STATISTICAL GENETICS

Gene Mapping Through Linkage and Association

Edited by
Benjamin M Neale
Manuel AR Ferreira
Sarah E Medland
Danielle Posthuma

Statistical Genetics: Gene Mapping through Linkage and Association

Edited by:

Benjamin M Neale

Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, UK
Broad Institute of MIT and Harvard University, USA
Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, USA

Manuel AR Ferreira

Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, USA
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Sarah E Medland

Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, USA
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Danielle Posthuma

Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands

http://www.genemapping.org

Taylor & Francis
Taylor & Francis Group

Published by:

Taylor & Francis Group

In US: 270 Madison Avenue
New York, N Y 10016
In UK: 2 Park Square, Milton Park
Abingdon, OX14 4RN

ISBN: 9780415410403

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

Library of Congress Cataloging-in-Publication Data

Statistical genetics: gene mapping through linkage and association / edited by Benjamin M. Neale . . . [et al.].
p. ; cm.
Includes bibligraphical references and index.
ISBN 978-0-415-41040-3 (alk. paper)
1. Gene mapping—Statistical methods—Congresses. 2. Linkage (Genetics)—Statistical methods—Congresses. I. Neale, Benjamin M.
[DNLM: 1. Chromosome Mapping—methods—Congresses. 2. Models, Statistical—Congresses. 3. Linkage (Genetics)—Congresses. QU 450 S797 2008]

QH445.2.S73 2008
572.8'633—dc22

2007042274

Editor: Elizabeth Owen
Editorial Assistant: Kirsty Lyons
Senior Production Editor: Simon Hill
Typeset by: Keyword Typesetting
Printed by: Cromwell Press

Printed on acid-free paper

10 9 8 7 6 5 4 3 2 1

Cover illustration credit: "DNA" ©iStockphoto.com/Scot Spencer

T&F informa

Visit our web site at http://www.garlandscience.com

Contents

ASSOCIATION ANALYSIS

About the editors

Benjamin M. Neale graduated with a BSc in psychiatric genetics from Virginia Commonwealth University in 2006. During his undergraduate education, he worked closely with Drs Patrick Sullivan, Cynthia Bulik and Kenneth Kendler on structural equation modeling and linkage analysis of psychiatric traits. In 2004, he moved to the Institute of Psychiatry (IOP) to work with Dr Pak Sham on developing methods for association analysis. While at the IOP, he began work with Dr Philip Asherson on the genetics of ADHD, with a focus on association analysis. Currently, he is visiting with Dr Mark Daly at the Center for Human Genetic Research at Massachusetts General Hospital and the Broad Institute of MIT and Harvard. The main focus of this work is developing association methodology especially with respect to genome-wide association.

Manuel A. R. Ferreira graduated in Biological Sciences from the University of Lisbon, Portugal, in 2000. His undergraduate work focused on behavioral ecology, particularly while a visiting student at the University of Oxford, UK. In 2001, he moved to Brisbane, Australia, to pursue a PhD in Human Genetics with Prof. Nick Martin and Dr David Duffy at the Queensland Institute of Medical Research and School of Medicine, University of Queensland. His thesis focused on the identification of genetic risk factors for asthma through linkage analysis. He has received a number of academic awards, including the Lodewijk-Sandkuijl 2004 award from the European Society of Human Genetics. In 2006, he was awarded a post-doctoral Sidney Sax fellowship from the National Health and Medical Research Council, Australia, to join Dr Shaun Purcell's group at the Center for Human Genetic Research, Harvard Medical School, Boston, where he currently is involved in the development of methods for the analysis of genome-wide association studies.

Sarah E. Medland graduated with a BA (Hons) in Neuropsychology from the University of Queensland, Australia, in 2000. Her undergraduate work focused on the lateralization of language centers within the brain. She undertook her PhD studies at the Genetic Epidemiology Unit of the Queensland Institute of Medical Research under the supervision of Dr David Duffy, Dr Margie Wright and Prof. Gina Geffen. During her PhD studies she also worked extensively with Prof. Nick Martin. Her PhD research

focused on the genetic epidemiology of behavioral laterality. However, she works on many topics including methodology development, substance abuse, political and social attitudes, infant development and obesity. In 2006, she was awarded a post-doctoral Sidney Sax fellowship from the National Health and Medical Research Council, Australia, to join Dr Mike Neale's group at the Virginia Institute for Psychiatric and Behavioral Genetics.

Danielle Posthuma currently works as an associate professor in the Department of Biological Psychology at the VU University, Amsterdam, The Netherlands. She received a triple MSc degree (Hons) in 1996, received her PhD in Behavior Genetics Cum Laude in 2002, and is the recipient of several national and international honors and awards, such as the Fuller and Scott award for early career outstanding achievements from the Behavior Genetics Association. She is PI on projects on gene-finding for cognition and the development of statistical genetic methods for GxE, scientific director of the Genetic Cluster Computer project dedicated to genome-wide analysis, and co-director of the annual European workshop on '*Methodology for Gene Finding and Genetic Epidemiology*'.

List of contributors

Gonçalo R. Abecasis
Center for Statistical Genetics, Dept. of Biostatistics, University of Michigan, USA

David Bender
Broad Institute of MIT and Harvard University, USA
Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, USA

Stéphanie M. van den Berg
Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands

Dorret I. Boomsma
Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands

Stacey S. Cherny
Department of Psychiatry and Genome Research Centre, The University of Hong Kong, China

Danielle M. Dick
Departments of Psychiatry and Psychology, Washington University in St. Louis, USA

Conor V. Dolan
Department of Psychology, University of Amsterdam, The Netherlands

David M. Evans
Wellcome Trust Centre for Human Genetics, University of Oxford, UK

Jesen Fagerness
Broad Institute of MIT and Harvard University, USA
Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, USA

Manuel A.R. Ferreira
Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, USA
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Jouke Jan Hottenga
Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands

Jo Knight
Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, UK

Johnny S.H. Kwan
Genome Research Centre, The University of Hong Kong, China

Hermine H. Maes
Virginia Institute of Psychiatric and Behavioral Genetics, Department Human Genetics and Massey Cancer Center, Virginia Commonwealth University, USA

Julian B. Maller
Broad Institute of MIT and Harvard University, USA
Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, USA

Nicholas G. Martin
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Sarah E. Medland
Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, USA
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Andrew P. Morris
Wellcome Trust Centre for Human Genetics, University of Oxford, UK

Benjamin M. Neale
Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, UK
Broad Institute of MIT and Harvard University, USA
Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, USA

Michael C. Neale
Virginia Institute of Psychiatric and Behavioral Genetics, Departments of Human Genetics, Psychiatry and Psychology, Virginia
Commonwealth University, USA
Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands

Dale R. Nyholt
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Danielle Posthuma
Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands

Shaun Purcell
Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Harvard Medical School, USA
Broad Institute of MIT and Harvard University, USA

Frühling V. Rijsdijk
Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, UK

Pak C. Sham
Department of Psychiatry and Genome Research Centre, The University of Hong Kong, China

Sophie van der Sluis
Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands

Patrick F. Sullivan
Department of Genetics, University of North Carolina, USA

Peter M. Visscher
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Naomi R. Wray
Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Australia

Preface

The idea for this book occurred in March 2005 during a faculty meeting at the Eighteenth International Workshop of Twin and Family Studies, held in Boulder, Colorado, US. Many students attending the workshop enquired about a handbook that collected the presented materials on the genetic linkage and association analyses of human complex traits. The faculty, as a group, decided such a project would aid in the effectiveness of the workshop, and we agreed to take this project on.

When we took up this challenge, our aim was to provide a handbook that would help researchers interested in human gene mapping to navigate through the challenging fields of linkage and association analysis. The format we chose to achieve this was to invite leading researchers in these fields, many of whom are faculty members on this Workshop, to contribute general theoretical chapters and/or practical, hands-on sections. We hope that the former will provide a reasonable overview of the theoretical foundations fundamental to gene mapping, and the latter a glimpse of how genetic linkage and association analyses are conducted in practice, what kind of obstacles are likely to be encountered and how these can be resolved. Example scripts and data files used in these tutorials are available via the book's homepage at http://www.genemapping.org.

Most of the chapters in this book were contributed by researchers who authored commonly used genetic software programs. Therefore, the software discussed throughout this book represents only a limited number of the many excellent packages available. Obviously, the credits of this book should go to all the authors that willingly agreed to embark on this project. Thank you all, it was a pleasure to work with you.

The Editors
April 3rd, 2007, Amsterdam/Boston/Richmond

Acknowledgements

This book would not have been produced without the support and input of many of our colleagues. First of all we would like to thank the faculty members of the Boulder and Egmond workshops, many of whom contributed chapters to this book and helped crystallize its contents, especially John Hewitt, Nick Martin, Dorret Boomsma, Mike Neale, Lon Cardon and Hermine Maes. In addition, we are indebted to all those who have dedicated both a significant amount of time and effort to carefully read and comment on the chapters at various stages of the editorial process: Gonçalo Abecasis, Leo Beem, David Bender, Dorret Boomsma, Lon Cardon, Heather Cordell, Christopher Cotsapas, Nancy Cox, Mark Daly, Conor Dolan, Frank Dudbridge, David Evans, Charles Gardner, John Hewitt, Peter Holmans, Jeanine Houwing-Duistermaat, Mike Kearsey, Christoph Lange, Penelope Lind, Stuart MacGregor, Matt McQueen, Mike Neale, Hein Putter, Brien Riley, Eric Schmitt, and Peter Visscher.

We would also like to express our appreciation for the efforts made by our publishing editors, Kirsty Lyons and Simon Hill at the Taylor and Francis Group, who was always very supportive and patient with our inexperience in these matters throughout the last two years.

Foreword

This book is the fruit of the Workshops on Methodology for Twin and Family Studies, the first of which was held in Leuven, Belgium in 1987 (hosted by Robert Derom), and which have been held over twenty times since. The workshops were sparked by the enthusiasm of an international network of researchers dedicated to developing and applying structural equation modeling methods in the burgeoning fields of behavior genetics and genetic epidemiology. At that time the power of these methods was becoming known, but there were few places where they were taught formally and there was a considerable demand from students in a wide variety of disciplines in many countries for venues where they could learn them. It seemed a good idea to gather the leading exponents of the emerging subject together in congenial surroundings in the hope of imparting the arcane art to a wider audience. Leuven, then later Boulder, Colorado, Helsinki and most recently Egmond on the Dutch coast have provided ideal venues for the informal but intensive training for which the workshops have become known, and at last count, over 800 'participants' had benefited from the 1-week hothouse of SEM, matrix algebra, psychometrics and biometrical genetics which characterizes the workshop.

The workshops were an immediate success, both didactically and in stimulating research applications and development of new methods. A special issue of *Behavior Genetics*, edited by Nick Martin, Dorret Boomsma and Michael Neale (January 1989), was published with ten papers arising from the first workshop, setting out the general framework for SEM in the genetic context, dealing with practical issues such as how to model age regression and sex effects, and showing innovative applications to extended twin designs and the first glimmering of linkage analysis using twins.

The 1991 workshop, organized by Hermine Maes in Leuven (who had been a student at the first workshop), was funded by NATO as an Advanced Studies Workshop, which required a publication of the proceedings. This stimulated compilation of a book, which was roughed-out by participating faculty over the weekend between the introductory and advanced weeks of the workshop, but was then fleshed-out and polished into the superb 'Neale and Cardon' (M.C. Neale and L.R. Cardon (1992) *Methodology for*

Genetic Studies of Twins and Families, NATO ASI Series D: Behavioral and social. Kluwer Academic, Dordrecht, The Netherlands) which has been cited over 1000 times. It is worth mentioning that used photocopied versions of Neale and Cardon now sell for $200 US on eBay and the hard cover has sold for over $1000 US on Amazon, a market testimony to its value!

Neale and Cardon was written around the LISREL software program on which the course was based for the first five years or so. Since 1990 however, Michael Neale has been developing his own program, Mx, based conceptually on LISREL but purpose-written and flexible to suit genetic applications, and this has now become the standard engine used by behavior geneticists and genetic epidemiologists. Mike, and his wife Hermine Maes, are close to completing a revision of Neale and Cardon (already known as Neale and Maes), adapting it to the Mx language.

In 1990 John Hewitt (who in 2001 became director of the Institute of Behavior Genetics at the University of Colorado in Boulder) obtained an NIMH training grant to fund the workshop in Boulder. Since then it has been held in Boulder fifteen times, usually in March. Beginning in 1999 there has been the regular alternation of the Introductory (even numbered years) and Advanced (odd numbered years) workshops. Both types of workshops have typically been oversubscribed, with an average of 55 students per workshop. Some of the 'students' are in fact senior researchers (including NIH directors) who attend to update on advances in the field. One key feature of the workshop, the hands-on nature of the workshops with theoretical presentations usually followed by a practical application, has been critical to the workshops' success, and is often cited as setting these workshops apart from other similar workshops offered. Given the success of the workshops, their popularity, and the need to train people outside the US, this model has also been emulated by a parallel European series of workshops, held in Egmond-aan-Zee in the Netherlands in September since 2003, organized by Danielle Posthuma who has gained supporting funding from the Netherlands Scientific Organization and the genomEUtwin project.

The syllabus of the Introductory workshop, biometrical genetics and SEM in a genetic context, is covered well in Neale and Cardon (or Neale and Maes as it will soon become). However, with the technical breakthroughs of high-throughput microsatellite typing, and now SNP typing, have come parallel developments in statistical genetic methods for linkage and association mapping of genes for complex traits. Through advances made in methods of analysis of large-scale phenotypic and genotypic data (and the continuous

dissemination of these methods through quickly evolving workshops), enormous gains have been made in our understanding of the etiology of a wide range of complex phenotypes. The impact of these statistical methods (and the associated training at annual workshops) can easily be judged in the rapid expanse of knowledge evidenced in thousands of publications. As an example, the Mx software, which is primarily used in genetic studies, has been cited over 1250 times since 1991 and the number of publications continues to grow every year in a range of research fields. Similarly, the Merlin software (Abecasis *et al.* 2002) used for linkage and association analyses has been cited 547 times since 2002. Many of the faculty and students who have participated in these workshops have applied the methods to produce published articles; some have taught them to their colleagues. Thus the workshops have had considerable influence in setting the research agenda in universities around the world. New leaders in the field have emerged from the ranks of students and junior faculty who have participated.

It is this interdisciplinary domain of molecular biology, statistical and population genetics, and SEM that has become the common ground of the Advanced course. Particularly gratifying has been the number of very bright younger scientists drawn to this field whose presentations to the course have set a formidable and inspiring standard. The feeling that what is being offered in the Advanced course is a unique and valuable 'take' on the subject is what has led to this volume. Each chapter is written by a different member of the faculty of the advanced course, but the volume is shaped and edited by four of the youngest and keenest. We hope that it will find a niche and serve a purpose as admirably as its predecessor, Neale and Cardon, has done.

Nicholas G. Martin, Queensland Institute of Medical Research, Brisbane
Dorret I. Boomsma, Vrije Universiteit, Amsterdam
Michael C. Neale, Virginia Commonwealth University, Richmond
Hermine H. Maes, Virginia Commonwealth University, Richmond

March 2007, Boulder, Colorado

Abbreviations

A	additive genetic variance component
ALS	amyotrophic lateral sclerosis
C	shared environmental variance component
CDCV	common disease common variant
CEPH	Centre d'Etude du Polymorphisme Humain collection
CNV	copy number variation
DNA	deoxyribonucleic acid
E	nonshared environmental variance component
EM	expectation maximization
FDR	false discovery rate
GAIN	Genetic Association Information Network
GPC	genetics power calculator
GWAS	genome-wide association study
G×E	gene-environment interaction
HapMap	haplotype map
HE	Haseman-Elston
HLOD	heterogeneity LOD
HRRT	haplotype relative risk test
HWE	Hardy-Weinberg equilibrium
IBD	identity-by-descent
IBS	identity-by-state
LD	linkage disequilibrium
LOD	log of odds
LRT	likelihood-ratio test
LSE	least squares estimation
MAF	minor allele frequency
MAR	missing at random
MCMC	Markov-Chain Monte-Carlo
MDS	multidimensional scaling
ML	maximum likelihood
MLE	maximum-likelihood estimation
MRV	multiple rare variants
NCP	noncentrality parameter
OLS	ordinary least squares
PAR	pseudoautosomal region
PCA	principal components analysis
PCR	polymerase chain reaction

POE	parent-of-origin effect
Q	variance component attributed to QTL effect
QC	quality control
QTDT	quantitative transmission disequilibrium test
QTL	quantitative trait locus
rGE	gene-environment correlation
RNA	ribonucleic acid
SEM	structural equation modeling
SNP	single nucleotide polymorphism
TDT	transmission disequilibrium test
VC	variance components
WGAS	whole genome association study
WLS	weighted least squares
WTCCC	Welcome Trust Case Control Consortium

Statistical symbols

General statistics

Term	*Symbol*
Mean	μ
Variance	σ^2
Degrees of freedom	df
Chi-square	χ^2
Summation	Σ
Multiplicative combination	Π
Normal distribution	$N(\mu,\sigma^2)$
Normal function	$\varphi(x)$
Probability of A	P(A)
Probability of A given B	$P(A \mid B)$
P (type-I error)	α
P (type-II error)	β
Correlation coefficient	ρ
Squared correlation coefficient	ρ^2
Variance covariance matrix	Σ
Likelihood of x	L(x)
Total sample size	N
Regression coefficients	β

Genetics

Term	*Symbol*
Additive genetic component of variance	σ_A^2
QTL component of variance	σ_Q^2
Additive QTL component of variance	σ_{Qa}^2
Dominance QTL component of variance	σ_{Qd}^2
Common environment component of variance	σ_C^2
Dominance component of variance	σ_D^2
Specific environment/error	σ_E^2
Total phenotypic variance	σ^2 or σ_P^2

Term	*Symbol*
Recombination fraction	θ
Biallelic locus system	1) A_1A_1, A_1A_2, A_2A_2 2) AA, Aa, aa.
Major/minor allele frequency in biallelic locus	p and q
Polymorphism information content	PIC
Heterozygosity	H
Additive deviation from 0 (biometrical model)	a for A_1A_1 and $-a$ for A_2A_2
Dominance deviation from midpoint of homozygotes	d for A_1A_2
Kinship coefficient	ϕ
Average IBD sharing	π
Estimated IBD sharing	$\hat{\pi}$

1 Introduction

Nicholas G. Martin

In this book, and in the field of quantitative trait loci (QTL) mapping in general, we are particularly concerned with study designs and analytical methods that enable us to address the following questions:

(i) Are there chromosomal regions that harbor genetic variants that influence the variation of a given *heritable* trait?
(ii) Can we precisely identify these genetic variants or polymorphisms?
(iii) How much do these variants contribute to trait variation in the population?

Our primary focus is on polygenic traits, that is, traits with a significant proportion of their variation attributable to a large number of genetic factors. These include well-defined traits such as body height or eye color, but also endophenotypes for common diseases, such as cholesterol levels for heart disease, insulin levels for type-2 diabetes or neurotic personality for depression. Our ultimate goal is to increase our understanding of the molecular mechanisms that underlie human trait variation. We hope that the identification of genetic risk factors for a complex trait is one major first step towards this goal. The theoretical background to the genetics of complex traits and the statistical issues involved in tackling them are covered in the opening chapters of this book (Chapters 3 through 6).

QTL mapping has a come a long way since Mendel's experiments, both conceptually and technologically. Morgan's discovery of linkage in 1910 (Morgan, 1910) led to the construction of a detailed genetic map of *Drosophila* and other experimental organisms in the following decades, but the shortage of good genetic markers in humans (mainly blood groups) meant that the first example of autosomal linkage in humans was only reported in 1955 (Renwick and Lawler, 1955). The development of electrophoretic protein markers in the 1950s and 1960s and then of restriction fragment length polymorphisms of DNA (RFLPs) in the 1970s produced some improvement

but still left human geneticists well behind their colleagues working on experimental organisms that could be mutagenised and crossed at will. All this changed with the discovery of microsatellite markers in 1989 (Litt and Luty, 1989; Weber and May, 1989) and, consequently, the development of high-throughput platforms for microsatellite genotyping. This effectively allowed statistical geneticists to apply standard linkage approaches, which had been proposed many years earlier in the context of gene mapping for Mendelian diseases, to the whole genome in large collections of families phenotyped for complex traits. In the last few years the dropping cost of high density SNP typing methods has improved the power of linkage analysis even more.

Linkage analysis, in its original parametric framework, formally tested whether the recombination fraction between two genetic markers, or between a genetic marker and a Mendelian disease locus, was different from 0.5 (that is the recombination fraction that would be expected between two loci that segregate independently according to Mendel's second law of inheritance). If the recombination fraction between the two loci was <0.5, the two loci were said to cosegregate, that is to be genetically linked. In this way, by sequentially testing many markers along the genome for linkage to an inferred Mendelian disease locus, linkage analysis allowed researchers to *localize* markers that cosegregated with the disease locus and thereby to locate a chromosomal region likely to harbor the disease locus. As we shall see in Chapters 7 through 13, linkage analysis methods for the analysis of complex traits have retained this fundamental framework but have been modified to account for the polygenic nature of most human phenotypes. Put simply, these methods test whether affected individuals *within a family* tend to share the same ancestral predisposing DNA segment at a given locus. For quantitative traits, for which 'affected' and 'unaffected' are not always clearly defined, linkage methods have been developed that test to what extent resemblance between relatives (often sibs) depends on their chromosomal sharing at a particular region.

Linkage analysis is particularly suited to address the first question raised above, that is, to localize regions of the genome that are likely to harbor disease loci. However, for reasons that we shall discuss later, it is of very limited use to the ultimate goal of *identifying* the genetic variants that contribute to phenotypic variation in humans. For the latter, association analysis has been the method of choice in recent years. Association mapping is conceptually very different from linkage analysis in that it tests the correlation between a specific variant and a trait or disease, independent of

ancestral considerations. From a statistical perspective, association methods are simpler and considerably more powerful, but only work if one has markers very close to the causal variant. This requirement of density of genetic information is being fulfilled by genome-wide association, which holds the most promise for mapping causal variation to date. Association mapping forms the second major topic of this book and is covered in detail in Chapters 14 through 21.

Many factors influence the success of both linkage and association analyses, but the impact of these can often be minimized by careful study design (Chapter 5) and choice of optimal statistical methods (Chapter 4). Failure to address these issues has led in the past to a plethora of type-I errors (reporting of QTL results that no one can replicate), and the reaction to this has often imposed severe significance criteria that discard many true results. Clearly, it is important to steer a course between the Scylla of high type-I error and the Charybdis of high type-II error, and Chapters 13 and 21 examine these issues for linkage and association respectively. Other chapters consider ways to increase power of both linkage and association by combining multiple measures in multivariate linkage analysis (Chapter 12) and multiple markers in haplotype analysis (Chapters 17 and 18).

Association analysis, as seen for linkage analysis in the early 1990s, has greatly benefited from significant technological breakthroughs in SNP genotyping platforms in recent years. Currently, it is feasible to genotype individual samples at almost a million SNPs across the genome (Chapter 16). This advance has not only allowed researchers to test markers all across the genome for association, but has opened a large array of analytical possibilities, such as genome-wide control for population stratification. In the last couple of months of 2006 there have been spectacular successes using genome-wide association studies, finding new QTLs for macular degeneration (Klein *et al.*, 2005), inflammatory bowel disease (Duerr *et al.*, 2006), and type-2 diabetes, (Sladek *et al.*, 2007). This trickle of early successes is likely to grow to a torrent in coming years, given that large epidemiologic samples of cases and controls have been collected over the last few decades by researchers around the world, their DNA is in the freezer, and that the price of high density SNP arrays will keep dropping.

In the near future, these traditional methods of linkage and association are likely to be augmented by new molecular tools, of which expression and methylation array studies are current examples. If discordance in monozygotic (MZ) twins for a complex trait can be correlated with discordance in expression or epigenetic

modification of a particular gene (either by DNA methylation or other mechanisms), this may provide a further layer of evidence for involvement of a particular gene or pathway. High density SNP arrays also allow the detection of copy number variation (CNV) over small or larger regions, and it is likely that CNV will prove to be important in explaining a portion of genetic variance in complex traits, though how important no one yet knows. There is increasing evidence that most polygenic variation is going to be due to a large number of rare variants which possibly have large effects within the few families in which they segregate, but account for trivial proportions of variance at the population level. Such variants will only be detected with larger and larger SNP arrays, or indeed with complete resequencing, which is within reach of being an economic proposition. All these technological developments will generate huge volumes of data which will need to be sifted and digested by powerful and sensitive statistical techniques.

References

Duerr, R.H., Taylor, K.D., Brant, S.R., *et al.* (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**: 1461–1463.

Klein, R.J., Zeiss, C., Chew, E.Y., *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385–389.

Litt, M., Luty, J.A. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**: 397–401.

Morgan, T.H. (1910) Chromosomes and heredity. *Am. Nat.* **44**: 449–496.

Renwick, J.H. and Lawler S.D. (1955) Genetic linkage between the ABO and nail-patella loci. *Ann. Hum. Genet.* **19**: 312–331.

Sladek, R., Rocheleau, G., Rung, J., *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**: 881–885.

Weber, J.L. and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**: 388–396.

Online resources

The URLs for freely available software, data and other online resources presented herein are as follows (listed alphabetically under the section where they appear first):

Section 1: The Basics

Software

Mx. Package for Structural Equation Modeling (http://www.vcu.edu/mx/).

The R Project for Statistical Computing (http://www.r-project.org/).

Data and other resources

International Workshop on Methodology of Twin and Family Studies. Homepage for the workshop that led to this book. Change year in the link for previous and subsequent workshops (http://ibgwww.colorado.edu/workshop2005/).

Statistical Genetics: gene mapping through linkage and association analysis. Scripts and most datasets used in this book can be downloaded from here (http://www.genemapping.org/).

Section 2: Linkage Analysis

Software

Cygwin. A Linux-like environment for Windows (http://www.cygwin.com/).

GRR. Graphical Representation of [pedigree] Relationships (http://www.sph.umich.edu/csg/abecasis/GRR/).

MERLIN. A user-friendly program to perform common pedigree genetic analyses, including linkage and association (http://www.sph.umich.edu/csg/abecasis/Merlin/index.html).

MERLIN-regress. Quantitative trait regression-based linkage analyses (http://www.sph.umich.edu/csg/abecasis/Merlin/tour/regress.html).

Pedstats. A program to summarize the contents of pedigree files (http://www.sph.umich.edu/csg/abecasis/PedStats/index.html).

GPC. A website that provides automated power analyses for common linkage and association analyses (http://pngu.mgh.harvard.edu/~purcell/gpc/).

Data and other resources

Chapter 8 examples (http://www.sph.umich.edu/csg/abecasis/Merlin/download/).

Mx library scripts. A compilation of Mx scripts for genetic analyses (http://www.psy.vu.nl/mxbib/).

Section 3: Association Analysis

Software

CaTS. Power calculation program for multistage case-control genome association analysis (http://www.sph.umich.edu/csg/abecasis/CaTS).

FBAT/PBAT. Programs for general family-based tests of association (http://biosun1.harvard.edu/~fbat).

GWAVA. Software platform tailored for genome association analysis (http://www.sph.umich.edu/csg/abecasis/gwava).

Haploview. A program to perform haplotype analysis and related association tests (http://www.broad.mit.edu/mpg/haploview/).

PLINK. Open-source whole-genome association analysis toolset (http://pngu.mgh.harvard.edu/~purcell/plink/).

Quanto. Power calculator for association studies incorporating gene-gene and gene-environment interactions (http://hydra.usc.edu/gxe).

QTDT. Interface to perform family-based analyses for quantitative and discrete traits (http://www.sph.umich.edu/csg/abecasis/QTDT/).

WHAP. A program to test SNP haplotype associations with qualitative and quantitative traits (http://pngu.mgh.harvard.edu/~purcell/whap/).

Data and other resources

International HapMap project. A public resource of human haplotypic variation (http://www.hapmap.org/).

The Genetic Association Information Network (GAIN) program. A public-private partnership established to understand the genetic factors influencing risk for complex diseases (http://www.fnih.org/GAIN/GAIN_home.shtml).

The Wellcome Trust Case Control Consortium (WTCCC). A collaboration of leading human geneticists to identify genetic variants that influence the risk of eight common diseases (http://www.wtccc.org.uk/).

STATISTICAL GENETICS

Gene Mapping Through Linkage and Association

Statistical Genetics is an advanced textbook focusing on genetic linkage and association analysis in the post-genomic era, where the emphasis is on conducting genome-wide studies into complex behaviors and diseases. Covering both established and new methodologies, this comprehensive volume provides the underlying genetics and statistical theory. The book has many pedagogical features including worked examples, using a problem-based approach, study design advice and sources of error. Each of the 22 chapters is written by a leading researcher, many of whom have authored the genetic software programs used.

An accompanying website **www.genemapping.org** provides supplementary online resources, including example scripts, data files and links to download the software used in the tutorials.

Statistical Genetics is a valuable resource for students of statistical genetics, genetic epidemiology or human molecular genetics with an interest in gene mapping of complex human traits.

ISBN 978-0-415-41040-3

9 780415 410403

Taylor & Francis
Taylor & Francis Group