav Genet*, 3:389–391.
569. Robnik-Šikonja, M. & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn*, 53:23–69.
570. Rothman, K. J. & Greenland, S. (1998). *Modern Epidemiology*, (2 ed.). Lippincott Williams & Wilkins: Baltimore (MD).
571. Rothman, K. J., Greenland, S. & Walker, A. M. (1980). Concepts of interaction. *Am J Epidemiol*, 112:467–470.
572. Rousseeuw, P. (1987). Silhouettes—a graphical aid to the interpretation and validation of cluster-analysis. *J Comput Appl Math*, 20:53–65.
573. Rücker, R., Schwarzer, G., Carpenter, J. R. et al. (2008). Undue reliance on i^2 in assessing heterogeneity may mislead. *BMC Med Res Methodol*, 8:79.
574. Saiki, R. K., Scharf, S., Faloona, F. et al. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230:1350–1354.
575. Salem, R. M., Wessel, J. & Schork, N. J. (2005). A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum Genomics*, 2:39–66.
576. Samani, N. J., Erdmann, J., Hall, A. S. et al. (2007). Genome-wide association analysis of coronary artery disease. *New Engl J Med*, 357:443–453.
577. Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics*, 53:1253–1261.
578. Satten, G. A. & Epstein, M. P. (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol*, 27:192–201.
579. Satten, G. A., Flanders, W. D. & Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet*, 68:466–477.
580. Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genet Epidemiol*, 27:348–364.
581. Schaid, D. J. (2004). Genetic epidemiology and haplotypes. *Genet Epidemiol*, 27:317–320.
582. Schaid, D. J., Olson, J. B., Gauderman, J. et al. (2003). Regression models for linkage: Issues of traits, covariates, heterogeneity, and interaction. *Hum Hered*, 55:86–96.
583. Schaid, D. J. & Rowland, C. (1998). Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *Am J Hum Genet*, 63:1492–1506.
584. Schaid, D. J., Rowland, C. M., Tines, D. E. et al. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*, 70:425–434.

585. Schaid, D. J. & Sommer, S. S. (1993). Genotype relative risks: Methods for design and analysis of candidate-gene association studies. *Am J Hum Genet*, 53:1114–1126.
586. Schaid, D. J. & Sommer, S. S. (1994). Comparison of statistics for candidate-gene association studies using cases and parents. *Am J Hum Genet*, 55:402–409.
587. Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78:629–644.
588. Scheinfeld, A. (1964). *Your Heredity and Environment*. J. B. Lippincott Co.: Philadelphia.
589. Scherag, A., Dempfle, A., Hinney, A. et al. (2002). Confidence intervals for genotype relative risks and allele frequencies from the case parent trio design for candidate-gene studies. *Hum Hered*, 54:210–217; Erratum in: *Hum Hered* 2003; 55:26.
590. Scherag, A., Hebebrand, J., Schäfer, H. et al. (2009). Flexible designs for genomewide association studies. *Biometrics*, 65:815–821.
591. Schillert, A., Schwarz, D. F., Vens, M. et al. (2009). ACPA: Automated cluster plot analysis of genotype data. *BMC Proc*, 3 Suppl 7:S59.
592. Schlesselman, J. J. (1982). *Case-Control Studies. Design, Conduct and Analysis*. Oxford University Press: New York.
593. Schlotterer, C. (2004). The evolution of molecular markers – just a matter of fashion? *Nat Rev Genet*, 5:63–69.
594. Schmutz, S. M. & Berryere, T. G. (2007). Genes affecting coat colour and pattern in domestic dogs: A review. *Anim Genet*, 38:539–549.
595. Schmutz, S. M., Berryere, T. G. & Goldfinch, A. D. (2002). *TYRP1* and *MC1R* genotypes and their effects on coat color in dogs. *Mamm Genome*, 13:380–387.
596. Schulz, K. F. & Grimes, D. A. (2002). Case-control studies: Research in reverse. *Lancet*, 359:431–434.
597. Schunkert, H., Götz, A., Braund, P. et al. (2008). Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. *Circulation*, 117:1675–1684.
598. Schwender, H. & Ickstadt, K. (2008). Identification of SNP interactions using logic regression. *Biostatistics*, 9:187–198.
599. Searle, S. R., Casella, G. & McCulloch, C. E. (1994). *Variance Components*. John Wiley & Sons: New York.
600. Sebastiani, P., Abad, M. M., Alpargu, G. et al. (2004). Robust transmission/disequilibrium test for incomplete family genotypes. *Genetics*, 168:2329–2337.
601. Sebastiani, P., Gussoni, E., Kohane, I. S. et al. (2003). Statistical challenges in functional genomics. *Stat Sci*, 18:33–70.
602. Seldin, M. F., Shigeta, R., Villoslada, P. et al. (2006). European population substructure: Clustering of northern and southern populations. *PLoS Genet*, 2:e143.
603. Self, S. G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc*, 82:605–610.
604. Self, S. G., Longton, G., Kopecky, K. J. et al. (1991). On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics*, 47:53–61.
605. Sengul, H., Weeks, D. E. & Feingold, E. (2001). A survey of affected-sibship statistics for nonparametric linkage analysis. *Am J Hum Genet*, 69:179–190.
606. Setakis, E., Stirnadel, H. & Balding, D. J. (2006). Logistic regression protects against population structure in genetic association studies. *Genome Res*, 16:290–296.

607. Sham, P. C. & Curtis, D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet*, 59:323–336.
608. Sham, P. C. & Purcell, S. (2001). Equivalence between Haseman–Elston and variance-components linkage analyses for sib pairs. *Am J Hum Genet*, 68:1527–1532.
609. Shaw, G. M., Rozen, R., Finnell, R. H. et al. (1998). Maternal vitamin use, genetic variation of infant methylenetetrahydrofolate reductase, and risk for spina bifida. *Am J Epidemiol*, 148:30–37.
610. Shen, G. Q., Abdullah, K. G. & Wang, Q. K. (2009). The TaqMan method for SNP genotyping. *Methods Mol Biol*, 578:293–306.
611. Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–398.
612. Sherry, S. T., Ward, M. H., Kholodov, M. et al. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res*, 29:308–311.
613. Shete, S., Beasley, T. M., Etzel, C. J. et al. (2004). Effect of Winsorization on power and type 1 error of variance components and related methods of QTL detection. *Behav Genet*, 34:153–159.
614. Shete, S., Tiwari, H. & Elston, R. C. (2000). On estimating the heterozygosity and polymorphism information content value. *Theor Popul Biol*, 57:265–271.
615. Siegmund, K. (2002). Age-of-onset estimation. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 10–17. John Wiley & Sons: New York.
616. Siemiatycki, J. & Thomas, D. C. (1981). Biological models and statistical interactions: An example from multistage carcinogenesis. *Int J Epidemiol*, 10:383–387.
617. Simon, T., Verstuyft, C., Mary-Krause, M. et al. (2009). Genetic determinants of response to clopidogrel and cardiovascular events. *N Engl J Med*, 360:363–375.
618. Skol, A. D., Scott, L. J., Abecasis, G. R. et al. (2007). Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol*, 31:776–788.
619. Slager, S. L. & Schaid, D. J. (2001). Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Hum Hered*, 52:149–153.
620. Sobel, E., Sengul, H. & Weeks, D. E. (2001). Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum Hered*, 52:121–131.
621. Song, K. & Elston, R. C. (2006). A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Stat Med*, 25:105–126.
622. Soubrier, F. & Cambien, F. (1994). The angiotensin I-converting enzyme gene polymorphism: Implication in hypertension and myocardial infarction. *Curr Opin Nephrol Hypertens*, 3:25–29.
623. Speed, T. (2002). Map functions. In: *Biostatistical Genetics and Genetic Epidemiology*, R. C. Elston, J. M. Olson & L. Palmer, ed., pp. 491–495. John Wiley & Sons: New York.
624. Spencer, C. C., Su, Z., Donnelly, P. et al. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5:e1000477.
625. Spielman, R. S. & Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet*, 59:983–989.
626. Spielman, R. S. & Ewens, W. J. (1998). A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *Am J Hum Genet*, 62:450–458.

627. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 52:506–516.
628. Sribney, W. M. & Swift, M. (1992). Power of sib-pair and sib-trio linkage analysis with assortative mating and multiple disease loci. *Am J Hum Genet*, 51:773–784.
629. Steemers, F. J. & Gunderson, K. L. (2007). Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J*, 2:41–49.
630. Steffens, M., Lamina, C., Illig, T. et al. (2006). SNP-based analysis of genetic substructure in the German population. *Hum Hered*, 62:20–29.
631. Stephens, M. & Donnelly, P. (2000). Inference in molecular population genetics. *J R Statist Soc B*, 62:605–655.
632. Stephens, M. & Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73:1162–1169.
633. Stephens, M. & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, 76:449–462.
634. Stephens, M., Smith, N. J. & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68:978–989.
635. Stoll, M., Corneliussen, B., Costello, C. et al. (2004). Genetic variation in *DLG5* is associated with inflammatory bowel disease. *Nat Genet*, 36:476–480.
636. Strachan, T. & Read, A. P. (2004). *Human Molecular Genetics* 3. John Wiley & Sons: New York.
637. Stram, D. O. (2005). Software for tag single nucleotide polymorphism selection. *Hum Genomics*, 2:144–151.
638. Stram, D. O., Lee, H. & Thomas, D. C. (1993). Use of generalized estimating equations in segregation analysis of continuous outcomes. *Genet Epidemiol*, 10:575–579.
639. Stram, D. O., Leigh Pearce, C., Bretsky, P. et al. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered*, 55:179–190.
640. Strasser, H. & Weber, C. (1999). On the asymptotic theory of permutation statistics. *Math Meth Stat*, 8:220–250.
641. Strauch, K. (2002). *Kopplungsanalyse bei genetisch komplexen Erkrankungen mit genomischer Imprinting und Zwei-Genort-Krankheitsmodellen*. Urban & Vogel: München.
642. Strauch, K. (2006). Multilocus linkage analysis. In: *Nature Encyclopedia of the Human Genome*, Vol. 4, pp. 166–169. Macmillan Publishers Ltd.: London.
643. Strauch, K., Fimmers, R., Kurz, T. et al. (2000). Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: Application to mite sensitization. *Am J Hum Genet*, 66:1945–1957.
644. Stricker, C., Fernando, R. L. & Elston, R. C. (1995). An algorithm to approximate the likelihood for pedigree data with loops by cutting. *Theor Appl Genet*, 91:1054–1063.
645. Stricker, C., Fernando, R. L. & Elston, R. C. (1995). Linkage analysis with an alternative formulation for the mixed model of inheritance: The finite polygenic mixture model. *Genetics*, 141:1651–1656.
646. Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool*, 14:43–59.
647. Sturtevant, A. H. (2001). *A History of Genetics*. Cold Spring Harbor Laboratory Press and Electronic Scholarly Publishing Project, <http://www.esp.org/books/sturt/history>.
648. Suarez, B. K. & Hodge, S. E. (1979). A simple method to detect linkage for rare recessive diseases: An application to juvenile diabetes. *Clin Genet*, 15:126–136.

649. Suarez, B. K., Rice, J. & Reich, T. (1978). The generalized sib pair IBD distribution: Its use in the detection of linkage. *Ann Hum Genet*, 42:87–94.
650. Sun, F., Flanders, W. D., Yang, Q. et al. (1999). Transmission disequilibrium test (TDT) when only one parent is available: The 1-TDT. *Am J Epidemiol*, 150:97–104.
651. Sun, L., Wildera, K. & McPeck, M. S. (2002). Enhanced pedigree error detection. *Hum Hered*, 54:99–110.
652. Sun, X. & Guo, B. (2006). Genotyping single-nucleotide polymorphisms by matrix-assisted laser desorption/ionization time-of-flight-based mini-sequencing. *Methods Mol Med*, 128:225–230.
653. Sutton, A. J., Abrams, K. R., Jones, D. R. et al. (2000). *Methods for Meta-Analysis in Medical Research*. John Wiley & Sons: Chichester.
654. Tang, H. K. & Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics*, 2:147–162.
655. te Meerman, G. J. & Van der Meulen, M. A. (1997). Genomic sharing surrounding alleles identical by descent: Effects of genetic drift and population growth. *Genet Epidemiol*, 14:1125–1130.
656. Templeton, J. W., Stewart, A. P. & Fletcher, W. S. (1977). Coat color genetics in labrador retriever. *J Hered*, 68:134–136.
657. Teo, Y. Y. (2008). Common statistical issues in genome-wide association studies: A review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol*, 19:133–143.
658. Teo, Y. Y., Fry, A. E., Clark, T. G. et al. (2007). On the usage of HWE for identifying genotyping errors. *Ann Hum Genet*, 71:701–703.
659. Teo, Y. Y., Small, K. S., Clark, T. G. et al. (2008). Perturbation analysis: A simple method for filtering snps with erroneous genotyping in genome-wide association studies. *Ann Hum Genet*, 72:368–374.
660. Terwilliger, J. D. & Ott, J. (1994). *Handbook of Human Genetic Linkage*. Johns Hopkins University Press: Baltimore (MD).
661. Thomas, D. C. & Witte, J. S. (2002). Point: Population stratification: A problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev*, 11:505–512.
662. Thompson, S. G. (1993). Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Stat Methods Med Res*, 2:173–192.
663. Tobler, A. R., Short, S., Andersen, M. R. et al. (2005). The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J Biomol Tech*, 16:398–406.
664. Todorov, A. A., Province, M. A., Borecki, I. B. et al. (1997). Trade-off between sibship size and sampling scheme for detecting quantitative trait loci. *Hum Hered*, 47:1–5.
665. Török, H. P., Glas, J., Endres, I. et al. (2009). Epistasis between toll-like receptor-9 polymorphisms and variants in *NOD2* and *IL23R* modulates susceptibility to Crohn's disease. *Am J Gastroenterol*, 104:1723–1733.
666. Tosteson, T. D., Rosner, B. & Redline, S. (1991). Logistic regression for clustered binary data in proband studies with application to familial aggregation of sleep disorders. *Biometrics*, 47:1257–1265.
667. Trégouët, D. A., Escolano, S., Tired, L. et al. (2004). A new algorithm for haplotype-based association analysis: The Stochastic-EM algorithm. *Ann Hum Genet*, 68:165–177.
668. Trégouët, D.-A., König, I. R., Erdmann, J. et al. (2009). A genome-wide haplotype association study identifies the *SLC22A3-LPAL2-LPA* gene cluster as a strong susceptibility locus for coronary artery disease. *Nat Genet*, 41:283–285.

669. Trikalinos, T. A., Salanti, G., Zintzaras, E. et al. (2008). Meta-analysis methods. *Adv Genet*, 60:311–334.
670. Tritchler, D., Liu, Y. & Fallah, S. (2003). A test of linkage for complex discrete and continuous traits in nuclear families. *Biometrics*, 59:382–392.
671. Tsai, W. Y., Heiman, G. A. & Hodge, S. E. (2005). New simple tests for age-at-onset anticipation: Application to panic disorder. *Genet Epidemiol*, 28:256–260.
672. Uspensky, J. V. (1948). *Theory of Equations*. McGraw-Hill: New York.
673. van den Oord, E. J. & Neale, B. M. (2004). Will haplotype maps be useful for finding genes? *Mol Psychiatry*, 9:227–236.
674. van Hoek, M., Dehghan, A., Witteman, J. C. et al. (2008). Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: A population-based study. *Diabetes*, 57:3122–3128.
675. van Oosterhout, C., Hutchinson, W. F., Wills, D. P. et al. (2004). MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. *Mol Ecol Notes*, 4:535–538.
676. Vaxillaire, M., Veslot, J., Dina, C. et al. (2008). Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. *Diabetes*, 57:244–254.
677. Vieland, V. & Huang, J. (1998). Statistical evaluation of age-at-onset anticipation: A new test and evaluation of its behavior in realistic applications. *Am J Hum Genet*, 62:1212–1227.
678. Villanueva, R. (2003). Sibling recurrence risk in autoimmune thyroid disease. *Thyroid*, 13:761–764.
679. Vogel, F. & Motulsky, A. G. (1996). *Human Genetics: Problems and Approaches*, (3 ed.). Springer: New York.
680. Wacholder, S., Chanock, S., Garcia-Closas, M. et al. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 96:434–442.
681. Wacholder, S., Rothman, N. & Caporaso, N. (2002). Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev*, 11:513–520.
682. Wain, H. M., Bruford, E. A., Lovering, R. C. et al. (2002). Guidelines for human gene nomenclature. *Genomics*, 79:464–470.
683. Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet*, 81:208–227.
684. Wald, A. (1947). *Sequential Analysis*. John Wiley & Sons: New York.
685. Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika*, 62:371–375.
686. Wan, Y., Cohen, J. & Guerra, R. (1997). A permutation test for the robust sib-pair linkage method. *Ann Hum Genet*, 61:79–87.
687. Wang, D., Lin, S., Cheng, R. et al. (2001). Transformation of sib-pair values for the Haseman–Elston method. *Am J Hum Genet*, 68:1238–1249.
688. Wang, H., Thomas, D. C., Pe'er, I. et al. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol*, 30:356–368.
689. Wang, K. (2008). Genetic association tests in the presence of epistasis or gene-environment interaction. *Genet Epidemiol*, 32:606–614.
690. Wang, K. & Huang, J. (2002). A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *Am J Hum Genet*, 70:412–424.
691. Wang, T. & Elston, R. C. (2004). A modified revisited Haseman–Elston method to further improve power. *Hum Hered*, 57:109–116.

692. Wang, T. & Elston, R. C. (2005). Two-level Haseman–Elston regression for general pedigree data analysis. *Genet Epidemiol*, 29:12–22.
693. Wassmer, G. & Rüger, B. (1998). Profile-a tests. In: *Encyclopedia of Statistical Sciences, Update 2*, S. Kotz & N. L. Johnson, ed., pp. 549–554. John Wiley & Sons: New York.
694. Watkins, W. S., Zenger, R., O'Brien, E. et al. (1994). Linkage disequilibrium patterns vary with chromosomal location: A case study from the von Willebrand factor region. *Am J Hum Genet*, 55:348–355.
695. Weale, M. E., Depondt, C., Macdonald, S. J. et al. (2003). Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene *SCN1A*: Implications for linkage-disequilibrium gene mapping. *Am J Hum Genet*, 73:551–565.
696. Weber, J. L. & Broman, K. W. (2001). Genotyping for human whole-genome scans: Past, present, and future. *Adv Genet*, 42:77–96.
697. Weeks, D. E. & Harby, L. D. (1995). The affected-pedigree-member method: Power to detect linkage. *Hum Hered*, 45:13–24.
698. Weeks, D. E. & Lange, K. (1988). The affected-pedigree member method of linkage analysis. *Am J Hum Genet*, 42:315–326.
699. Weeks, D. E. & Lange, K. (1992). A multilocus extension of the affected-pedigree-member method of linkage analysis. *Am J Hum Genet*, 50:859–868.
700. Weeks, D. E., Ott, J. & Lathrop, G. M. (1990). SLINK: A general simulation program for linkage analysis. *Am J Hum Genet*, 47:A204 (abstr).
701. Weeks, D. E., Valappil, T. I., Schroeder, M. et al. (1995). An X-linked version of the affected-pedigree-member method of linkage analysis. *Hum Hered*, 45:25–33.
702. Weinberg, C. R. (1999). Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet*, 64:1186–1193.
703. Weinberg, C. R. & Morris, R. W. (2003). Invited commentary: Testing for Hardy–Weinberg disequilibrium using a genome single-nucleotide polymorphism scan based on cases only. *Am J Epidemiol*, 158:401–403.
704. Weir, B. S. (1979). Inferences about linkage disequilibrium. *Biometrics*, 35:235–254.
705. Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer: Sunderland (MA).
706. Weir, B. S. & Cockerham, C. C. (1978). Testing hypotheses about linkage disequilibrium with multiple alleles. *Genetics*, 88:633–642.
707. Weir, B. S. & Cockerham, C. C. (1979). Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 42:105–111.
708. Wellcome Trust Case Control Consortium, W. T. C. C. C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678.
709. Wellek, S. (2003). *Testing Statistical Hypotheses of Equivalence*, Chapter 8. Chapman & Hall/CRC: Boca Raton.
710. Wellek, S. (2004). Tests for establishing compatibility of an observed genotype distribution with Hardy–Weinberg equilibrium in the case of a biallelic locus. *Biometrics*, 60:694–703.
711. Wellek, S., Goddard, K. A. B. & Ziegler, A. (2010). A confidence-limit-based approach to the assessment of Hardy–Weinberg equilibrium. *Biom J*, in press.
712. Wellek, S. & Ziegler, A. (2009). A genotype-based approach to assessing the association between single nucleotide polymorphisms. *Hum Hered*, 67:128–139.
713. Wells, R. D., Dere, R., Hebert, M. L. et al. (2005). Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res*, 33:3785–3798.

714. Wheeler, E. & Cordell, H. J. (2007). Quantitative trait association in parent offspring trios: Extension of case/pseudocontrol method and comparison of prospective and retrospective approaches. *Genet Epidemiol*, 31:813–833.
715. Whitehead, A. (2002). *Meta-Analysis of Controlled Clinical Trials*. John Wiley & Sons: Chichester.
716. Whittemore, A. S. & Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics*, 50:118–127.
717. Whittemore, A. S. & Halpern, J. (1994). Probability of gene identity by descent: Computation and applications. *Biometrics*, 50:109–117.
718. Whittemore, A. S. & Halpern, J. (2003). Logistic regression of family data from retrospective study designs. *Genet Epidemiol*, 25:177–189.
719. Whittemore, A. S. & Tu, I. P. (1998). Simple, robust linkage tests for affected sibs. *Am J Hum Genet*, 62:1228–1242.
720. Wilcox, M. A., Pugh, E. W., Zhang, H. et al. (2005). Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: Presentation groups 1, 2, and 3. *Genet Epidemiol*, 29 Suppl 1:S7–S28.
721. Williamson, J. A. & Amos, C. I. (1990). On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. *Genet Epidemiol*, 7:309–318.
722. Williamson, J. A. & Amos, C. I. (1995). Guess LOD approach: Sufficient conditions for robustness. *Genet Epidemiol*, 12:163–176.
723. Wilson, A. F. & Elston, R. C. (1993). Statistical validity of the Haseman–Elston sib-pair test in small samples. *Genet Epidemiol*, 10:593–598.
724. Wilson, R. K., Chen, C., Avdalovic, N. et al. (1990). Development of an automated procedure for fluorescent DNA sequencing. *Genomics*, 6:626–634.
725. Witte, J. S., Elston, R. C. & Schork, N. J. (1996). Genetic dissection of complex traits. *Nat Genet*, 12:355–356.
726. Witte, J. S., Gauderman, W. J. & Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: Basic family designs. *Am J Epidemiol*, 149:693–705.
727. Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. (2005). Rational inferences about departures from Hardy–Weinberg equilibrium. *Am J Hum Genet*, 76:967–986.
728. Woolf, B. (1955). On estimating the relationship between blood group and disease. *Ann Hum Genet*, 19:251–253.
729. Wray, N. R., Goddard, M. E. & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res*, 17:1520–1528.
730. Wright, A. F., Carothers, A. D. & Pirastu, M. (1999). Population choice in mapping genes for complex diseases. *Nat Genet*, 23:397–404.
731. Wright, F. A. (1997). The phenotypic difference discards sib-pair QTL linkage information. *Am J Hum Genet*, 60:740–742.
732. Wright, S. (1951). The general structure of populations. *Ann Eugen*, 15:323–354.
733. Yu, L., Li, J., Wu, X. et al. (2002). Study on the mutation of human short tandem repeats at three loci. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*, 19:308–312.
734. Yuan, K.-H., Bentler, P. M. & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis. *Sociol Methods Res*, 34:240–258.
735. Zaykin, D. V. & Zhivotovsky, L. A. (2005). Ranks of genuine associations in whole-genome scans. *Genetics*, 171:813–823.

736. Zeegers, M. P., Rice, J. P., Rijdsdijk, F. V. et al. (2003). Regression-based sib pair linkage analysis for binary traits. *Hum Hered*, 55:125–131.
737. Zeggini, E., Rayner, W., Morris, A. P. et al. (2005). An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat Genet*, 37:1320–1322.
738. Zeng, Z.-B. (2003). Quantitative trait loci (QTL) mapping. In: *Nature Encyclopedia of the Human Genome*, Vol. 4, pp. 963–968. Macmillan Publishers Ltd.: London.
739. Zhang, S., Sha, Q., Chen, H. S. et al. (2003). Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet*, 73:566–579.
740. Zhang, W., Collins, A., Maniatis, N. et al. (2002). Properties of linkage disequilibrium (LD) maps. *PNAS USA*, 99:17004–17007.
741. Zhang, W., Collins, A. & Morton, N. E. (2004). Does haplotype diversity predict power for association mapping of disease susceptibility? *Hum Genet*, 115:157–164.
742. Zhao, H., Merikangas, K. R. & Kidd, K. K. (1999). On a randomization procedure in linkage analysis. *Am J Hum Genet*, 65:1449–1456.
743. Zhao, H., Zhang, S., Merikangas, K. R. et al. (2000). Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet*, 67:936–946.
744. Zhao, J., Jin, L. & Xiong, M. (2006). Test for interaction between two unlinked loci. *Am J Hum Genet*, 79:831–845.
745. Zhao, L. P. (1994). Segregation analysis of human pedigrees using estimating equations. *Biometrika*, 81:197–209.
746. Zhao, L. P., Davidov, O., Quiaoit, F. et al. (2002). Combined association, segregation and aggregation analysis on case-control family data. *Biostatistics*, 3:315–329.
747. Zhao, L. P., Hsu, L., Davidov, O. et al. (1997). Population-based family study designs: An interdisciplinary research framework for genetic epidemiology. *Genet Epidemiol*, 14:365–388.
748. Zhao, Z., Timofeev, N., Hartley, S. W. et al. (2008). Imputation of missing genotypes: An empirical evaluation of IMPUTE. *BMC Genet*, 9:85.
749. Zheng, G., Freidlin, B., Li, Z. et al. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biom J*, 45:335–348.
750. Zheng, G., Freidlin, B., Li, Z. et al. (2005). Genomic control for association studies under various genetic models. *Biometrics*, 61:186–192.
751. Zheng, G., Joo, J., Lin, J. P. et al. (2007). Robust ranks of true associations in genome-wide case-control association studies. *BMC Proc*, 1 Suppl 1:S165.
752. Zheng, G., Joo, J. & Yang, Y. (2009). Pearson's test, trend test, and MAX are all trend tests with different types of scores. *Ann Hum Genet*, 73:133–140.
753. Zheng, G., Joo, J., Zhang, C. et al. (2007). Testing association for markers on the X chromosome. *Genet Epidemiol*, 31:834–843.
754. Zheng, G. & Ng, H. K. (2008). Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics*, 9:391–399.
755. Zheng, G. & Tian, X. (2006). Robust trend tests for genetic association using matched case-control design. *Stat Med*, 25:3160–3173.
756. Zheng, S. L., Sun, J., Wiklund, F. et al. (2008). Cumulative association of five genetic variants with prostate cancer. *New Engl J Med*, 358:910–919.
757. Zhou, X.-H. & Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons: New York.
758. Zhu, C. & Yu, J. (2009). Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics*, 182:875–888.

759. Zhu, X., Li, S., Cooper, R. S. et al. (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet*, 82:352–365.
760. Ziegler, A. (2009). Genome-wide association studies: Quality control and population-based measures. *Genet Epidemiol*, 33:S45–S50.
761. Ziegler, A., Barth, N., Coners, H. et al. (2006). Practical considerations on the use of extreme sib-pairs for obesity. *Methods Inf Med*, 45:419–423.
762. Ziegler, A., Blettner, M., Kastner, C. et al. (1998). Identifying influential families using regression diagnostics for generalized estimating equations. *Genet Epidemiol*, 15:341–37–353.
763. Ziegler, A., DeStefano, A., König, I. R. et al. (2007). Data mining, neural nets, trees—problems 2 and 3 of Genetic Analysis Workshop 15. *Genet Epidemiol*, 33:S51–S60.
764. Ziegler, A. & Hebebrand, J. (1998). Sample size calculations for linkage analysis using extreme sib pairs based on segregation analysis with the quantitative phenotype body weight as an example. *Genet Epidemiol*, 15:577–593.
765. Ziegler, A., Kastner, C. & Blettner, M. (1998). The generalised estimating equations: An annotated bibliography. *Biom J*, 40:115–139.
766. Ziegler, A., König, I. R., Deimel, W. et al. (2005). Developmental dyslexia—recurrence risk estimates from a German bi-center study using the single-proband sib pair design. *Human Hered*, 59:136–143.
767. Ziegler, A., König, I. R. & Thompson, J. R. (2008). Biostatistical aspects of genome-wide association studies. *Biom J*, 50:8–28.
768. Ziegler, A., Schäfer, H. & Hebebrand, J. (1997). Risch's lambda values for human obesity estimated from segregation analysis. *Int J Obes Relat Metab Disord*, 21:952–953.
769. Zinn-Justin, A., Ziegler, A. & Abel, L. (2001). Multipoint development of the weighted pairwise correlation (WPC) linkage method for pedigrees of arbitrary size and application to the analysis of breast cancer and alcoholism familial data. *Genet Epidemiol*, 21:40–52.
770. Zöllner, S., Wen, X., Hanchard, N. A. et al. (2004). Evidence for extensive transmission distortion in the human genome. *Am J Hum Genet*, 74:62–72.
771. Zou, G. Y. (2006). Statistical methods for the analysis of genetic association studies. *Ann Hum Genet*, 70:262–276.
772. Zschiedrich, K., König, I. R., Brüggemann, N. et al. (2009). *MDR1* variants and risk of Parkinson disease: Association with pesticide exposure? *J Neurol*, 256:115–120.

Index

- 1000 Genomes Project, 370
- 1 d.f. tests, **201**
- 1-transmission disequilibrium test, 334–335
- Absolute risk reduction, **276**
- ACCE, **383**
- Additive genetic effect, **136**
- Additive inheritance, **30**, 133, 136, 208, 270, 329
- Additive variance
 - for dichotomous traits, 207–**208**
 - for quantitative traits, **136**, 225, 227, 239
- Adoption, 68
- Adoption studies, 141–143
- Affected pedigree member method, 214
- Affected sib-pair, 173, **190**, 242, 244
 - weighting, 213
- Age-dependent penetrance, 31–33
- Age of onset, 32–34, 37, 69, 176, 266, 333
- AIC, see Akaike information criterion
- Akaike information criterion, 150, **313**
- Alliegro, 172, 216
- Allele, **10**, 247–248, 250
- Allele frequency, 38–41, 48, 172, 253, 258, 267, 273
 - effect of Hardy–Weinberg disequilibrium, 85
 - variance, 85
- Allele sharing, 95, **191**
- Allelic heterogeneity, **35**
- Alttest, **69**
- Amino acid, 5, 7, 350
- Amplification refractory mutation system, 55
- Analytical validity, **384**
- Anticipation, **31**, 126
- Antisense strand, 6
- APM, see Affected pedigree member method
- Area under the curve, **386**
- Ascertainment correction, 140
- ASP, see Affected sib-pair
- Association
 - definition, 156, 247–249
 - direct, **248**, 289, 350, 359
 - indirect, **248**–250, 259, 289, 359–360
- Assortative mating, **41**, 77
- Attributable fraction, **277**
 - in the exposed, **278**
- Attributable risk, **276**, 305
 - in the exposed, **278**
 - in variant carriers, **278**
- AUC, see Area under the curve
- Autosomal inheritance, 24–26
- Autosome, 3, 24
- B locus, 302
- Basepair, **2**, 113
- Binning algorithm, 360–361
- Biological interaction, **304**
- Bland-Altman plot, 93
- Bonferroni correction, 181, 374
- Bowker test of symmetry, 331
- Box–Cox transformation, 229
- Broad sense heritability, **137**
 - due to trait locus, **136**
- Brown locus, 302
- C-statistic, **386**

- Calinski–Harabasz Index, 103
- Call fraction, **95**
- Call rate, **95, 97**
- Calling algorithm, 94
- Cardon–Fulker algorithm, 212, 412–414
- Carrier, **26–28**
- Case-control study, 77, **249**, 357, 361
 - effect measures, 274–280
 - multiallelic association, 273
 - sample size calculation, 289–291
 - selection of probands, 266
 - tests for association, 266–274
 - X chromosome, 287–288
- Case-only design, **308**
- CDCV, see Common disease—common variant
- CentiMorgan, 260
- Central dogma of molecular biology, 5, 37
- Centromere, 5
- Chain-termination method, 12
- Chapman–Kolmogorov equations, 411
- χ^2 test
 - expected frequencies, 267
 - for affected sib-pairs, 194, 206
 - for allele frequencies, 268
 - for deviation from HWE, 78
 - for genotype frequencies, 268–270
 - for heterozygotes, 270
 - for population stratification, 294
 - goodness-of-fit, 194
 - in trios, 329
 - observed frequencies, 267
- Chiasma, 9
- Chromatid, 8
- Chromosome, **2–5**, 7–9
- Cis position, **349–350**
- Clark’s algorithm, 352
- Class A model, 147
- Class D model, 147
- Clinical utility, **384**
- Clinical validity, **384**, 386–392
- Cluster algorithm, 93
- Cluster homogeneity, 99
- Cluster plot, 98–109
- Cluster separation, 99
- Cluster separation criterion, **103**
- Cluster stability, 100
- CNV, see Copy number variation
- Coalescence-based algorithm, 352–353
- Cochran’s Q, **382**
- Co-dominant inheritance, 29, 48
- Codon, **5**
- Coefficient of determination, 136
 - for linkage disequilibrium, 253
- Cohort study, **249**, 258
 - effect measures, 274–280
 - multiallelic association, 273
 - sample size calculation, 289–291
 - tests for association, 266–274
 - X chromosome, 287–288
- Combined transmission disequilibrium test, **336**
- Common disease—common variant, **37**, 363, 390
- Common disease, 37
- Compactness, 99
- Complex disease, **36–38**, 118, 175, 223, 266
- Component, **145**
- Component allele, **145**
- Concordance fraction, 70
- Concordance proportion, 70
- Concordant sib-pair, see Sib-pair, concordant
- Confounder, **291**
- Connectedness, 99
- Connectivity, **101**
- Consanguinity, 77, 173, 177
- Contrast, **93**, 100
- Copy number variation, 368, 390
- Coverage, **369**
- Crossing over, **8–9**, 23, 114
- Cross-validation, **316**
- Cryptic relatedness, **95**, 293
- Cycle sequencing, 13
- Davies–Bouldin Index, **103**
- DbSNP, 54
- De Finetti triangle, 40, 42, 86–87, 89
 - for IBD distribution, 206
- Deficiency of heterozygosity, see Genotyping
 - error, false homozygosity
- Deletion, 10
- Deoxyribonucleic acid, **2–7**, 10–11
 - amplification, **14–16**
 - coding, **7**
 - contamination, 95
 - non-coding, 7, 52
 - primer, 12–13, 16, 56
 - replication, 8, 10
 - variation, 10–16
- DerSimonian and Laird approach, **382**
- Design effect, **211**
- Diallelic, **24**, 56
- Dideoxy sequencing, 12
- Dinucleotide, **52**
- Diploid, **3**, 8, 10
- Directional test, **288**
- Discordant sib-pair, see Sib-pair, discordant
- Disequilibrium coefficient, **84**
- Distance, **114**
 - genetic, **113**
 - linkage disequilibrium, **118**
 - map, **115**, 120
 - see also Map distance
 - meiotic, 114
 - physical, **113–114**, 116–117, 260
- DNA, see Deoxyribonucleic acid

- Dogma of molecular biology, see Central dogma of molecular biology
- Dominance genetic effect, **136**
- Dominance variance
 - for dichotomous traits, 200, 207–**208**
 - for quantitative traits, **136**, 225, 227, 229, 239
- Dominant inheritance, **23–25**, 136, 269, 329, 343
- Double helix, **2**, 6
- Double recombinant, **74–76**
- DSL approach, **382**
- DSP, see Sib-pair, discordant
- Dunn Index, **103**
- E locus, **302**
- EC, see Sib-pair, extreme concordant
- ED, see Sib-pair, extreme discordant
- EEA, see Equal environment assumption
- Effect measures, association, see Cohort study, effect measures
- Effect measures, family-based association, see Family-based association
- Effective number of tests, **377**
- Electropherogram, 13–14
- Electrophoresis artefact, **72**
- ELSI, **384**
- Elston–Stewart algorithm, 177, 212, **394**, 396, 398, 400–401
 - anterior, 398
 - posterior, 398
- EMEA, 385
- Empirical heritability, **425**
- Ensembl, 54
- Environment
 - non-shared, **138**
 - shared, **138**
- Epigenetics, **126**
- Epistasis, **36**, **299**
 - see also Interaction
- Equal environment assumption, **142**
- Equivalence test, **86**
 - non-inferiority, 86
 - true equivalence, **86**
- ESP, see Extreme sib-pair
- Ethnicity, 293
- Etiologic fraction, **277**
- Euclidian distance, **100**, 224
- Exact replication, **380**
- Excess fraction, **278**
- Excess heterozygosity, **85**, 97, 269
- Excess homozygosity, **84–85**, **85**
- Excess incidence, **276**
- Exon, 6–7
- Expectation maximization algorithm, 238
 - for haplotype inference, 353–354, 356–357
- Expected LOD score, 180
- Expected maximized LOD score, **180**
- Expressivity, **31**
- Extended transmission disequilibrium test, 332
- Extension locus, **302**
- External validity, **99**
- Extreme concordant sib-pair, see Sib-pair, extreme concordant
- Extreme discordant sib-pair, see Sib-pair, extreme discordant
- Extreme sib-pair, **240**
- Falconer model, 134–141, 232
- False discovery rate, 375
- False positive report probability, 375
- Familial aggregation, 127–146
- Familial correlation, 125, 143–144
- Familial resemblance, 129
- Familiality, 127–131
- Family-based association, 293
 - effect measures, 325–327
 - in pedigrees, 341–344
 - in sibships, 336–341
 - in trios, 319–320, 322, 329, 333, 336
 - missing parental data, 334–335, 338, 340, 343
 - multiallelic association, 330–333, 337, 341, 343, 345
 - quantitative traits, 342, 344–346
 - risk estimation, 325–327
 - sample size calculation, 327–328
 - selection of siblings, 336
- Family-based association test, **342–343**, 345
- Family-wise error fraction, **180**, **374**
- Family-wise error rate, **374**
- Family history method, **127**
- Family study method, **127**
- FastLink, 172
- FastPHASE, 353
- FBAT, see Family-based association test
- FDA, 385
- FDR, see False discovery rate
- FE model, see Fixed effect model
- Fine mapping, 84
- Fixed effect model, **282**, **381–382**
 - inadequacy, **382**
- Fluorescence, 92
- Founder, **166**, 213, 402
- FPRP, see False positive report probability
- Fulker–Cardon algorithm, see Cardon–Fulker algorithm
- Functional study, 125
- FWER, see Family-wise error rate
- Gamete, 7–9
- GC, see Genomic control
- GC-content, **13**
- Gel electrophoresis, 12–14, 56
- Genassoc, 259
- Gene, **7**
 - nomenclature, 11
- Gene-environment interaction, see also Interaction

- GeneFinder, 217
- Gene-gene interaction, *see also* Interaction
- Genehunter, 75, 172, 216
- Generalized linear model, 356–357
- Generalized recurrence risk ratio, **134**
- Genetic code, 5, 7, 11
- Genetic heterogeneity, **35**
- Genetic imprinting, 126
- Genetic marker, **47**
- Genetic model, **30**
 - additive, **30**
 - dominant, **30**
 - multiplicative, **30**
 - recessive, **30**
 - relation to IBD probabilities, 208
 - selection, 280–287
- Genetic variance, **136**
- Genetics, 17
- Genome-wide association, 315, **367**
 - meta-analysis, 380–383
 - p-value, 374–378
 - significance level, 374–378
 - standard quality control, 94–98
 - workflow, 368
- Genome-wide association study, 61, 67–68, 91, 126, 296–298, **367**
- Genome-wide scan, 71, 122, 174, 180–182, 210–212
- Genomic control, 94, 295–297
- Genomic coverage, **369**
- Genotype, 7, 10, 250
- Genotype calling, 94, 98
- Genotype calling algorithm, 94, 98
- Genotype frequency, 38–42, 163, 267–268, 273
- Genotype stability, 100
- Genotypic relative risk, **30**, 34–35, 325–326, 328
- Genotypic similarity, 190, 224
- Genotyping, **57**
- Genotyping error, **71**, 172, 175
 - false homozygosity, 72, 75, 81
 - family-based study, 70
 - frequency, 70
 - human cause, 72
 - null allele, 71, 82
 - population-based study, **77**
 - stutter band, 72
 - technological factor, **71**
- Genotyping technology, 57–63
 - chip-based, **61**
 - MALDI-TOF, **61**
 - MassARRAY iPLEX, **61**
 - RFLP, **58**
 - RT-PCR, **58**
- GLM, *see* Generalized linear model
- GoId, 122, 259
- GRR, *see* Genotypic relative risk
- GWA study, *see* Genome-wide association study
- H^2 , h^2 , *see* Heritability
- Hadamard product, 407
- Hamming distance, 407
- Hapinfern, 352
- Haplo.stat, 355, 357–358
- Haploid, **3**, 7–8
- Haplotype, 40, 157, 261, **349**, 359
 - advantages, 350–351
 - ambiguity, 250–**251**, 251, 351
 - block, 122, **261**–262, 353, 359
 - inference, 251, 351–356
 - tests for association, 351, 356–358
- Haplotype relative risk, 321–322, 325, 327
- Haploview, 259
- HapMap project, 360
- Hardy–Weinberg
 - disequilibrium, 84
 - effect on allele frequency, 85
 - equilibrium, 16, **38**–43, 48, 68, 77–93, 135, 145, 163, 225, 273, 326, 352
 - compatibility with, 86–91
 - measures of deviation, 83–85
 - law, **38**–43
- Haseman–Elston, **225**
 - empirical *p*-value, 243
 - model, 225
 - new, **232**–**233**
 - power comparison, 233
 - revisited, 231
 - sample size calculation, 234–237
 - standard, **225**
 - two-level, 234
- HE, *see* Haseman–Elston
- Heat map, 122
- Hemizygous, **27**–28
- HEr, *see* Haseman–Elston, revisited
- Heritability, 131, 134–147
 - broad sense, **137**
 - broad sense due to trait locus, **136**
 - narrow sense, **137**
 - narrow sense due to trait locus, **136**
- Het, 49
- Heterogeneity, **35**
- Heterogeneity LOD score, *see* LOD score, heterogeneity
- Heterozygosity, **48**, 51, 95, 97, 211–212
 - relative excess, **85**
- Heterozygote deficiency, **85**, **97**, 269
- Heterozygous, **10**, 23, 29–30
- Hidden Markov model, 407
- High allele, **136**
- Hit fraction, **316**
- Homogeneity of clusters, 99
- Homologous, **8**–9
- Homozygous, **10**, 23, 29–30

- HRR, see Haplotype relative risk
- Hubert's Index
 - modified, **103**
 - standardized modified, **104**
- HWG-Test, 80
- HWE, see Hardy-Weinberg, equilibrium
- HWL, see Hardy-Weinberg, law
- IBD, see identical by descent
- IBS, see identical by state
- Identical by descent, 83, 95, **139**, **190**, 206, 210, 224, 228, 238, 243, 412
- Identical by state, 83, 95, **190**, 206
- Imprinting, 6, **33**–35, 176
- Imputation, 370
- Impute, 370
- In-silico replication, **378**
- Inbreeding, 77, 83
- Inbreeding coefficient, **83**, 273, 298
 - within population, **83**
- Incidence, **274**
- Index proband, **21**
- Informative-transmission disequilibrium test, 341
- Informativity, 53, 56, 210–211, 351
 - measures of, 48–52
- Inheritance distribution, 402
 - graph theoretical approach, 403–405
 - multi-marker, 407
 - single-marker, 404
- Inheritance pattern, **24**, 29–30, 36, 172
- Inheritance vector, **213**, **401**–403
- Insertion, 10
- Interaction, **303**–306
 - statistical tests, 307–314
- Interaction in public health, 305
- Inter-cluster distance, **102**
 - average, **102**
 - minimum, **102**
- Interference, **74**, 116, 174, 181
 - positive, **74**, 116
- Intermediate phenotype, 37, **222**–223
- Internal control, 320
- Internal validity, **99**
- International Organization for Standardization, 70
- Intra-cluster variance, 101
- Intron, 6–7
- Inverse variance weighting, **382**
- ISO, 70
- Jackpot effect, **379**
- Jacquard's Δ_7 coefficient, **139**, 208
- Karyogram, 4
- Karyotype, **5**
- Kb, see Kilobase
- Kernel of likelihood, see Likelihood, kernel
- Kilobase, **113**
- Kinship coefficient, 138–141, 208
- L-pop, 295
- Lander-Green algorithm, 177, 212, 244, **401**
- Large sibships, 213–214, 230–231
- LD, see Linkage disequilibrium
- LdSelect, 361
- Lead-bias, 33
- Liability class, 176
- LIC, see Linkage information content
- Likelihood ratio test, 158–159, 194, 237, 240
- Likelihood
 - kernel, **158**, 160–161, 239
- Linkage, 23, **156**, 248, 325
 - a posteriori probability for, **184**
 - a priori odds for, 183
 - forward odds for, 183
 - hypothesis, 158
 - phase, **157**
 - phase unknown, **160**
 - significant, 180–184
- Linkage, 172
- Linkage analysis, 125
 - assumptions, 172–173
 - causal model, **168**
 - counting recombinants and non-recombinants, 156
 - disease definition, 168
 - loops, 177
 - model-free
 - p-value, 243–244
 - model-based, **167**, 207
 - model-free, 175, **190**–**191**
 - model-free including unaffected individuals, 218
 - model misspecification, 172–173
 - multipoint, **174**–175, 212–213, 217, 407
 - non-parametric, **191**
 - optimal model-free, 206
 - phase-known pedigrees, 156–160
 - phase-unknown pedigrees, 160–161
 - power considerations, 177–180
 - quantitative trait, 177
 - single-marker, **169**
 - two-locus model, 176
 - two-point, **168**
 - with missing genotypes, **161**
 - X chromosome, 177
- Linkage disequilibrium, 173, **248**–**250**, 289, 325, 327–328, 353–354, 359, 369
 - allelic, 119, **250**–255
 - complete, **253**, 258
 - composite, **256**
 - covariance, 119, 252
 - decay, 259–260, 262
 - digenic, **256**
 - extent, 259–262
 - genotypic, 250, **255**–259
 - map, **118**

- measures, 250, 255
- probability, **119**
- Linkage disequilibrium unit, 118–123
- Linkage equilibrium, see Linkage disequilibrium
- Linkage information content, **51**, 211
- Locus, **10**, 47, 248
- Locus heterogeneity, **35**
- Locus homogeneity, 176
- LOD score, **159**
 - additivity, 160
 - at θ , 158
 - expected, **180**
 - expected maximized, 180
 - function, **158**
 - heterogeneity, **176**, 216
 - maximized, **180**
 - table, **165**
- LRT, see Likelihood ratio test
- MA-plot, 93
- Mach, 371
- Machine learning, **315**
- MAF, see Minor allele frequency
- MALDI-TOF, see Matrix-assisted desorption/ionization time-of-flight
- Malecot equation, **121**
- Malecot model, **120**, 254
- Map distance, **113–114**, 114–118, 120
 - meiotic, 114
 - pseudoautosomal region, 116
 - X chromosome, 117
 - Y chromosome, 117
- Map function, 115–116, 119
 - Felsenstein, **116**
 - Haldane, **115**, 118
 - Kosambi, **116**, 118
 - Morgan, **116**
- Map length, 114
- Marginal homogeneity test, 331
- Markov chain
 - inhomogeneous, 406
- MassARRAY iPLEX, 61–62
- Matrix assisted laser desorption/ionization time of flight, 56, **61–62**
- Maximized LOD score, **180**
- Maximum likelihood, 157, 221
- Maximum likelihood binomial test, 206, 214
 - for quantitative traits, 221, 231
- Maximum likelihood ratio test, **159**
- Maximum LOD score, **194**, **207**, 214
 - no dominance variance, **200**
- Mb, see Megabase
- McNemar test
 - for transmission frequencies, **324**
 - with multiple alleles, 331
- MDR, see Multifactor dimensionality reduction
- Mean test, **202**, 206
 - power, see Power calculation, for mean test
- Megabase, **113**
- Meiosis, 7–10
- Mendel's laws, 22–23, 156, 248
- Mendel check, 72–76
- Mendel, 75
- Mendelian error, **69**
- Merlin, 172, 216
- Meta-analysis, see Genome-wide association, meta-analysis
- Methylation, 6, 34
- Microarray, 67
- Microsatellite marker, see Short tandem repeat
- Midparent value, **154**
- MiF, see Missing frequency
- Migration, **41**
- Minimax test, **202–203**, 207
- Minor allele frequency, 54, 88, **98**
- Minsage, 17
- Missing frequency, **97**
- Missing genotypes, 94
- ML, see Maximum likelihood
- MLB, see Maximum likelihood binomial test
- MLbGH, 216
- MLBQT, see Maximum likelihood binomial test, for quantitative traits
- MLS, see Maximum LOD score
- Mode of inheritance, see Inheritance pattern
- Monogenic, **29**, 35
- Monomorphic SNP, see Single nucleotide polymorphism, monomorphic
- Mononucleotide, **52**
- Morgan, 115
- Multifactor dimensionality reduction, **315–316**
- Multilocal, **29**
- Multilocus feasibility, 117–**118**, 118, 411
- Multiplicative inheritance, **30**, 132, 270
- Multipoint analysis, see Linkage analysis, multipoint
- Multistage design, **378**
- Mutation, **10–11**, 41–42, 259, 350
 - frameshift, 10–11
 - frequency, 11, 48, 53, 55
 - nomenclature, 11
 - non-synonymous, 11
 - point, 10
 - probability to detect, 16
 - rate, 72, 172
 - synonymous, 11
- Narrow sense heritability, **137**
 - due to trait locus, **136**
- Negative predictive value, **385**
- Neutral marker, see Null marker
- No dominance variance, **136**
- Nomenclature
 - for genes, 11

- for genomic DNA, 11
- for mutations, 11
- Non-inferiority, **86**
- Non-parametric linkage test, 206, 214
 - based on affected pairs, **215**
 - based on all affected individuals, **215**
 - for quantitative traits, 221
- Non-paternity, 68, 243
- Non-shared environment, **138**
- NPL test, see Non-parametric linkage test
- Nucleoside, **2**
- Nucleotide, **2**
- Null allele, **71**
- Null marker, 293–295, 297
- Odds ratio, **262**, 274, 278
 - for allele frequencies, 268, 289
 - for linkage disequilibrium, 254, **262**
 - for parental haplotypes, **321**
- Oligogenic, **29**, 36
- Oligogenic model, **137**
- OMIM, see Online Mendelian Inheritance in Man
- Online Mendelian Inheritance in Man, 24
- Optimal test, see Linkage analysis, optimal
 - model-free
- OR, see Odds ratio
- Ornstein–Uhlenbeck process, 181
- Overdominance, **136**
- Overfitting, **316**
- Panmixia, 41
- PAF, see Population attributable fraction
- PAR, see Population attributable risk
- Parent of origin effect, see Imprinting
- Parental effects, 34
- Paucilocal, **29**
- PCA, see Principal components analysis
- PCR, see Polymerase chain reaction
- PDT, see Pedigree disequilibrium test
- PedCheck, **73–74**
- Pedigree, 21–22
- Pedigree disequilibrium test, **341–342**
- Pedigree error, **68–69**, 172
- Peeling, **396**, 398–399
- Penetrance, **29–36**, 132
- Penetrance function, 146, 168, 172
- Pharmacogenetics, 389
- Phase, see Linkage, phase
- PHASE, 353
- Phenocopy, **29–30**
- Phenotype, **7**, 172, 247–248
- Phenotypic heterogeneity, **35**
- Phenotypic similarity, 190, 224–225, 227, 230–232, 241
- PIC, see Polymorphism information content
- Pivot element, **399**
- Pleiotropy, **35**
- Polygenic component, **137**, 225, 238
- Polygenic effect, **29**, 36
 - additive, **137**
 - dominance, **137**
- Polymerase, 6, 14, 16
- Polymerase chain reaction, **14–16**, 53, 55
 - real-time, **58**
- Polymorphic, **48**, 53
- Polymorphism, **11**
- Polymorphism information content, **49–51**, 210–211
- Pooled analysis, 381
- Population attributable fraction, **277**
- Population attributable risk, **277**
 - for genotype frequencies, 326
 - for linkage disequilibrium, 254
- Population prevalence, 132
- Population stratification, 41–42, 77, **291–298**, 335
 - genomic control, see Genomic control
 - principal components, see Principal components analysis
 - structured association, see Structured association
 - test for, 293–295
- Positional mapping, **118**, 175
- Positive predictive value, **385**
- Possible triangle, **198–199**
- Possible triangle test, **198**
- Power calculation
 - for Haseman–Elston, 234–237
 - for mean test, 211
 - for model-based linkage analysis, 179
 - for variance components approach, 234
- Power of a test, **210**
- Predictive genetic test, 383
- Predictive value, **385**
- Prest, **69**
- Prevalence factor, **132**
- Principal components analysis, 297–298, 361
- Principle of similarity, **190**, 224, 241–242
- Profile-a test, 183
- Promoter, 6
- Proportion of missing genotypes, **94**
- Proportion test, **202**, 206
- Protein, 5, 7
- Pseudoautosomal region, **4**, 116
- Pseudo-control, 320
- Pter, **5**
- Q-Q plot, see Quantile-quantile plot
- Qter, **5**
- Quantile-quantile plot, **94**, **296**
- Quantitative trait, additive variance, see Additive variance
- Quantitative trait
 - advantages, 222–223
 - disadvantages, 222–223
 - dominance variance, see Dominance variance

- intermediate phenotype, **222–223**
- model-based linkage analysis, **177**
- model-free linkage analysis, **222**
- R^2 , see Coefficient of determination
- Random effect model, **381**, **282**
- Random mating, **41–42**, **50**, **135**, **142**, **225**
- Random sib-pair, **240**
- RC-TDT, see Reconstruction combined transmission disequilibrium test
- RE model, see Random effect model
- Reading frame, **5**
- Real-time polymerase chain reaction, **58–60**
- Receiver operating characteristic, **386**
- Recessive inheritance, **23–25**, **136**, **269**, **329**, **343**
- Recombination, **9–10**, **23**
 - cold spot, **116**
 - hot spot, **116**, **122**, **261**, **359**
- Recombination fraction, **10**, **70**, **114**, **120**, **174**, **217**, **259–261**, **322**, **412**
 - estimation, **157**
 - relationship to identical by descent, **192**
 - sex-specific, **175**, **400**
- Reconstruction combined transmission disequilibrium test, **335**, **338–340**
- Reconstruction of genotypes, **335**, **338**
- Recurrence risk ratio, **125**, **127–130**, **132**, **207–209**, **275**
 - factor, **133**
- Reference SNP number, **54**
- Regional replication, **380**
- Regressive model, **146**
- Relationship estimation, **69**
- Relative excess heterozygosity, **85**
- Relative risk, **274–275**, **279**, **321**
 - for genotype frequencies, **329**
 - for parental haplotypes, **321**
 - genotypic relative risk, see Genotypic relative risk
- Repeatability error, **70**
- Replication, **301**, **379–380**
 - exact, **380**
 - regional, **380**
- Reproducibility error, **70**
- Residual correlation, **225**
- Restriction fragment length polymorphism, **58**
- RFLP, see Restriction fragment length polymorphism
- Ribonucleic acid, **2–7**
- Ribosome, **7**
- Risch's λ -value, see Recurrence risk ratio
- Risk, **274**
- Risk difference, **276**
- Risk estimation, **125**
- Risk ratio, see Recurrence risk ratio
- RNA, see Ribonucleic acid
- Robust transmission disequilibrium test, **335**
- ROC curve, see Receiver operating characteristic
- Root mean square distance, **101**
- RR, see Relative risk
- Rs-number, **54**
- RSP, see Random sib-pair
- RT-PCR, see Real-time polymerase chain reaction
- SA, see Structured association
- S.A.G.E., **172**, **216**, **231**
- Sample size calculation
 - for association in unrelated individuals, **289–291**
 - for Hardy–Weinberg equivalence test, **89**
 - for Haseman–Elston, **234–237**
 - for number of permutations, **79**
 - for variance components approach, **234**, **238**
 - in trios, **327–328**
- Sample swap, **68**
- Sanger method, **12**
- SAO, see Single affected offspring
- SAP, see Statistical analysis plan
- Score function, **443–444**
- Score test, **201**, **203**, **213–214**
 - linkage disequilibrium, **252**
 - model-free linkage, **201**
- SDT, see Sibship disequilibrium test
- Secular trend, **143**
- Segregation analysis, **125**, **144–154**, **173**
- Selection, **41**, **77–78**, **81**, **97**, **260**, **269**
- Selection of markers, **359–364**
 - based on linkage disequilibrium, **360–361**
 - evaluation, **363–364**
- Sense strand, **6**
- Sensitivity, **385**
- Separability, **99**
- Separation of clusters, **99**
- Sequencing, **12–14**
- Sequential probability ratio test, **180**
- Sex chromosome, **3–4**, **26**, **117**
- Shared environment, **138**
- Shared environmental effect, **225**
- Short tandem repeat, **52–54**, **71**, **81**, **174**, **273**, **330–331**
 - probability to detect, **16**
- Sib-pair
 - ascertainment, **240–243**
 - concordant, **190**, **218**
 - discordant, **190**, **218**, **242**, **336**
 - extreme, **240**
 - extreme concordant, **242**
 - extreme discordant, **242**
 - random, **240**
 - single proband, **231**, **240**
- Sibship disequilibrium test, **336–337**
- Sib transmission disequilibrium test, **336–338**
- Signal intensity, **91–92**
- Signal intensity plot, **98–109**
- Significance level, **180–184**, **211**, **243**

- in genome-wide association, 374–378
- Silhouette Index, **104**
- Simlink, **179**
- SimWalk2, 75
- Single affected offspring, **322**
- Single-marker analysis, **169**
- Single nucleotide polymorphism, 16, **54–57**, 71, 78–86, 175, 203, 210, 360
 - coding, **55**, 359
 - haplotype tagging, **360**
 - monomorphic, **95**
 - non-coding, 55
 - probability to detect, 16
 - reference number, **54**
 - rs-number, **54**
 - ss-number, **54**
 - submitted number, **54**
 - tagging, **360**
- Single proband sib-pair, 231, **240**
- SLink, 179
- SNP, see Single nucleotide polymorphism
- SNPHAP, 355
- Specificity, **385**
- Spillover, 412
- Splicing variant, 6
- SPSP, see Single proband sib-pair
- Ss-number, **54**
- S-TDT, see Sib transmission disequilibrium test
- Stability of clusters, 100
- Stability of genotypes, 100
- Standardized effect size, **211**
- Statistical analysis plan, 381
- Statistical interaction, 303
- STR, see Short tandem repeat
- Strat, 295
- Structure, 294
- Structured association, 294–295, 297
- Stutter band, **72**
- Submitted SNP number, **54**
- SuperLink, 172
- Sushi, 291–292
- Swept radius, **121**
- tagSNP, see Single nucleotide polymorphism, tagging
- TaqMan, 58–60, 104
- Telomere, 5, 260
- Test data set, **316**
- Testing accuracy, **316**
- Tetranucleotide, **52**, 330
- Training data set, **316**
- Transcription, **5–7**
- Transition, 10
- Transition probability, **145**, 395, 400, 407
- Translation, **7**
- Transmission disequilibrium test, **322–325**, 327–329, 333–334
- Transmission probability, **146**, 322–323, 328, 331, 400
- Trans position, **349**
- Transversion, 10
- Trend test, **270**, 289
- Triangle restriction, **198**
- Triangle test, **198–199**
- Trinucleotide, **52**, 330
- Trio design, see Family-based association
- Triplet, **5**
- TTS, see Triangle test statistic
- Twin studies, 141–142
- Two-point analysis, see Linkage analysis, two-point
- Type of an individual, **144**
- Underdominance, **136**
- Unilocal, **29**
- Utility
 - clinical, **384**
- Validation, **380**
- Validity
 - analytical, **384**
 - clinical, **384**, 386–392
- Variable number of tandem repeat, **52**
- Variance components approach, 221, 227, 233–234, 237–240
 - for association analysis, 344
 - identifiability of parameters, 239
 - model, 134–141
 - power calculation, 234
 - sample size calculation, 234, 238
- Vasarely, 81
- Vitesse, 172
- VNTR, see Variable number of tandem repeat
- Wahlund effect, 77
- Wald test, **201**, 203, 213
 - for linkage disequilibrium, 252
 - for model-free linkage, 201
- Weighted least squares, 230
- Weighted pairwise correlation test, 221
- Whole-genome genotyping, 58
- Whole-genome sequencing, 390
- Wilcoxon signed rank test, 221
- Winner's curse, 379
- Winsorization, **229**
- X-chromosomal inheritance, 26–28
- Y-chromosomal inheritance, 28