Shili Lin • Hongyu Zhao
Editors

Handbook on Analyzing Human Genetic Data

Computational Approaches and Software

Springer

Editors

Professor Dr. Shili Lin
Department of Statistics
The Ohio State University
Columbus, Ohio 43210
USA
shili@stat.ohio-state.edu

Professor Dr. Hongyu Zhao
Department of Epidemiology and Public Health
Yale University
School of Medicine
60 College St.
New Haven, CT 06520-8034
USA
hongyu.zhao@yale.edu

ISBN 978-3-540-69263-8 e-ISBN 978-3-540-69624-5
DOI 10.1007/978-3-540-69264-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009931713

Cover design: WMX Design GmbH, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface and Introduction

The discipline of statistical genetics is highly computational. Be it exact computational methods, simulation based, or a hybrid of the two, computational packages are indispensable tools and constant companions of researchers in the field. This handbook is intended to provide human geneticists and other biomedical researchers with guidance on selections of appropriate computational methods and software packages for their specific genetic problems. It may also be used by students and other learners as a reference in conjunction with a more theoretical and/or methodologically oriented text book. This book tries to strike a balance between methodological expositions and practical guidelines for software selections. Wherever possible, comparisons among the competing methods and software are made to highlight the relative advantages and disadvantages of the approaches, so that the reader can make informed choices to best match their specific needs.

Human genetics has been undergoing an evolution in the past several years as new knowledge and technologies are transforming the field, leading to numerous new discoveries of genes associated with complex traits such as cancer, obesity, and diabetes. Many recent genome-wide association studies employ the case–control design, where the study subjects consist of unrelated affected individuals and normal controls. For each individual, a large number of genetic markers are queried. A genetic marker refers to a location in the human genome where people may differ in the genetic material they carry. Genetic markers can come in different forms, with the single nucleotide polymorphisms (SNPs) most commonly used due to their high abundance in the genome and the availabilities of reliable and affordable technologies to genotype them. For a SNP, two different forms (called alleles) generally exist at a single nucleotide position. Because each person carries two chromosomes, for a given SNP with two alleles A and a, there are three possible genotypes a person can have: AA, Aa, and aa. In this setting, a genetic association study amounts to identifying markers that are associated with disease status. This can be accomplished by examining whether there is a statistical association between the marker genotype and the disease status. Although this analysis resembles a standard epidemiological study where each marker can be treated as a potential risk factor, there are many issues that are unique to genetics studies that need to be addressed. For example, one major concern in these studies is sample heterogeneity in their genetic background, and ignoring this issue may result in many false positive findings that have nothing to do with disease etiology. On the other hand, much research has been done to empirically characterize and theoretically model the distributions and dependencies of genetic markers, and such knowledge is very beneficial for association analysis. In fact, a thorough genetic association analysis is not possible without a good understanding of the basic principles in population genetics, a field devoted to the study of the allele frequency distribution and change under various factors that can impact them, including mutations, random sampling, migrations, and natural selections. The chapter by Dr. Weir provides an overview of the basic concepts of population genetics and serves as the starting point of the analysis of human genetics data.

Although current genotyping platforms can genotype up to one million markers, there are many more markers in the genome that are not queried on these platforms. The reason that these typed markers can provide a good coverage of the genome is the dependence among physically close markers, and such dependence is called linkage disequilibrium. For example, if one SNP has alleles A and a each with allele frequency 50%, and another marker with alleles B and b each with frequency 50%. If the two markers are independent of each other, we would expect that 25% of chromosomes carry both A and B in the population, and similarly for all other three possible combinations: Ab, aB, and ab. However, it is often the case that if these two markers are very close to each other on the same chromosome, the two alleles carried on the same chromosome are not independent. In the most extreme case, there are only two types of chromosomes, those carrying AB and those carrying ab, a phenomenon called perfect linkage disequilibrium. Haplotypes refer to the combination of alleles on the same chromosome, and the presence of such marker dependency is the key underlying recent successes of genetic association studies collecting the genotypes from only a small fraction of all known markers. There are many statistical challenges presented in the analysis of haplotypes, both for population genetics studies and for more effective genetic association studies. These topics are discussed in the chapter by Drs. Zhang and Niu focusing on population genetics and in the chapter by Drs. Epstein and Kwee in the context of disease association analysis.

Genetic association studies can be performed on unrelated individuals using traditional epidemiological designs, for example, case–control design and cohort design, or designs unique to genetic studies, for example, family-based association design. Because sample heterogeneity in genetic background is one major concern in the validity of a genetic association study based on unrelated individuals, various statistical methods have been proposed to utilize genetic information in the collected marker genotypes to make appropriate adjustments in association analysis. For example, with enough marker information, it is possible to infer genetic background for each individual and such inferred background information can be incorporated in association analysis to make the results less susceptible to sample heterogeneity. This issue is thoroughly studied and addressed in the chapter by Drs. Zhu and Zhang.

With data from related individuals, genetic association tests may be conducted in a manner that is valid (i.e., not subject to bias due to sample heterogeneity) even without utilizing genetic markers to infer genetic background. The basic principle is to detect whether there is a departure from random marker segregation at a candidate locus. For example, if a study population consists of affected children and their parents and a marker with two alleles A and a is studied for its potential involvement in the disease. If the marker has nothing to do with disease phenotype, we expect that a parent who is heterozygous Aa would have equal chance to transmit allele A or a to his/her affected offspring. On the other hand, if allele A increases disease risk, we would expect to observe a preferential transmission of allele A to the affected offspring. This testing procedure is robust to sample heterogeneity as the inference is conditional on each parent's genotype and the only genetic principle tested is random marker allele transmission from parents to offspring, the Mendel's first law. Many statistical developments along this research route are discussed in the chapter by Drs. Zhang and Zhao.

Both population-based and family-based association studies examine statistical associations between a phenotype and the genotypes at a marker. One implicit assumption is that the same marker genotype would exert the same or similar effects on a phenotype. While this is expected to be the case for most genetic markers that have direct functional impact, this assumption may well be violated for many markers. For example, consider a marker with two alleles A and a studied is not functional but rather is in linkage disequilibrium with a truly functional one with two alleles D and d. It is possible that A is positively associated with D in one population, that is, someone carrying A on one chromosome is also more likely to carry D on the same chromosome, but A is negatively associated with D in another population. In this case, an analysis using samples from these two populations together may not even be able to detect a genetic association. More importantly, when the markers are sparse and not expected to provide a good coverage of the genome, the association analysis paradigm discussed above will not be effective as a large proportion of the genome that likely harbors disease genes may be missed due to poor coverage. This was in fact the case only a few years ago when only fewer markers could be used for genetic analysis. In this scenario, although the markers were not dense enough to cover the genome for association analysis, they were more than adequate to allow geneticists to infer whether two relatives in an ascertained pedigree share a segment in the genome from the same ancestor. For example, if two siblings have the same marker genotypes across a set of closely linked markers on the same chromosome, then they likely have inherited the same genetic materials from both their parents. A genetic linkage analysis is to statistically assess whether there is a co-segregation of genetic materials within a candidate region and the phenotype within a family. For example, this can be done by studying whether there is a correlation between trait similarities and inheritance similarities at a candidate region among a set of individuals from the same family. Consider a study enrolling affected sib pairs. If majority of them share the same genetic materials from their parents in a region, then this region is likely involved in disease etiology. Note that in contrast to association analysis that is performed across all study subjects, linkage analysis is conducted within families and evidence is then summed over across individual families. Statistical methods for linkage analysis can be conducted for either qualitative traits (the chapter by Dr. Li and Abecasis) or quantitative traits (the chapter by Drs. Amos, Peng, Xu, and Ma).

Exact inference of inheritance patterns within a pedigree is tractable either for a small pedigree or for a few markers, but such inference becomes computationally prohibitive for large pedigree with many genetic markers. In this case, the exact probabilities may be estimated by Monte Carlo simulations. In the chapter by Drs. Igo, Luo, and Lin, the principles and implementations underlying the simulation methods for linkage analysis in large complex pedigrees are discussed.

One central topic in statistical inference is the control of false positive results so as to minimize any consequences resulting from false leads. This issue has been well addressed when only one or a small number of statistical hypotheses are tested. However, hundreds of thousands of markers are tested for their associations with disease in a genome-wide association study, and false positive control at the individual marker levels will not be adequate. For example, if a study considers 500,000 markers and the statistical significance level is set at 0.01, we would expect to see 5,000 false positive results even when there is no association between disease status and any of the markers. Similar issue exists in the linkage analysis context, although not to the same great extent as association analysis. The chapter by Drs. Zhang and Ott presents some recent developments on appropriately controlling overall false positive results in genetic studies at the genome level.

The identifications of disease genes can lead to biological insights on pathways involved in disease etiology, and these findings can also be used to predict an individual's disease risk. In the chapter by Drs. Gail and Chatterjee, they discuss statistical methods that can be used to make use of findings from genetic studies to identify individuals at higher risks for disease.

The book concludes with the last chapter by Drs. Molony, Sieberts, and Schadt, where they discuss integrating genetics and genomics data to better delineate biological pathways underlying complex traits. In addition to disease status and possibly other clinical outcomes, they consider gene expression data that can now be routinely gathered to measure the expression levels of tens of thousands of genes simultaneously for each study subject. These gene expression data add another whole new dimension of statistical analysis and are very information rich. In principle, the expression level of each gene can be thought as a quantitative trait, and linkage/association analysis can be conducted to identity genes regulating a gene's expression level. Therefore, based on this perspective, we would be in a position to conduct genetic analysis for tens of thousands of traits. Some of these expression levels may be associated with disease outcome, and so it is natural to investigate how a genetic variation affects the expression levels as well as disease outcomes. Many biological questions on the underlying genetic networks relating genetic variations, expression variations, and phenotype variations can be posed and answered with these data. This chapter discusses topics falling into the domain of systems biology where the whole biological system is the focus of a study and genome-level data of different types are needed to dissect the networks.

We hope that this book will provide an overview of the most important areas in genetic data analysis methods. We focus on fundamental principles and, when possible, demonstrate these principles with real data examples. Despite our efforts, this is not an encyclopedia of statistical methods in human genetics, and some topics are not included such as the experimental design of a genetic study, data preprocessing from high-throughput genotyping platforms, and copy number variations. Most importantly, this is a very rapidly developing field and new technologies are constantly introduced that demand novel statistical approaches to make the most use of the data collected. For example, the statistical methods discussed in this book may not be the most effective for inferring inheritance patterns in a pedigree using high density SNP data. On the other hand, the availabilities of re-sequencing data from a large number of study subjects lead to a new set of informatics and statistical challenges, such as the incorporation of SNP annotation information and the dealing of rare genetic variations. We hope the basic principles and statistical methods discussed in this book will motivate the readers to develop their own approaches if necessary to accelerate our progresses in mapping disease genes.

Contents

Contributors

Gonçalo R. Abecasis Department of Biostatistics, Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA, goncalo@umich.edu

Christopher I. Amos Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, camos@mdanderson.org

Nilanjan Chatterjee Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, MD 20852, USA

Michael P. Epstein Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322, USA, mepstein@genetics.emory.edu

Mitchell H. Gail Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, MD 20852, USA

Robert P. Igo Department of Epidemiology and Biostatistics, Division of Genetics and Molecular Epidemiology, Case Western Reserve University, Cleveland, OH, USA, rigo@darwin.EPBI.CWRU.edu

Lydia C. Kwee Department of Biostatistics, Emory University, Atlanta, GA, USA

Mingyao Li Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA, mingyao@mail.med.upenn.edu

Shili Lin Department of Statistics, The Ohio State University, OH, USA, shili@stat.osu.edu

Yuqun Luo Department of Epidemiology and Biostatistics, Division of Genetics and Molecular Epidemiology, Case Western Reserve University, Cleveland, OH, USA, yuqun.luo@case.edu

Jianzhong Ma Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, jzma@mdanderson.org

Cliona Molony Rosetta Inpharmatics, LLC, (a wholly owned subsidiary of) Merck & Co., Inc., Seattle, WA 98109, USA

Tianhua Niu Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02215, USA

and Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA, tniu@hsph.harvard.edu

Jurg Ott Beijing Institute of Genomics, Chinese Academy of Sciences, No. 7 Bei Tu Cheng West Road, Beijing 100029, China, ottjurg@yahoo.com

Bo Peng Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, bpeng@mdanderson.org

Eric E. Schadt Rosetta Inpharmatics, LLC, a wholly owned subsidiary of Merck & Co., Inc., Seattle, WA 98109, USA

Solveig K. Sieberts Rosetta Inpharmatics, LLC, (a wholly owned subsidiary of) Merck & Co., Inc., Seattle, WA 98109, USA

Bruce Weir Department of Biostatistics, University of Washington, Seattle, WA 98185.

Yaji Xu Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Blvd, Unit 1340, Houston, TX, 77030, USA, yajixu@mdanderson.org

Yu Zhang Department of Statistics, the Pennsylvania State University, 422A Thomas Building, University Park, PA 16802, USA, yuzhang@stat.psu.edu

ShuangLin Zhang Department of Mathematical Science, Michigan Technological University, Houghton, MI, USA

Kui Zhang Section on Statistical Genetics, Department of Biostatistics University of Alabama at Birmingham, Birmingham, AL 35294, USA

Qingrun Zhang Chinese Academy of Sciences, Beijing Institute of Genomics, Beijing, China

Hongyu Zhao Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, 06520, USA

Xiaofeng Zhu Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Population Genetics

Bruce Weir

Abstract Understanding population genetics is critical to designing and interpreting results for human genetic studies. Much research has been done in this area in the past century, often involving sophisticated mathematical and computational tools. However, there has been a detachment between theoretical developments and real data analyses primarily due to the lack of data for population genetics studies. The landscape has changed completely due to the recent advances in molecular technologies allowing high-throughput and affordable sequencing and genotyping of a large number of samples. For example, the on-going HapMap project can be considered a very large-scale population genetics study where genetic variation throughout the genome in diverse populations is thoroughly studied. This chapter covers basic statistical tests and procedures involved in the analysis of population genetics data, such as the tests of Hardy–Weinberg equilibrium and linkage equilibrium and characterization of population structure.

1 Introduction

Human population genetic studies have entered a new era with substantial amounts of data being reported on large samples. Early in the twentieth century, data were emerging on human blood group frequencies, reflecting variation at very few loci in samples that rarely exceeded 100 individuals. One hundred years later, we have public access to data at up to one million single nucleotide polymorphisms (SNPs) for samples in the thousands [1], and the Archon X Prize of $10 million has been established for the first group to completely sequence 100 individuals in 10 days (http://genomics.xprize.org). These new data sets offer both challenges and opportunities. The obvious benefit is the increased precision they allow to characterize relationships between people, between populations, and between genetic variants and human disease. On the other hand, large data sets bring problems of data handling, multiple testing and accommodating interactions among sets of genes.

B. Weir
Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195–7232.

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_1, © Springer-Verlag Berlin Heidelberg 2010

This chapter describes some of the basic statistical analyses that can be performed on human population genetic data, and illustrates those analyses with data collected by the FBI for forensic purposes [2]. We will adopt two principal frames of reference: analyses for data collected from a single population that lead to inferences for just that population and then analyses for data that allow for variation among populations. In the first place, the analyses reflect the sampling strategies adopted by the investigator. Sample size and family relationships among sampled individuals, for example, determine properties of estimates of genetic parameters. In the second place, the investigator's strategies may be less important than the sampling that is inherent in the evolutionary process that has led to the current divergence among populations. There is less guidance in this situation from standard statistical theories. In both situations, however, it is convenient to describe many of the analyses in terms of correlations between pairs of alleles, whether the pairs are within individuals, or within populations or between populations.

2 Within-Population Analyses

2.1 Genotype and Allele Frequencies

Our initial development is for samples of n individuals taken randomly from a single population. Genotypes G_i are represented n_i times in the sample, so $\sum_i n_i = n$. Sample genotype frequencies are $\tilde{P}_i = n_i/n$ and the corresponding population frequencies are P_i. Because of the random sampling assumption, the genotype counts are multinomially distributed so that means, variances and covariances of the frequencies are:

$$\begin{aligned}
\mathcal{E}(\tilde{P}_i) &= P_i \\
\mathrm{Var}(\tilde{P}_i) &= \frac{1}{n}P_i(1-P_i) \\
\mathrm{Cov}(\tilde{P}_i, \tilde{P}_{i'}) &= -\frac{1}{n}P_iP_{i'}, \;\; i \neq i'
\end{aligned}$$

If we focus on a specific genotype, say I, then the count n_I has a binomial distribution and it is often appropriate to approximate that by a normal distribution. For the frequency, then,

$$\tilde{P}_I \sim N\left(P_I, \frac{1}{n}P_I(1-P_I)\right)$$

and confidence intervals can be constructed for the population genotype frequency. As 95% of the standard normal distribution lies between ± 1.96, the 95% confidence interval for P_I is $\tilde{P}_I \pm 1.96\sqrt{\tilde{P}_I(1-\tilde{P}_I)/n}$.

When multinomial categories are combined into a smaller number of categories, the distribution remains multinomial. So, if data are collected for pairs of loci **A** and **B**, each of which has two alleles A, a and B, b there are nine two-locus genotype classes. The one-locus genotype classes can be found by summation, for example,

$$n_{AA} = n_{AABB} + n_{AABb} + n_{AAbb}$$

Both the two-locus and the one-locus genotype counts are multinomially distributed. The same preservation does not hold, however, when categories are subdivided. Consider the allele counts n_A, n_a for two alleles at a single locus:

$$\begin{array}{lll} n_A = 2n_{AA} + n_{Aa} & \tilde{p}_A = \frac{n_A}{2n} & \tilde{p}_A = \tilde{P}_{AA} + \frac{1}{2}\tilde{P}_{Aa} \\ n_a = 2n_{aa} + n_{Aa} & \tilde{p}_a = \frac{n_a}{2n} & \tilde{p}_a = \tilde{P}_{aa} + \frac{1}{2}\tilde{P}_{Aa} \\ n_A + n_a = 2n & & \tilde{p}_A + \tilde{p}_a = 1 \end{array}$$

The heterozygote count n_{Aa} contributes to both allele counts, and this means that the count for any one allele is not binomially distributed. This can be seen most easily by considering the variance of an allele frequency:

$$\begin{aligned} \mathrm{Var}(\tilde{p}_A) &= \frac{1}{4n^2}\left[\mathrm{Var}(2n_{AA}) + \mathrm{Var}(n_{Aa}) + 2\mathrm{Cov}(2n_{AA}, n_{Aa})\right] \\ &= \frac{1}{2n}p_A(1-p_A) + \frac{1}{2n}(P_{AA} - p_A^2) \end{aligned} \tag{1}$$

The first term on the right-hand side looks like a binomial variance, but there is an additional term that reflects possible departures from Hardy–Weinberg equilibrium (HWE) in the population, $P_{AA} \neq p_A^2$. The HWE law states that in the absence of forces such as genetic drift, selection, mutation and migration, genotype probabilities are products of allelic probabilities. If there is HWE in a population, then a random sample of n individuals (genotypes) is equivalent to a random sample of $2n$ alleles at each locus as the alleles within individuals are independent in that case.

2.2 Maximum Likelihood Estimation

The multinomial probability $\Pr(\{n_i\}|\{P_i\})$ of a set of genotype counts depends on the population frequencies:

$$\Pr(\{n_i\}|\{P_i\}) = \frac{n!}{\prod_i n_i!}\prod_i (P_i)^{n_i}$$

and this leads to the likelihood of the frequencies given the counts:

$$L(\{P_i\}|\{n_i\}) = C\prod_i (P_i)^{n_i} \tag{2}$$

Here, C is an arbitrary constant. The likelihood contains only the terms in the probability that involve the parameters P_i. The maximum likelihood estimates (MLE) $\hat{P}_i$ of the parameters are those values that maximize the likelihood and these are found to be $\hat{P}_i = n_i/n = \tilde{P}_i$. Note that, if there is not HWE, the MLE of an allele frequency is not generally the sample frequency because the allele counts do not follow a multinomial distribution. We will see an exception to this rule for the case of loci with only two alleles.

MLEs have many desirable properties, including sufficiency (they make use of all information in the data) and efficiency (they have smaller variances than other estimates), but they need not be unbiased. To show this, consider the estimates of "heterozygosity." If this term means the proportion of heterozygotes, as it should, then things are simple. If H is the probability that a random individual is heterozygous, then the sample frequency of heterozygotes $\tilde{H}$ is just the sum of the frequencies of all heterozygous genotypes and it is the MLE of H. Moreover, it is unbiased since it has expectation $\mathcal{E}(\tilde{H}) = H$. However, the term "heterozygosity" is often applied to the expression $1 - \sum_u p_u^2$ or one minus the sum of squares of allele frequencies at a locus. This quantity should be referred to as "allele diversity" D, and it has an MLE of

$$\hat{D} = 1 - \sum_u \tilde{p}_u^2$$

To find the expectation of this estimate, we note that $\mathcal{E}(\tilde{p}_u^2) = \mathrm{Var}(\tilde{p}_u) + p_u^2$. This leads to

$$\mathcal{E}(\hat{D}) = D - \frac{1}{2n}(2D - H)$$

Even if there is HWE, and $H = D$, the MLE of D has a small bias.

2.3 Inbreeding Coefficient

It is convenient to introduce parameters that measure the departure from HWE in a population. For a locus with two alleles, a single inbreeding coefficient f can play this role. This quantity can be defined by the following equations:

$$\begin{aligned} P_{AA} &= p_A^2 + f p_A p_a \\ P_{Aa} &= 2p_A p_a - 2f p_A p_a \\ P_{aa} &= p_a^2 + f p_A p_a \end{aligned}$$

These equations preserve the property that $p_A = P_{AA} + P_{Aa}/2$, but what is f? In the first place, we note that it has bounds imposed by the fact that genotype frequencies are bounded by zero and allele frequencies: $0 \leq P_{AA} \leq p_A$, $0 \leq P_{aa} \leq p_a$, $0 \leq P_{Aa} \leq \min(2p_A, 2p_a)$. These lead to

$$\max\left(-\frac{p_A}{p_a}, -\frac{p_a}{p_A}\right) \leq f \leq 1.$$

The best description of f is that it is an intra-class correlation coefficient. For the ith individual in a sample, replace the two alleles by indicator variables $x_{ij}, j = 1, 2$ that take the values 1 for A alleles and 0 for a alleles. Genotypes AA, Aa, aa are coded $11, 10, 00$. Taking expectations of x's

$$\begin{aligned}
\mathcal{E}(x_{ij}) &= p_A \\
\mathcal{E}(x_{ij}^2) &= p_A \\
\text{Var}(x_{ij}) &= p_A p_a \\
\mathcal{E}(x_{i1}x_{i2}) &= P_{AA} \\
\text{Cov}(x_{i1}, x_{i2}) &= P_{AA} - p_A^2 = f p_A p_a
\end{aligned}$$

The correlation of the two x's for the same individual, $\text{Corr}(x_{i1}, x_{i2})$ is the covariance of these two x's divided by the square root of the product of their variances. The correlation is f.

The number of A alleles in the sample is the sum of these indicators: $n_A = \sum_{i=1}^{n} \sum_{j=1}^{2} x_{ij}$. For random samples, the x's from different individuals are independent: $\mathcal{E}(x_{ij}x_{i'j'}) = p_A^2, i \neq i'$, and using this result as well as the expectations for x's in the same individual leads to the variance of a sample allele frequency,

$$\text{Var}(\tilde{p}_A) = \frac{p_A p_a (1 + f)}{2n}$$

which is equivalent to the result in (1).

The MLE of f can be found from replacing genotype probabilities in (2) by expressions involving f and allele frequencies

$$\begin{aligned}
L(p_A, f | n_{AA}, n_{Aa}, n_{aa}) = C\big[p_A^2 + f p_A(1 p_A)\big]^{n_{AA}} \big[2(1-f)p_A(1-p_A)\big]^{n_{Aa}} \\
\times\big[(1-p_A)^2 + f p_A(1-p_A)\big]^{n_{aa}}
\end{aligned}$$

and then maximizing this expression with respect to f and p_A. It is simpler in this case to note that the degrees of freedom equal the number of parameters and equate observed and expected genotype proportions. There are three genotypic categories and therefore two df, and there are two parameters f and p_A since $p_a = (1 - p_A)$. This provides

$$\begin{aligned}
\hat{p}_A &= \tilde{p}_A \\
\hat{f} &= 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a}.
\end{aligned}$$

Because $\hat{f}$ is an MLE it can be regarded as having a normal distribution for large sample sizes. Because it is a ratio, however, it is difficult to derive its mean and variance. Appealing to large-sample theory does provide the results [3].

$$\begin{aligned}\mathcal{E}(\hat{f}) &= f \\ \mathrm{Var}(\hat{f}) &= \frac{1}{2np_Ap_a}(1-f)\big[2(1-f)(1-2f)p_Ap_a + f(2-f)\big]\end{aligned} \tag{3}$$

With multiple alleles, the use of inbreeding coefficients is not quite as simple. With m alleles at a locus there are $m(m+1)/2$ genotypes, so there is a need for $m(m-1)/2$ inbreeding coefficients and it is convenient to define one for each heterozygote. Genotype frequencies for A_uA_u homozygotes and $A_uA_{u'}, u \neq u'$ heterozygotes can be re-parameterized as:

$$\begin{aligned}P_{uu} &= p_u^2 + \sum_{u' \neq u} f_{uu'}p_up_{u'} \\ P_{uu'} &= 2p_up_{u'}(1-f_{uu'}), \;\; u' \neq u\end{aligned}$$

HWE holds when all the f's are zero. With this parameterization, MLEs are

$$\begin{aligned}\hat{p}_u &= \tilde{p}_u \\ \hat{f}_{uu'} &= 1 - \frac{\tilde{P}_{uu'}}{2\tilde{p}_u\tilde{p}_{u'}}, \;\; u \neq u'.\end{aligned}$$

For neutral genetic markers, we are more likely to want to use a single inbreeding coefficient f. This imposes a constraint on the system since m allele frequencies and one inbreeding coefficient are being use to parameterize $m(m+1)/2$ genotype frequencies. It is still possible to find MLEs for the allele frequencies and the single f, but none of these now has analytic expressions. Numerical methods are needed to maximize the likelihood.

2.4 Testing for Hardy–Weinberg Equilibrium

For loci with two alleles there are several equivalent ways of testing the hypothesis that the sampled population is in Hardy–Weinberg equilibrium. Appealing to the asymptotic normality of the MLE for the inbreeding coefficient parameter f_A, the hypothesis that $f_A = 0$ can be tested for with the standard normal statistic $z = \hat{f}_A/\sqrt{\mathrm{Var}(\hat{f}_A)}$. Squaring this gives a statistic with a 1 df chi-square distribution:

$$X_A^2 = n\hat{f}_A^2$$

which uses the result that $\mathrm{Var}(\hat{f}_A) = 1/n$ when $f_A = 0$. The same underlying assumption of a large sample size leads to this test statistic from a goodness-of-fit perspective: the three observed counts $O : n_{AA}, n_{Aa}, n_{aa}$ are compared to the expected counts $E : n\tilde{p}_A^2, 2n\tilde{p}_A\tilde{p}_a, n\tilde{p}_a^2$ by means of the equation $X^2 = \sum(O - E)^2/E$.

For small samples, it is usual instead to perform an exact test, where "exact" refers to calculating the probability of a false rejection. The chi-square tests calculate this significance level from the chi-square distribution, which must be an approximation since the chi-square is a continuous distribution and the data are discrete. The exact test uses the multinomial distribution of the genotype counts, and works specifically with the probability of the genotype counts conditional on the observed allele counts. Because the allele counts are just sums of genotype counts

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a) = \frac{\Pr(n_{AA}, n_{Aa}, n_{aa})}{\Pr(n_A, n_a)}.$$

If there is HWE the allele counts are binomially distributed and

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(p_A^2)^{n_{AA}}(2p_Ap_a)^{n_{Aa}}(p_a^2)^{n_{aa}}$$

$$\Pr(n_A, n_a) = \frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a}$$

so

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a) = \frac{n!2^{n_{Aa}}n_A!n_a!}{(2n)!n_{AA}!n_{Aa}!n_{aa}!}.$$

This quantity is evaluated for a dataset to give a probability p. All sets of genotypic counts with the same allelic counts are then considered and the corresponding probability is added to p if it is not greater than p. The resulting sum is the significance level, or p-value, for the test. For loci with multiple alleles the probability is:

$$\Pr(\{n_{ij}\}|\{n_i\}) = \frac{n!}{\prod_{i\leq j} n_{ij}!}\frac{2^H \prod_i n_i!}{(2n)!},$$

where H is the number of heterozygotes in the sample and n_i, n_{ij} are allelic and genotypic counts. It may not now be possible to do a complete enumeration of all genotypic arrays, and a permutation procedure is used instead to give a random sample of arrays [4]. The $2n$ alleles in a sample of n genotypes are permuted to form a new genotype array, and the proportion of a large number of permutations (say 2,000) that lead to a smaller probability than the data is an estimate of the p value.

When there is not HWE, the conditional probability of the genotype counts given the allele counts in the two-allele cases can be written as:

$$\Pr(n_{AA}, n_{Aa}, n_{aa}|n_A, n_a) = \frac{C\phi^{n_{Aa}}}{n_{AA}!n_{Aa}!n_{aa}!} \tag{4}$$

[5], where $\phi = P_{Aa}/\sqrt{P_{AA}P_{aa}}$ and C is a normalizing constant that ensures the probabilities sum to one over all genotype arrays consistent with the allele counts n_A, n_a. Note that $\phi = 2$ for HWE. Once the rejection region of the exact test has been determined, the sum of the probabilities in (4) for that region gives the power of the test.

2.5 Linkage Disequilibrium

The inbreeding coefficient f was introduced as a means of describing the dependence between the two alleles at a locus carried by an individual. There is an analogous quantity, linkage disequilibrium, for a pair of alleles at different loci but transmitted together from one parent to an individual. For alleles A, B at loci **A** and **B**, the frequency of AB gametes (the material transmitted from parent to child) is written as p_{AB} and (gametic) linkage disequilibrium D_{AB} is defined as:

$$D_{AB} = p_{AB} - p_A p_B.$$

If the frequencies on the right-hand side are replaced by sample values, the left-hand side is the MLE. It has mean and (large-sample) variance of

$$\begin{aligned}\mathcal{E}(\hat{D}_{AB}) &= D_{AB}\\ \mathrm{Var}(\hat{D}_{AB}) &= \frac{1}{2n}\big[p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)\\ &\quad \times D_{AB} - D_{AB}^2\big],\end{aligned} \tag{5}$$

where $2n$ is the number of gametes in the sample.

As with the inbreeding coefficient, there are bounds on D_{AB} because gamete frequencies are bounded by allele frequencies, e.g., $0 \leq p_{AB} \leq \min(p_A, p_B)$. These lead to

$$\max\big[-p_A p_B, -(1-p_A)(1-p_B)\big] \leq D_{AB} \leq \min\big[p_A(1-p_B), (1-p_A)p_B\big].$$

The correlation nature of linkage disequilibrium is demonstrated by considering two indicator variables, x_i, y_i, defined as $x_i = 1$ if the **A** allele on the ith gamete in a sample is A and $y_i = 1$ if the **B** allele on the gamete is B. Otherwise, the variables are zero and the relevant expectations are:

$$\begin{aligned}\mathcal{E}(x_i) = p_A &\quad \mathcal{E}(y_i) = p_B\\ \mathcal{E}(x_i^2) = p_A &\quad \mathcal{E}(y_i^2) = p_B\\ \mathrm{Var}(x_i) = p_A(1-p_A) &\quad \mathrm{Var}(y_i) = p_B(1-p_B)\\ \mathcal{E}(x_i y_i) = p_{AB} &\quad \mathrm{Cov}(x_i, y_i) = D_{AB}\end{aligned}$$

The correlation between x_i and y_i is $\rho_{AB} = D_{AB}/\sqrt{p_A(1-p_A)p_B(1-p_B)}$.

In the case of no inbreeding at either locus, the estimated inbreeding coefficients $\hat{f}_A, \hat{f}_B$ at the two loci are correlated to an extent [5].

$$\mathrm{Corr}(\hat{f}_A, \hat{f}_B) = \frac{D_{AB}^2}{[p_A(1-p_A)p_B(1-p_B)]} = \rho_{AB}^2.$$

2.6 Composite Linkage Disequilibrium

Gametic linkage disequilibrium is estimated from sample gametic frequencies, while data are collected at the genotypic level. Because genotypic phase is generally unknown, meaning that only the genotype and not the constituent gametes are observed, double heterozygotes $AaBb$ cannot be partitioned into those that are formed by the union of AB and ab gametes from those formed by the union of Ab and aB gametes. If HWE can be assumed at each locus, then there are algorithms that allow gamete frequencies to be estimated from genotypic frequencies and these estimates can be used to estimate linkage disequilibrium (e.g., [6]).

We often prefer to avoid the HWE assumption and instead work with a measure of composite linkage disequilibrium that measures the dependence of alleles A and B in the same individual whether the alleles were transmitted to the individual on one or two gametes. If $p_{A,B}$ is the probability with which an individual receives alleles A and B from different parents, then the composite linkage disequilibrium Δ_{AB} is the sum of the gametic and the non-gametic disequilibria

$$\Delta_{AB} = (p_{AB} - p_A p_B) + (p_{A,B} - p_A p_B) = D_{AB} + D_{A,B},$$

This has an MLE of

$$\hat{\Delta}_{AB} = \frac{1}{n}\left(2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}\right) - 2\tilde{p}_A\tilde{p}_B$$

with mean and large-sample variance of

$$\begin{aligned}
\mathcal{E}(\hat{\Delta}_{AB}) &= \Delta_{AB} \\
\mathrm{Var}(\hat{\Delta}_{AB}) &= \frac{1}{n}[p_A(1-p_A)(1+f_A)p_B(1-p_B)(1+f_B) \\
&\quad + (1-2p_A)(1-2p_B)\Delta_{AB} + (1-2p_A)D_{ABB} \\
&\quad + (1-2p_B)D_{AAB} + \Delta_{AABB}],
\end{aligned}$$

where n is the number of individuals in the sample [7]. The quantities D_{AAB}, D_{ABB}, Δ_{AABB} are higher-order measures of association for three or four alleles. These are all zero when there is HWE, and then $\Delta_{AB} = D_{AB}$.

The is also a correlation aspect of composite linkage disequilibrium. If the **A** and **B** indicator variables x_{ij} and y_{ij} for the jth gamete of the ith sampled individual are defined earlier, then the sums $x_i = (x_{i1} + x_{i2})/2$, $y_i = (y_{i1} + y_{i2})/2$ are half the numbers of A and B alleles in the ith individual. They take the values 0, 0.5, or 1, and they have expectations

$$\begin{aligned} \mathcal{E}(x_i) &= p_A \quad \mathcal{E}(y_i) = p_B \\ \mathcal{E}(x_i^2) &= p_A \quad \mathcal{E}(y_i^2) = p_B \\ \mathrm{Var}(x_i) = \frac{1}{2}p_A(1-p_A)(1+f_A) &\quad \mathrm{Var}(y_i) = \frac{1}{2}p_B(1-p_B)(1+f_B) \\ \mathcal{E}(x_i y_i) = \frac{1}{2}(p_{AB} + p_{A,B}) &\quad \mathrm{Cov}(x_i, y_i) = \frac{1}{2}\Delta_{AB} \end{aligned}$$

The correlation between x_i and y_i is [7].

$$\rho_{AB_c} = \Delta_{AB}/\sqrt{p_A(1-p_A)(1+f_A)p_B(1-p_B)(1+f_B)}$$

Under HWE, this reduces to the correlation for indicator variables on the same gamete.

2.7 Testing for Linkage Equilibrium

For loci with two alleles, there are several equivalent ways of testing the hypothesis that the sampled population is in linkage equilibrium. Appealing to the asymptotic normality of the MLE for the linkage disequilibrium coefficient D_{AB}, the hypothesis that $D_{AB} = 0$ can be tested for with the standard normal statistic $z = \hat{D}_{AB}/\sqrt{\mathrm{Var}(\hat{D}_{AB})}$. Squaring this gives a statistic with a 1 df chi-square distribution:

$$X_{AB}^2 = 2nr_{AB}^2 = \frac{2n\hat{D}_{AB}^2}{\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)}$$

which uses the result that $\mathrm{Var}(\hat{D}) = p_A(1-p_A)p_B(1-p_B)/2n$ for a sample of $2n$ gametes when $D_{AB} = 0$. The same underlying assumption of a large sample size leads to this test statistic from a goodness-of-fit perspective: the four observed counts $O : n_{AB}, n_{Ab}, n_{aB}, n_{ab}$ are compared to the expected counts $E : 2n\tilde{p}_A\tilde{p}_B, 2n\tilde{p}_A\tilde{p}_B, 2n\tilde{p}_A\tilde{p}_B, 2n\tilde{p}_A\tilde{p}_B$ by means of the equation $X_{AB}^2 = \sum(O-E)^2/E$. It is also possible to perform an exact test using gamete counts.

For the composite linkage disequilibrium coefficient, there is the complication of the higher-order three- and four-allele disequilibria. By analogy to the test for gametic linkage disequilibrium though, the test statistic

$$X_{AB_c}^2 = 2nr_{AB_c}^2 = \frac{2n\hat{\Delta}_{AB}^2}{\tilde{p}_A(1-\tilde{p}_A)(1+\hat{f}_A)\tilde{p}_B(1-\tilde{p}_B)(1+\hat{f}_B)}$$

can be regarded as having (approximately) a chi-square distribution with 1 df [7]. There is not a goodness-of-fit route to this statistic as there was for HWE and for gametic linkage equilibrium.

2.8 Application to Data

Although genetic data are widely used throughout human genetics, it may be that the largest number of published data sets has been produced by forensic scientists. DNA profiles have proven to be of considerable use in issues of human identity determination, whether this is the forensic setting of associating a suspect with a crime, assessing the strength of evidence in a paternity dispute or in identifying remains after a mass disaster. Although the forensic uses of DNA used minisatellites in the early 1990s and the use of SNPs is now being investigated, it is microsatellites that are in common current use. As part of the process of validating these markers, forensic scientists publish analyses of data they collect to establish allele frequencies. On this occasion they publish the complete data, and one such set is from the FBI [2]. These data can be downloaded from the URL given in the Reference section, and they are also distributed, in .nex format, with the GDA package [8].

The data consist of genotypes at nine loci from six populations: self-identified as African-American (FBIA), Caucasian (FBIC), or Hispanic (FBIH) in the US, and from Bahamas (FBIB), Jamaica (FBIJ) and Trinidad (FBIT) in the Caribbean. The following results were obtained with GDA, but there are many other packages that could equally well be used.

Hardy–Weinberg Tests

The p-values shown in Table 1 were obtained by the exact test with 10,000 permutations. As expected, there is little evidence for departures from HWE although there

Table 1 HWE exact test p-values for FBI data

	Population					
Locus	FBIA	FBIC	FBIH	FBIB	FBIJ	FBIT
D3S1358	0.794	0.088	0.332	0.759	0.287	0.224
vWA	0.315	0.066	0.921	0.742	0.650	0.235
FGA	0.992	0.258	0.630	0.927	0.300	0.836
D8S1179	0.702	0.791	0.057	0.248	0.269	0.897
D21S11	0.492	0.414	0.647	0.019	0.922	0.915
D18S51	0.918	0.633	0.517	0.297	0.551	0.527
D5S818	0.443	0.576	0.525	0.259	0.770	0.973
D13S317	0.304	0.402	0.992	0.208	0.057	0.070
D7S870	0.487	0.324	0.421	0.192	0.257	0.370
Sample size	210	203	209	160	244	85

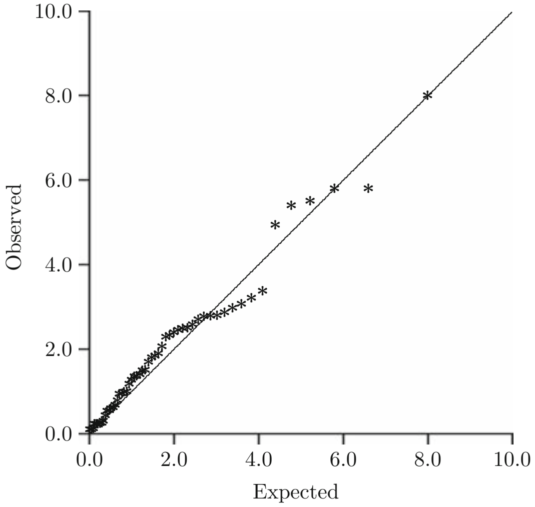


Fig. 1 Q-Q plot of HWE exact $-2\ln(p)$-values for FBI data

are some values less than the conventional $p = 0.05$ level. If all nine loci were in HWE in all six populations, we would expect to get two or three such "significant" values and we see there is only one. There appears to be no need for any corrections for multiple testing, although two procedures can be mentioned. The very conservative Bonferroni correction requires each p-value to be multiplied by the number of tests (with the products truncated at one) and those values used to assess significance. In this case, no value is anywhere close to being as small as 0.05. A better procedure (for independent tests) is to recognize that p-values are uniformly distributed when the hypotheses are true and construct a Q-Q plot of ranked p-values against their expected values. Because we have most interest in small p-values, we construct these plots for values of $-2\ln(p)$ which have a chi-square distribution with 2 df when the hypotheses are true. This transformation accentuates the small values, and leads to Fig. 1. The largest value of $-2\ln(p)$ is 7.93, for D21S11 in the FBIB sample, and this is seen to be not all unusual for a set of 54 tests.

Linkage Disequilibrium

To illustrate the similarities between gametic and composite linkage disequilibrium, we reduce the FBI data to two alleles at each locus: the most common allele vs. the rest. In Fig. 2, we show the estimated linkage disequilibrium coefficients when HWE is either assumed or not assumed – it is the composite linkage disequilibrium in the latter case. There is very little difference between the two estimates over the set of 216 values (36 pairs of loci for each of the six samples), as was expected because of the overall agreement with HWE. The composite coefficient requires

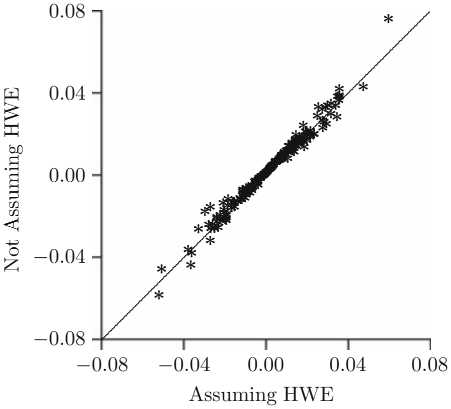


Fig. 2 Linkage disequilibrium estimates for FBI data

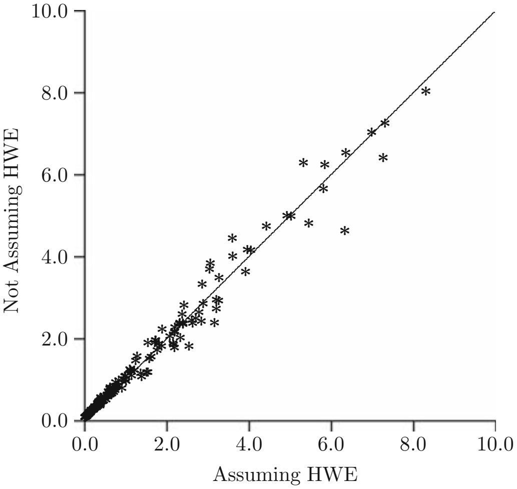


Fig. 3 Linkage disequilibrium test statistics for FBI data

very much less computation. The corresponding test statistics are shown in Fig. 3. Should any of the 216 test statistics be regarded as significant? The Q-Q plots for the two linkage disequilibrium test statistics are shown in Figs. 4 and 5, and there does not seem to be any substantial evidence for disequilibrium in any of the 216 tests.

The GDA package provides another approach to testing for association between pairs of loci. An exact test is constructed for the hypothesis that two-locus genotype frequencies are equal to the products of allele frequencies at both loci. For loci **A** and **B** the test statistic is:

$$\Pr(\{n_{ABijkl}\}|\{n_{Ai}\},\{n_{Bk}\}) = \frac{n!}{\prod_{i\leq j,k\leq l} n_{ABijkl}!}\frac{2^{H_A}\prod_i n_{Ai}!}{(2n)!}\frac{2^{H_B}\prod_k n_{Bk}!}{(2n)!}.$$

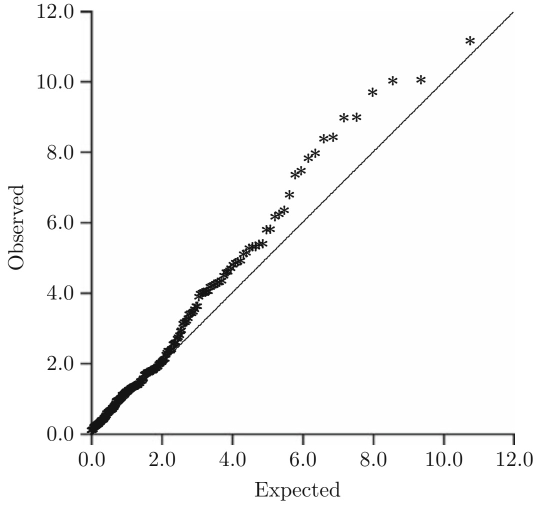


Fig. 4 Q-Q plot for linkage disequilibrium test statistic $-2\ln(p)$ values (assuming HWE) for FBI data

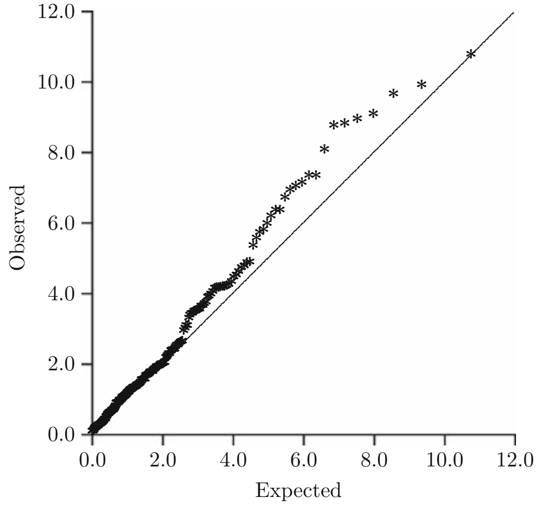


Fig. 5 Q-Q plot for linkage disequilibrium test statistic $-2\ln(p)$ values (not assuming HWE) for FBI data

Here, n_{ABijkl} is the count of $A_iA_jB_kB_l$ genotypes and n_{Ai}, n_{Bk} are the sample counts for A_i, B_k alleles. The hypothesis being tested is a composite of HWE at each locus, LD gametic and nongametic linkage disequilibrium and any other association between alleles taken three or four at a time. The Q-Q plot for the 216 tests on the FBI data is shown in Fig. 6. A goodness-of-fit test for this hypothesis would have 6 df.

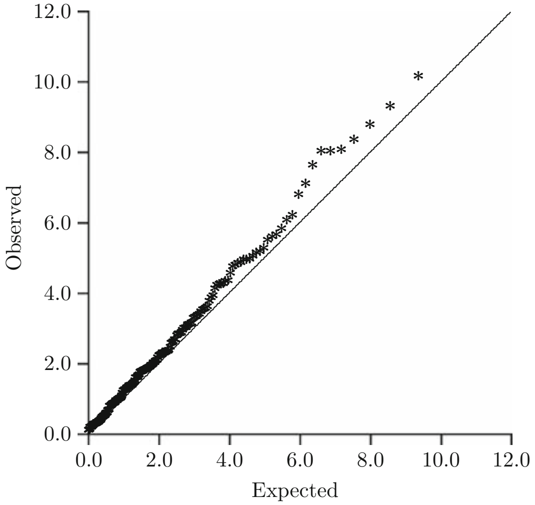


Fig. 6 Q-Q plot for two-locus allelic disequilibria test $-2\ln(p)$ values for FBI data

GDA also tests the hypotheses that multi-locus genotype frequencies are products of single-locus frequencies. The exact test statistic is:

$$\Pr(\{n_{ABijkl}\}|\{n_{Aij}\},\{n_{Bkl}\}) = \frac{n!}{\prod_{i\le j,k\le l} n_{ABijkl}!}\frac{\prod_{i\le j} n_{Aij}!}{n!}\frac{\prod_{k\le l} n_{Bkl}!}{n!}.$$

A goodness-of-fit test for this hypothesis would have 4 df. The Q-Q plot for the exact test shown in Fig. 7 is the first we have seen that shows the effect of the discrete nature of these test statistics. With the single-locus genotype counts fixed, there may be very few possible sets of two-locus genotype counts and therefore very few possible test statistic values of p-values. The continuous uniform distribution is no longer valid. The problem is less extreme when allelic counts are marginals, but to illustrate the problem consider the situation when the alleles are collapsed into the most frequent A, B vs. the rest a, b for **A**: D21S11 and **B:** D18S51 in the African-American FBI sample. The nine two-locus counts, with one-locus marginals are:

	BB	Bb	bb	**A** – Marginals
AA	0	4	2	6
Aa	2	21	42	65
aa	2	34	72	108
B – Marginals	4	58	117	179

The counts $AABB, AaBB, aaBB$ must each lie between 0 and 4, and there are only 15 possible values for the trio.

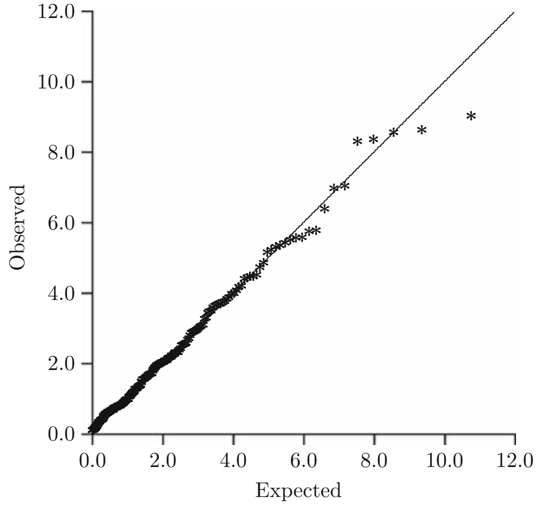


Fig. 7 $Q-Q$ plot for two-locus genotype disequilibria test $-2\ln(p)$ values for FBI data

3 Between-Population and Analyses

3.1 F-statistics

The use of correlation coefficients between pairs of alleles at the same or different loci is very useful for genetic analyses within populations. No appeal was made to evolutionary mechanisms in defining these quantities. When data are available from more than one population, it is possible to take into account the variation among populations caused by past evolutionary events. The major effect of considering the past is that individuals can no longer be considered independent – they are all affected by the history of their population.

The simplest model for analyzing population structure supposes that all populations have allele frequencies sampled independently from the same distribution. This sampling refers to the evolutionary process, not to actions of the investigator. It is sufficient to specify only the mean and variance of this distribution, but specifying the whole distribution allows for MLEs. Once again, it is convenient to develop statistical procedures by considering indicator variables. If there is random mating within populations, it may be sufficient to consider only allele frequencies instead of genotype frequencies, so define x_{ij} to be equal to 1 if the jth allele sampled from the ith population is of type A and is zero otherwise. The expected values of these variables introduce the population structure parameter θ (also written as F_{ST}):

$$\begin{aligned}\mathcal{E}(x_{ij}) &= p_A \\ \mathcal{E}(x_{ij}^2) &= p_A\end{aligned}$$

$$\mathcal{E}(x_{ij}x_{ij'}) = p_A^2 + p_A(1-p_A)\theta, \;\; j' \neq j$$
$$\mathcal{E}(x_{ij}x_{i'j'}) = p_A^2, \;\; i' \neq i$$

so that alleles drawn from different populations are independent. As $\mathrm{Var}(x_{ij}) = p_A(1-p_A)$ and $\mathrm{Cov}(x_{ij}, x_{ij'}) = p_A(1-p_A)\theta$, it is seen that θ is the correlation for alleles in the same population and so is analogous to the within-population inbreeding coefficient f. The difference between the two is that f refers only to the single sampled population, whereas θ refers to the average over the collection of populations of which the current population is but one member.

Allele A has frequency $\tilde{p}_{Ai}$ in the sample from the ith population, and if the sample size is n alleles, $\tilde{p}_{Ai} = \sum_{j=1}^{n} x_{ij}/n$. This leads to the variance of sample allele frequencies

$$\mathrm{Var}(\tilde{p}_{Ai}) = p_A(1-p_A)\left(\theta + \frac{1-\theta}{n}\right). \tag{6}$$

Whether or not there is HWE within a population, (1) shows the within-population variance of sample allele frequencies decreasing as the sample size increases. By contrast, the total variance in (6) never decreases below $p_A(1-p_A)\theta$ no matter how large a sample is taken. The evolutionary variation cannot be reduced by an investigator.

Manipulating the various expectations for the indicator variables suggests the use of two mean squares for estimating θ. If n_i alleles are sampled from the ith of r populations, we define mean squares for allele A among and within populations as

$$\mathrm{MSA}_A = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\bar{x}_{i.} - \bar{x}_{..})^2 = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\tilde{p}_{Ai} - \bar{p}_A)^2$$
$$\mathrm{MSW}_A = \frac{1}{\sum_{i=1}^{r}(n_i-1)}\sum_{i=1}^{r}\sum_{j=1}^{n_i} n_i(x_{ij} - \bar{x}_{i.})^2$$
$$= \frac{1}{\sum_{i=1}^{r}(n_i-1)}\sum_{i=1}^{r} n_i\tilde{p}_{Ai}(1-\tilde{p}_{Ai}),$$

where

$$\bar{x}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} = \tilde{p}_{Ai}\ , \ \bar{x}_{..} = \frac{1}{\sum_{i=1}^{r} n_i}\sum_{j=1}^{n_i} x_{ij} = \bar{p}_A.$$

Taking expectations, and writing $n_c = \left(\sum_{i=1}^{r} n_i - \sum_{i=1}^{2} n_i^2 / \sum_{i=1}^{r} n_i\right)/(r-1)$, provides

$$\mathcal{E}(\mathrm{MSA}_A) = p_A(1-p_A)[(1-\theta) + n_c\theta]$$
$$\mathcal{E}(\mathrm{MSW}_A) = p_A(1-p_A)(1-\theta)$$

and a moment-estimator of θ from allele A is

$$\hat{\theta}_A = \frac{\mathrm{MSA}_A - \mathrm{MSW}_A}{\mathrm{MSA}_A + (n_c - 1)\mathrm{MSW}_A}. \tag{7}$$

The large-sample mean and variance of the moment estimate is

$$\begin{aligned} \mathcal{E}(\hat{\theta}_A) &\approx \theta \\ \mathrm{Var}(\hat{\theta}_A) &\approx \frac{2\theta^2(1-\theta)^2}{r-1}. \end{aligned}$$

For loci with only two alleles, the estimate of θ is given by (7) since the same value would be found if sample frequencies of the alternative allele were used instead. With multiple alleles, assuming that the parameter θ is the same for all alleles

$$\hat{\theta} = \frac{\sum_A(\mathrm{MSA}_A - \mathrm{MSW}_A)}{\sum_A[\mathrm{MSA}_A + (n_c - 1)\mathrm{MSW}_A]}.$$

Assuming that θ is the same for all loci, the same equation can be used if A ranges over all the alleles at all the loci being considered. There is empirical evidence, however, [9] that this is not true and certainly natural selection would lead to differences among loci.

For large sample sizes the moment estimator becomes

$$\hat{\theta}_A \approx \frac{r\sum_{i=1}^r \tilde{p}_{Ai}(1-\bar{p}_A)}{\left[r\sum_{i=1}^r(\tilde{p}_{Ai}-\bar{p}_A)^2 + (r-1)\sum_{i=1}^r \tilde{p}_{Ai}(1-\tilde{p}_{Ai})\right]}$$

and for a large number of samples it reduces further to

$$\hat{\theta}_A \approx \frac{\sum_{i=1}^r(\tilde{p}_{Ai}-\bar{p}_A)^2}{r\bar{p}_A(1-\bar{p}_A)}. \tag{8}$$

It would be preferable to have a maximum-likelihood estimate of θ as this would allow more to be said about the properties of the estimate. In general, it is difficult to derive the distribution of allele frequencies over populations, but a useful approximation is to assume normality. For large sample sizes, where n_i can be taken to be large and equal, we assume

$$\tilde{p}_{Ai} \sim N[p_A, p_A(1-p_A)\theta].$$

If the r samples are independent, the likelihood for parameters p_A, θ is

$$L(p_A,\theta) = \prod_{i=1}^r \frac{e^{-\left(\frac{(\tilde{p}_{Ai}-p_A)^2}{2p_A(1-p_A)\theta}\right)}}{\sqrt{2\pi p_A(1-p_A)\theta}}$$

and the MLEs are

$$\hat{p}_A = \frac{1}{r}\sum_i \tilde{p}_{Ai} = \bar{p}_A$$

$$\hat{\theta} = \frac{\sum_i (\tilde{p}_{Ai} - \bar{p}_A)^2}{r\bar{p}_A(1-\bar{p}_A)}.$$

The second of these is the same as (8).

The multiple-allele version of the normal-based MLEs is

$$\hat{\theta}_l = \frac{1}{rm_l}\sum_{i=1}^{r}\sum_{u=1}^{m}\frac{(\tilde{p}_{liu} - \bar{p}_u)^2}{\bar{p}_u}$$

if there are m_l alleles at the locus. Combining over loci is just by averaging over loci. The MLEs are asymptotically chi-square distributed [10]. At a single locus

$$\hat{\theta}_l \sim \frac{\theta}{(r-1)(m_l-1)}\chi^2_{[(r-1)(m_l-1)]}.$$

This indicates that $\hat{\theta}$ is unbiased with variance $2\theta^2/[(r-1)(m_l-1)]$. Averaging over independent loci retains unbiasedness and decreases the variance to

$$\mathrm{Var}\left(\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_l\right) = \frac{2\theta^2}{L^2(r-1)}\sum_{l=1}^{L}\frac{1}{m_l-1}.$$

The chi-square result points to the asymmetric distribution of estimates about the mean unless the df are large and the chi-square tends to become normal. There is the most uncertainty about true values of θ when estimates are based on SNPs ($m = 2$) and small numbers of populations ($r = 2, 3, 4$).

The analysis of allele frequencies is appropriate when HWE can be assumed, but otherwise a distinction needs to be made between pairs of alleles within individuals and pairs of alleles from different individuals. It is then appropriate to define indicator variables x_{ijk} for the kth allele ($k = 1, 2$) in the jth individual ($j = 1, 2, \ldots, n_i$) sampled from the ith population. The expected values of these variables introduce the total inbreeding coefficient F (also writhen as F_{IT}) for alleles in the same individual in addition to θ for alleles in different individuals:

$$\begin{aligned}
\mathcal{E}(x_{ijk}) &= p_A \\
\mathcal{E}(x_{ijk}^2) &= p_A \\
\mathcal{E}(x_{ijk}x_{ijk'}) &= p_A^2 + p_A(1-p_A)F, \;\; k' \neq k \\
\mathcal{E}(x_{ijk}x_{ij'k'}) &= p_A^2 + p_A(1-p_A)\theta, \;\; j' \neq j \\
\mathcal{E}(x_{ijk}x_{i'j'k'}) &= p_A^2, \;\; i' \neq i
\end{aligned}$$

There are now three sources of variation, alleles within individuals, individuals within populations and among populations. If $\bar{H}_A = 1 - \sum_i n_i \tilde{P}_{AAi} / \sum_{i=1}^r n_i$ is the average over populations of the sample proportion of heterozygotes and $\bar{n} = \sum_{i=1}^r n_i$ is the average sample size, the three mean squares have these expectations

Source	d.f.	Sum of squares	Expected mean square
Populations	$(r-1)$	$2\sum_i n_i(\tilde{p}_{Ai} - \bar{p}_A)^2$ $= 2(r-1)\bar{n}s_A^2$	$p(1-p)[(1-F) + 2(F-\theta)$ $+ 2n_c\theta]$
Individuals in populations	$\sum_{i=1}^r (n_i - 1)$	$2r\bar{n}\bar{p}_A(1-\bar{p}_A) - \frac{1}{2}r\bar{n}\bar{H}_A$ $-2(r-1)\bar{n}s_A^2$	$p(1-p)[(1-F) + 2(F-\theta)]$
Alleles in individuals	$\sum_{i=1}^r n_i$	$\frac{1}{2}r\bar{n}\bar{H}_A$	$p(1-p)(1-F)$

Moment estimates of F and θ can be found from equating the mean squares to their expectations.

3.2 *Application to Data*

The GDA package implements the moment estimation procedure for population structure parameters. The estimates obtained for each allele at D3S1358 in the FBI data are shown in Table 2, firstly on the basis of genotypic data so that both F and θ can be estimated and then on the basis of allelic data so that only θ can be estimated. The within-population inbreeding coefficient f or F_{IS} is calculated as $f = (F - \theta)/(1 - \theta)$.

Table 2 Allele-specific F-statistics for D13S1358 in FBI data

	Not assuming HWE			HWE
Allele A	$\hat{f}_A(F_{IS})$	$\hat{F}_A(F_{IT})$	$\hat{\theta}_A(F_{ST})$	$\hat{\theta}(F_{ST})$
15.2	−0.000047	−0.000569	−0.000522	−0.000522
12	−0.001385	0.000860	0.000525	0.000521
<12	−0.001922	−0.000165	0.001754	0.001748
>19	0.000470	−0.000681	−0.001152	−0.001151
19	−0.006650	−0.006138	0.000509	0.000489
13	−0.009375	−0.004967	0.004367	0.004340
16	0.032688	0.037250	0.004716	0.004811
15	−0.026606	−0.009535	0.016628	0.016552
14	−0.035042	−0.024538	0.010149	0.010048
18	−0.053503	−0.027975	0.024231	0.024080
17	−0.045407	−0.039967	0.005204	0.005073
All alleles	−0.017504	−0.006497	0.010818	0.010768

Table 3 Locus-specific F-statistics for FBI data

	Not assuming HWE			HWE
Locus	$\hat{f}(F_{IS})$	$\hat{F}(F_{IT})$	$\hat{\theta}(F_{ST})$	$\hat{\theta}(F_{ST})$
D3S1358	–0.017504	–0.006497	0.010818	0.010768
vWA	0.001439	0.012150	0.010727	0.010731
FGA	–0.010243	–0.005117	0.005074	0.005043
D8S1179	–0.007740	0.006382	0.014013	0.013990
D21S11	0.000726	0.013279	0.012561	0.012563
D18S51	–0.004309	0.009942	0.014190	0.014177
D5S818	–0.004985	0.017716	0.022588	0.022574
D13S317	0.036872	0.062208	0.026306	0.026409
D7S870	0.016383	0.022371	0.006087	0.006132
All loci	0.000891	0.014273	0.013394	0.013396
Lower limit*	–0.007473	0.003858	0.009390	0.009363
Upper limit*	0.011991	0.028134	0.017939	0.017946

* 95% confidence interval by bootstrapping over loci

The very large variation among the allele-specific estimates at a locus mean that they are of little value, and attention is generally restricted to the estimates combined over alleles as shown in Table 3. Even these values, however, are very variable and the estimates combined over loci are likely to be of most value. Table 3 shows 95% confidence intervals generated by bootstrapping over loci and reported by GDA. In essence, nine loci are sampled with replacement from the data and new estimates are calculated. This is repeated 1,000 times and the central 95% of these values serve as a confidence interval. This procedure is computationally intensive and requires a large number of loci. An alternative is to make use of the chi-square approximation. The number of alleles at the nine loci are 11, 10, 28, 11, 23, 21, 11, 9, and 11 so $\sum_{l=1}^{L} 1/[(r-1)(m_l-1)] = 0.1537$ and the confidence interval for θ is approximately $\hat{\theta}(1 \pm 1.96\sqrt{2(0.1537)/81}) = (0.011788, 0.015004)$ which is a little narrower than that given by bootstrapping over loci.

Table 3 shows that there is little within-population inbreeding, as was already indicated by the non-significant HWE test results, but that F and θ are both different from zero. This is not surprising given the historical separation of the African and Caucasian populations. The relationship among populations can be explored further by estimating θ for each pair of populations separately and using the estimates as measures of genetic distances between the populations. Under a pure drift model of evolution $-\ln(1-\theta)$ is proportional to the time of the most recent common ancestral population of the populations under study [11]. GDA can calculate the pairwise estimates (Table 4) and use these to reconstruct the evolutionary history of the populations (Fig. 8). The African-American and Caribbean samples cluster together, as do the Caucasian and Hispanic samples.

Table 4 Population-pairwise estimates of θ for FBI data (θ above diagonal, $-\ln(1-\theta)$ below diagonal)

	FBIA	FBIC	FBIH	FHIB	FBIJ	FBIT
FBIA		0.014393	0.025581	0.000599	0.000554	0.004754
FBIC	0.014497		0.012169	0.013204	0.020197	0.007395
FBIH	0.025914	0.012243		0.022389	0.030719	0.014028
FBIB	0.000599	0.013292	0.022643		0.000278	0.002644
FBIJ	0.000554	0.020404	0.031201	0.000278		0.006082
FBIT	0.004765	0.007422	0.014127	0.002647	0.006100	

```
                       +---------------------------------------------FBIC
+----------------------10
|                      +---------------------------------------------FBIH
|
11                                                               +---FBIA
|                                                  +-------------8
|                                                  |             |+--FBIB
+--------------------------------------------------9             +7
                                                   |              +--FBIJ
                                                   |
                                                   +-----------------FBIT

|----------------|----------------|----------------|----------------|
0.0093           0.0070           0.0047           0.0023           0.0000
```

Fig. 8 UPGMA phenogram for FBI data

4 Discussion

This chapter has presented some basic methods in the statistical analysis of population genetic data. A much more complete account is given in the two-volume work edited by Balding et al. [12].

Acknowledgements This work was supported in part by NIH Grant GM 075091.

5 Web Resources

The data discussed in this chapter are available at

`http://www.fbi.gov.programs/hq/lab/fsc/backissu/`

`july1999/budowle.htm`

The package GDA which generates the numerical results in this chapter is available from

`http://lewis.eeb.uconn.edu/lewishome/software.html`

The X-prize for DNA sequencing is described at

```
http://genomics.xprize.org
```

References

1. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
2. Budowle B, Moretti, TR (1999) Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR-based loci. Forensic Science Communications 1999. Available at http://www.fbi.gov.programs/hq/lab/fsc/backissu/july1999/budowle.htm
3. Curie-Cohen M (1982) Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics 100:339–358
4. Guo SW, Thompson, EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372
5. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76:887–893
6. Weir BS, Hill WG, Cardon LR (2004) Allelic association patterns for a dense SNP map. Genet Epidemiol 24:442–450
7. Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 35:235–254
8. Lewis PO, Zaykin D (2001) Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html
9. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome Res 15:1468–1476
10. Weir BS, Hill (2002) Estimating F-statistics. Ann Rev Genet 36:721–750
11. Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779
12. Balding DJ, Bishop M, Cannings C (2007) Handbook of statistical genetics, 3rd ed. Wiley, New York
13. Weir BS (1996) Genetic Data Analysis II. Sinauer, Sunderland, MA

Haplotype Structure

Yu Zhang and Tianhua Niu

Abstract This chapter consists of five parts. In the first part, we provide definitions for important terms and concepts used in studies of population haplotype structures. In the second part, we introduce the user to valuable publicly available genotype/haplotype databases, such as databases generated by the International HapMap Project. In the third part, we provide concise guides to the user on how to download genotype data from the HapMap web site, how to use the Haploview program, as well as how to perform haplotype simulation. In the fourth part, we provide guides to several widely used haplotype inference Inference methods, including the Clark's algorithm, PHASE, HAPLOTYPER, and CHB. In the fifth part, we present to the user two popular software packages, LDhat and HOTSPOTTER, for estimation of recombination rates.

1 Population Haplotype Structure

1.1 Haplotype Block Structure in Human Populations

Based on empirical studies, the human genome can be viewed as a series of high linkage disequilibrium (LD) regions separated by discrete segments of very low LD [28, 30, 41]. Those genetic markers located within a high LD region are inherited from generation to generation essentially as a single unit. For example, Daly et al. [28] found that, a 500-kb region covering 103 single nucleotide polymorphisms (SNPS) on chromosome 5q31 could be partitioned into 11 haplotype blocks (99 SNPS were within these blocks, and four SNPS were outside the blocks). They found that within each block, two to four haplotypes account for at least 90% of haplotype variations in their sample [28]. In another study of SNPs located in a

T. Niu (✉)
Department of Psychiatry and Neurobehavioral Sciences, University of Virginia, 1670 Discovery Drive, Suite 110, Charlottesville, VA 22911,
e-mail: tn7b@cms.mail.virginia.edu

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_2, © Springer-Verlag Berlin Heidelberg 2010

216-kb region of the major histocompatibility complex (MHC) II complex in 50 British male sperm samples, Jeffreys et al. [34] revealed that recombination hotspots had caused block-like LD structures. These results lead to the conceptualization of haplotype blocks.

Haplotype blocks are defined as long stretches of DNA along a chromosome that have low recombination rates, which exhibit high LD and are characterized by relatively few haplotypes [39]. Furthermore, adjacent blocks are presumably separated by recombination hotspots, which are short regions with high recombination rates. Recombination hotspots (or coldspots) are defined as regions of the human genome with higher (or lower) recombination fractions than would be expected on the basis of the genome average recombination rate, 1 cM/Mb [27]. However, it should be noted that recombination hotspots (or coldspots) can also be defined relative to their local recombination rates. DNA segments that undergo more (or less) recombinations than their surrounding regions can also be defined as recombination hotspots (or coldspots). It should also be noted that the term recombination hotspots (or coldspots) can correspond to chromosomal segments that vary considerably in size. Popular software packages for identifying haplotype blocks include: HapBlock [47], HaploBlock [32], and HaploBlockFinder [46].

Although the presence of recombination hotspots can result in discrete haplotype blocks [28,30,31,43], a notion which appears to be supported by sperm typing studies of class II region of MHC [34], coalescent simulations demonstrate that a model assuming randomly distributed recombinations can also explain haplotype block-like structures [45]. Furthermore, by using the four-gamete test (FGT [14]) for defining haplotype blocks, Wang et al. [45] showed that the empirical chromosome 21 SNP dataset [41] is also congruent with a randomly distributed recombination model (i.e., without hotspots) with a varying recombination rate across the chromosome.

Fig. 1 shows a schematic diagram for an idealized haplotype block structure. This structure implies that a disease-causing mutation is often introduced on a specific haplotype background. Delineation of the haplotype block structure would help selecting a minimal set of haplotype-tagging SNPs (htSNPs) in searching for disease-causing mutations [35]. The selection of htSNPs not only ensures that the majority of haplotypic variations are captured but also dramatically reduces the genotyping cost in comparison with an exhaustive SNP coverage approach. It should be cautioned that recombination hotspot intensities vary such that haplotype block boundaries are often not sharp, and typically each hotspot corresponds to a genomic region of 1–2 kb in length [34].

1.2 Wright–Fisher Model

In the 1930s, both Ronald A. Fisher [18] and Sewall Wright [19] developed a stochastic model that allows a mathematical description of population reproduction. This model has become known as the Wright–Fisher model and is widely used

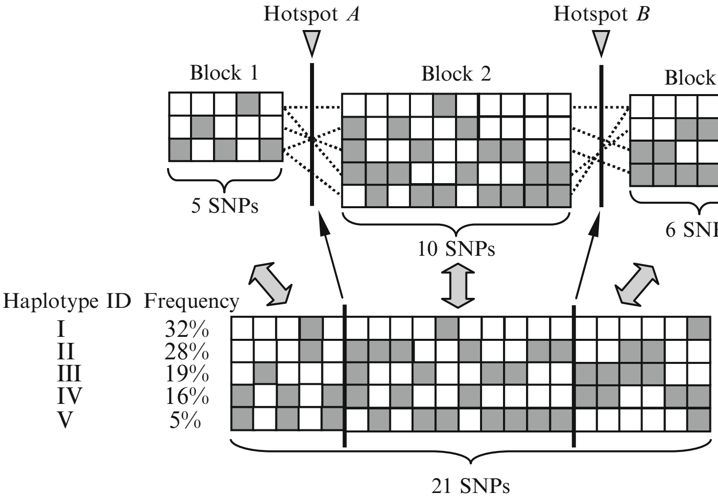


Fig. 1 Recombination disrupts the configurations of ancestral haplotypes when they are passed on from generation to generation. Each square represents a specificallele (*white*: wild-type allele; *gray*: variant allele) at a pahular SNP position. The entire haplotype encompasses 21 SNPs. The presence of two recombination hotspots (*A* and *B*) results in a block-wise structure of this region, which is composed of three discrete blocks, Blocks 1 (5 SNPs), 2 (10 SNPs), and 3 (6 SNPs), respectively. Recombination hotspots *A* and *B* reshuffle respective sub-haplotypes across the three blocks to create the overall block-like haplotype structure

in population genetic studies. The Wright–Fisher model is the canonical model of genetic drift in populations, which has the following assumptions:

1. Constant diploid population of size N ($2N$ alleles)
2. Synchronized and nonoverlapping generations
3. Random mating
4. No recombination
5. No selection
6. No migration to or from other populations and
7. Mutations are neutral and occur at a constant rate μ per generation

A schematic illustration of the Wright–Fisher model and the genealogical tree of two gene copies of the present generation are shown in Fig. 2.

The Wright–Fisher model is a simple binomial model of the amount of genetic randomness in a population of alleles created due to sampling. Assuming a haploid population size $2N$, the distribution of alleles can be described by a Markov model with binomial transition probabilities (for bi-allelic SNPs). More specifically, let X_i denote the number of a particular allele in the ith generation, then the distribution of the allele number in the next generation, X_{i+1}, can be expressed as

$$P(X_{i+1} = b|X_i = a) = \binom{2N}{b} \left(\frac{a}{2N}\right)^b \left(1 - \frac{a}{2N}\right)^{2N-b}.$$

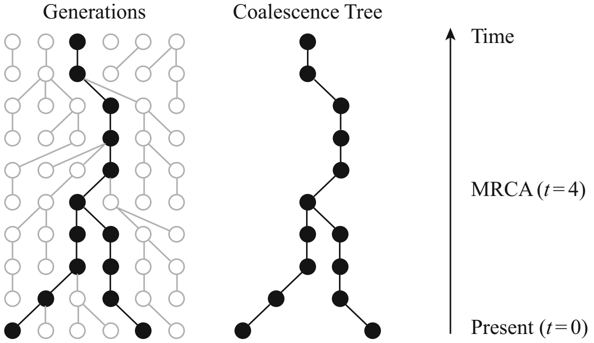


Fig. 2 The Wright–Fisher population model. *Left panel*: going backward in time, two randomly selected gene copies (*filled circles*) in the sample of the present generation trace back to the most recent common ancestor (MRCA) four generations ago. *Right panel*: the coalescence tree of two gene copies in the present generation

In fact, X_i is a Martingale bounded by 0 and $2N$. Mutations in the population can either become extinct ($X_i = 0$) or reach fixation ($X_i = 2N$) throughout generations, without the presence of selective forces. The phenomenon of allele frequency fluctuation under neutral conditions is called genetic drift.

Using the stopping time theorem for bounded Martingales, it can be shown that the probability that a newly arisen mutation eventually fixes and replaces the ancestral allele is $\frac{1}{2N}$. In fact, under the Wright–Fisher model, the fixation probability of any allele is simply the relative frequency of that allele in the population.

The probability that two individuals (each individual denotes a haploid gene copy) coalesce to a common individual in the parental generation is $\frac{1}{2N}$. This can be easily seen from the fact that the probability of two individuals sharing a particular parent is $\frac{1}{(2N)^2}$ and there are $2N$ individuals in the parental population. Generalizing this result, we get the distribution of coalescence time t of two individuals as

$$P(\text{coalescence in generation } t) = \left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right) \approx \frac{1}{2N}\,\mathrm{e}^{-\frac{t}{2N}}.$$

The mean coalescence time is therefore $E(T) = 2N$ generations with variance $\sigma_T^2 = 4N^2$.

1.3 Coalescent Theory

In the Wright–Fisher model depicted in Fig. 2, we are considering the simple case with only two gene copies. Kingman [37, 38] generalized the two gene copies to $n(n \geq 2)$ gene copies, realizing that when we look backward in time and change the discrete time scale to a continuous time scale (using exponential distributions instead of geometric distributions), we would have the "Kingman n-coalescent." The coalescent model is a standard population genetic model that allows us to construct and analyze random genealogies [42]. The development of coalescent theory has provided an important theoretical framework for capturing historical relationships among different gene copies, and indeed, over the past two decades, coalescent theory has revolutionized the field of molecular population genetics [42]. The basic coalescent operates under several assumptions that include constant population size, no selection, random mating, and no population structure [33]. Another assumption of the coalescent is that the sample size (n) is much smaller than the effective haploid population size ($2N$) of the population (i.e., $n << 2N$). A number of tests of the coalescent null model have been proposed, among them Tajima's D [44] and the statistics of Fu and Li [29].

In "Kingman n-coalescent," for a constant effective population size, $2N$, a genealogical tree is created for n gene copies. In any generation, we have

$$P(n \text{ copies coalesce to } n-1 \text{ copies per generation}) = \frac{n(n-1)}{4N}.$$

The expected time of a coalescent event is the reciprocal of this probability, that is,

$$E(T_{n \to (n-1)}) = \frac{4N}{n(n-1)}.$$

Simple induction reveals the expected time for n lineages to coalescence to their most recent common ancestor (MRCA) as

$$E(T_{\mathrm{MRCA}}) = 4N \sum_{i=2}^{n} \frac{1}{i(i-1)} = 4N\left(1 - \frac{1}{n}\right).$$

A special case is the expected coalescence time for two gene copies, which is $\frac{4N}{2(2-1)} = 2N$.

Note that in the coalescent process for a sample of size n, after the first coalescent event, we are faced with an identical coalescent process for a sample of size $n-1$. This lack of "memory" is called the Markovian property. It indicates that to follow the fate of lineages back in time all we need to know is the number of lineages currently active in the population. The reason we call coalescent as "Kingman n-coalescent" is to emphasize the dependence of coalescence time on sample size.

Kingman's recipe for constructing a genealogical tree of k gene copies is simply:

1. Go back a number of generations drawn from an exponential distribution with mean $\frac{4N}{k(k-1)}$.
2. Combine two randomly chosen lineages from the current sample.
3. Decrease k by 1.
4. If $k = 1$, construction is done. Otherwise goto step 1.

Mutations occur randomly at a rate proportional to the product of the time to coalescence and the mutation rate per generation. Because the genealogical process could be separated from the mutational process, we first build a genealogical tree of coalescence, and then add mutations to the genealogical tree. The results presented earlier have assumed the infinite-sites model of sequence evolution [36]. Assume that the mutation rate is μ per generation per gene, the number of mutations occurring during t generations is Poisson distributed with mean μt (the Poisson nature results directly from the independence of mutations). The expected number of differences between a pair of sequences for a diploid population of size N is $2\mu E[T_{\mathrm{MRCA}}] = 4N\mu$, which is often written as a single parameter: $\theta = 4N\mu$.

The standard coalescent theory does not allow for intragenic recombination. Generalization of the coalescent theory to include recombination events leads to a graph of lineages (rather than a tree), called the Ancestral Recombination Graph (ARG). Effects of recombination depend on the per gene per generation rate of crossing-over (i.e., genetic map length) and population size. Due to recombination, different genes located across the human genome may have different genealogies, and k lineages of a gene may split to $k + 1$ lineages in the parental generation. As a result, coalescence inference accounting for recombination events is computationally very complicated.

The strengths of the coalescent theory are (1) the coalescent is an enormously powerful and efficient way of looking at population genetic data; (2) the coalescent is fast and easy to simulate from and is very flexible; and (3) full likelihood analysis based on the coalescent theory uses all possible information in the data, and can be used to estimate the ages of mutations and the expected time taken to coalescence to the MRCA for gene copies of the present generation.

One weakness of the coalescent theory occurs when the fates of lineages depend on their allelic states (i.e., in the presence of selection) [40]. In addition, full likelihood analysis based on coalescence can be very computationally intensive (although this is an inherent feature for modeling population evolutionary histories), and devising an efficient algorithm is a challenging task [5, 23].

2 Public Genotype/Haplotype Databases

Large-scale genetic databases of human populations, containing data on genome-wide SNPs, genotypes, inferred haplotypes, mutation and recombination structures, have become publicly available. These databases provide rich information for understanding genetic variation patterns and for inferring evolutionary histories of human populations. In the following, we introduce two such resources where the genome-wide data are freely downloadable to researchers. We further introduce a haplotype simulation software that can simulate SNP data according to the coalescent theory, in which the user can specify settings allowing for both mutation and recombination events.

2.1 International HapMap Project

The International HapMap Project [25, 26] is a world-wide effort to identify and catalog genetic variants in human populations. It describes what these variants are, where they occur in the genome, and how they are distributed among individuals within and among populations distributed around the world. The web site of the International HapMap Project is available at **http://www.hapmap.org**.

The goal of the International HapMap Project is to compare the DNA sequences among individuals to identify chromosomal regions where genetic variants are shared. There are approximately ten million SNPs estimated to be present in the human genome (**http://www.hapmap.org/abouthapmap.html**). Testing all of these SNPs in chromosomes of individuals, however, can be extremely expensive and cost-inefficient. The development of the HapMap enables geneticists to take the advantage of how SNPs and other genetic variants are organized on the same chromosome.

The DNA samples for the International HapMap Project come from four populations with African, Asian, and European ancestries. DNA samples were collected from a total of 270 people: (1) YRI: the Yoruba population of Ibadan, Nigeria, provided 30 sets of samples from two parents and an adult child (each such set is called a trio); (2) JPT: 45 unrelated Japanese individuals from the Tokyo area provided samples; (3) CHB: 45 unrelated Han Chinese individuals from Beijing provided samples; and (4) CEU: 30 U.S. trios provided samples, which were collected in 1980 from U.S. residents with Northern and Western European ancestry by the Centre d'Etude du Polymorphisme Humain (CEPH) [25].

The goal of the International HapMap Project is to identify common haplotypes and to select so-called tag SNPs that uniquely identify those common haplotypes in human populations. The number of tag SNPs that capture most of the information of genetic variation patterns is estimated to be between 300,000 and 600,000, far fewer than the ten million common SNPs.

In the current phase of the project, approximately six million common SNPs have been genotyped for each of the four ethnic groups. The number of genotyped SNPs and the number of genotyped SNPs that passed the quality control (QC+) are summarized in Table 1 (HapMap Public Release #19):

The International HapMap Project can help researchers find functional regions that influence human health outcomes as well as responses to therapeutic drugs and environmental factors. The HapMap itself will not identify such regions directly. Instead, the HapMap provides a tool that can be used in both population-based and family-based disease association studies (topics presented in Chaps. 6–8).

Table 1 HapMap genotyping results (HapMap Public Release No. 19)

Populations	CEU	CHB	JPT	YRI
Total QC+ SNPs	3,901,408	3,903,524	3,902,623	3,806,910
Total Genotyped SNPs	5,894,684	5,812,990	5,812,990	5,857,466

Download Genotype Data from HapMap

The HapMap web site, located at **http://www.hapmap.org**, is an online resource for the display, retrieval, and analysis of the high-quality and high-throughput data generated by the International HapMap Project [25, 26]. The genome browser at the HapMap web site provides access to small-to-medium-sized regions of the genome for interactive exploration. In the following, we provide a step-by-step guide for downloading genotype data from the HapMap web site:

1. Goto **http://www.hapmap.org**.
2. Click on the "HapMap Genome Browser (Phase 1 & 2, B36)" link under "Project Data" on the left-hand side of the web site. This will lead the user to a genome browser based on the GBrowse package [48].
3. In the "Landmark or Region" field, enter the query term "IL10" which is the common name for interleukin 10, then click on the "Search" button to the right. Then, as shown in Fig. 3, the genome browser will display information about the region of the genome surrounding this gene. There are three panels: (1) an upper "Overview" panel providing a panoramic display of the whole chromosome with the region of interest indicated by a vertical yellow line; (2) a middle "Region"

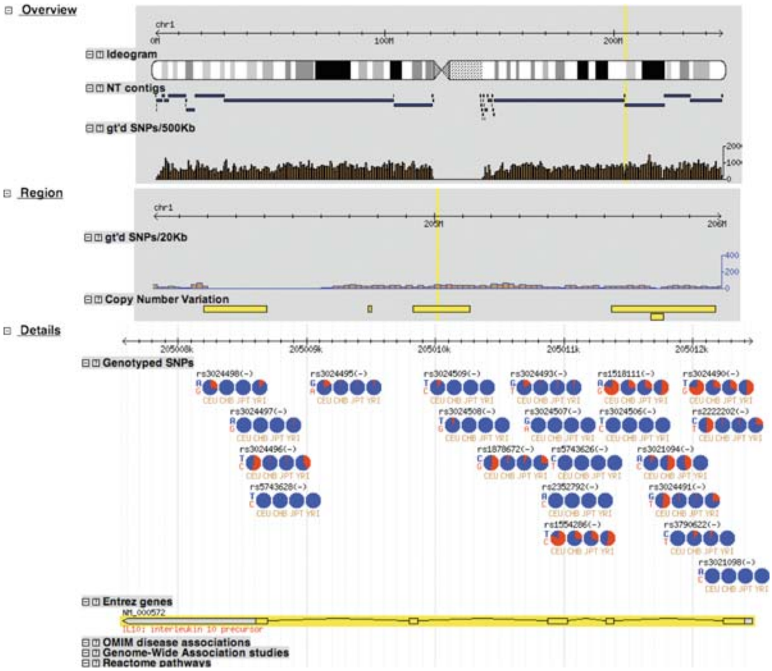


Fig. 3 HapMap Genome Browser displaying IL10 region. The genotyped SNPs track shows pie charts representing the allele frequency for each of the four genotyped HapMap populations. The *blue wedge* of the pie chart indicates the frequency of the allele that appears in the reference genome sequence. The *red wedge* is the frequency of the alternative allele

panel showing distributions of SNPs and copy number variations around the selected region; and (3) a lower ”Details” panel showing detailed information of SNPs, genes, and disease associations within the selected region. Under the three panels, there are several tracks options that, if selected, provide additional information about the region.

4. Choose the magnification at “Show 4.892 kbp” level. Find the “Reports & Analysis” menu and select the menu item “Download SNP genotype data.” Next, click on the “Configure” button. This will open a configuration page that allows the user to select the desired HapMap population, and whether to save the data to disk or view it in the web browser. Choose “CEU” for “Population” parameter, choose “rs” for “Strand” parameter, and choose “Save to Disk” for “Output format” parameter, and click on the “Go” button, and save the dumped data file as “dumped_region_IL10_CEU.txt”. Similarly, choose “YRI” for “Population” parameter, click on the “Go” button, and save the dumped data file as “dumped_region_IL10_YRI.txt”. These downloaded files known as “space-delimited text dumps” can be easily loaded into the Haploview program [49] for detailed analysis on the researcher’s local computer.

2.2 *The HapMap ENCODE Resequencing and Genotyping Project*

Another rich source of genetic information of human populations is provided by the HapMap ENCODE resequencing and genotyping project, which we will refer to as the HapMap ENCODE project hereafter. The term ENCODE stands for “ENCyclopedia Of DNA Elements” that aims to identify all functional elements in the human genome.

A total of 44 ENCODE regions were selected, consisting of 30-millionbase (Mb) with sizes ranging from 500 kb to 2 Mb each, roughly 1% of the human genome. These regions serve as a foundation on which to test and to evaluate the effectiveness and efficiency of a diverse set of methods and technologies for finding functional elements in human DNA. Half of the ENCODE regions were selected manually to include well-studied genes and known sequence elements that are located within conserved regions. The remaining regions were selected randomly based on a wide range of gene density and nonexonic conservation levels such that the selected ENCODE regions can be a good representative of the human genome.

Compared to the genotype data available from the International HapMap Project, the HapMap ENCODE project provides a much denser set of genotypes across large genomic regions. Ten 500-kb ENCODE regions of the genome were resequenced in 48 unrelated DNA samples from the International HapMap Project (16 YRI, 8 JPT, 8 CHB, and 16 CEU). All identified SNPs, either rare or common, were genotyped in 269 HapMap DNA samples (90 YRI, 44 JPT, 45 CHB, and 90 CEU). SNPs in the remaining 34 ENCODE regions in all of the HapMap DNA samples will be identified and genotyped.

Thanks to resequencing, genetic variants revealed by the HapMap ENCODE project are much less subjective to the ascertainment bias. Among the ten resequenced ENCODE regions, a total of 24,828 SNPs from the National Center for Biotechnology Information (NCBI) dbSNP and 6,256 SNPs outside the NCBI dbSNP were identified and genotyped in four HapMap ethnic groups. The average spacing of identified SNPs is less than 200 bp, much denser than that of the common SNPs provided by the International HapMap Project.

More details can be found at **http://www.hapmap.org/downloads/encode1.html.en**

2.2.1 Download ENCODE Genotype Data

The genotype data of the HapMap ENCODE project are distributed in the same way as the other HapMap genotype data. Readers can refer to the downloading procedures described in the HapMap section for how to download the HapMap ENCODE genotype data. Instead of entering a gene name or symbol, the user should enter the ENCODE region ID.

An alternative way to download HapMap ENCODE genotype data is called "bulk download." In the following, we demonstrate how to locate and download ENCODE datasets using "bulk download":

1. Goto **http://www.hapmap.org/downloads/encode1/html.en**.
2. Click on the "ENCODE genotype data dumps" link under "ENCODE Links" on the left-hand side of the web site. The link leads the user to a folder where all ENCODE genotype data are stored.
3. Click on the "nonredundant" folder and a list of ENCODE data files will appear.
4. Click on a file that the user want to download. For example, a file named "genotypes_ENm010.7p15.2_YRI.txt.gz" contains the ENCODE genotype data from the region "ENm010" of YRI samples. "7p15.2" indicates the chromosome location.
5. Save the file in the user's local machine. The suffix of the data file is ".gz," which means the file is zipped. For Unix/Linux users, the file can be unzipped using the command "gunzip genotypes_ENm010.7p15.2_YRI.txt.gz". For Windows users, the file can be unzipped by double clicking on it. As these files are in the same format of other HapMap genotype data files, they can be conveniently loaded into the Haploview program [49] for further analysis.

2.3 Haplotype Simulation

Stochastic population genetic models, such as the Wright–Fisher model, constitute an important class of models for the interpretation of molecular variation within populations. Samples drawn from a particular population contain both evolutionary variations and sampling variations. Statistical properties of such samples are often difficult to obtain from either analytical or numerical methods. A program that is able to simulate independent samples according to a stochastic population model can thus be very helpful in terms of studying statistical properties of such samples and evaluating different statistical methods.

A commonly used program to simulate population haplotypes is called **mksample**, developed by Hudson [17]. The program uses Monte Carlo techniques to draw haplotype samples from a population evolving according to the Wright–Fisher model. The program assumes an infinite-sites model of mutation and assumes a constant recombination rate. The program can simulate various evolutionary events such as gene conversion or symmetric migration among subpopulations.

The program assumes the standard coalescent approximation to the Wright–Fisher model under neutrality. The approximation works well as long as the sample sizes are small relative to the population size. For each sample (a set of haplotypes), the program first generates a random genealogical history for a segment of a chromosome. Conditional on the genealogy, mutations are randomly placed on each genealogical branch according to a Poisson process. Since the infinite-sites model is assumed, each mutation event gives rise to a new polymorphic site on the segment and no recurrent mutation occurs.

The output file of **mksample** is a sample of haplotypes, with each allele represented by “0” (ancestral allele) or “1” (mutated allele), respectively. The segment is mapped in the interval of (0,1), and the physical positions of polymorphic sites, where mutation events take place, are mapped on this (0,1) interval. The mutation parameter is $4N\mu$, where N denotes the effective diploid population size and μ denotes the neutral mutation rate for the entire segment being modeled.

The program is freely available for researchers from
http://home.uchicago.edu/∼rhudson1/source/mksamples.html

To simulate samples of haplotypes for a constant population size without cryptic population substructure, recombination, or gene conversion, one can simply specify the number of haplotypes to be collected in each sample, the number of samples to be produced, and the mutation parameter $4N\mu$. For example, by typing in the command line “ms 5 100 -t 3.0,” the program will simulate 100 samples of five haplotypes with $4N\mu = 3.0$.

To incorporate complicated scenarios, such as recombination, gene conversion, migration, or to change population size, additional options need to be used. We refer the user to the manual of **mksample** for details.

The first two lines of the output of **mksample** consist of the command line and the random number generator’s seed value. Following these two lines are the samples of haplotypes. Each sample is preceded by a line containing only “//,” followed by lines containing the number of simulated segregating sites, the position of each site on a scale of (0,1), and the configurations of haplotypes within the sample. For example, by typing in the command line “ms 4 2 -t 5.0,” one will get the following output:

```
ms 4 3 -t 5.0
18820382
```

```
//
segsites: 5
positions: 0.0227 0.5520 0.6190 0.9200 0.9459
10001
00010
00000
01100

//
segsites: 4
positions: 0.6760 0.7866 0.9056 0.9606
0101
1000
0101
0110
```

3 Haploview

Given the enormous amount of public genotype data such as HapMap and other data genotyped from genetic association studies, software tools for analyzing, interpreting, and visualizing these data are in pressing demand [49]. Here, we introduce a platform-free software, called Haploview.

3.1 What is Haploview?

Haploview is a software package that provides computation of LD statistics and population haplotype patterns from primary genotype data through a visually appealing and interactive interface [49]. The following procedures of usage are referring to Haploview version 4.1.

3.2 How to Download and Install Haploview

1. Download Haploview
   Haploview requires the installation of the Java Runtime Environment (JRE). The user can download the newest version of the JRE at: **http://www.java.com/**.
   After installing the JRE, goto the Haploview download web site:
   http://www.broadinstitute.org/haploview/haploview-downloads.

   (i) For Windows users, download the executable “hapinstall.exe” by a single click on the hyperlink “HapInstall.exe,” and save “hapinstall.exe” in a local directory.
   (ii) For Mac and Unix users, download the Haploview JAR file “Haploview.jar.”

2. Install Haploview

   (i) For Windows users, double click on "hapinstall.exe" saved in the user's local directory. Follow the installation instructions to install Haploview in the user's local computer.
   (ii) For Mac and Unix users, the program virtually needs no installation, and can be run by typing in the command line "**java -jar Haploview.jar**," or by clicking on the jar file.

3.3 How to Run Haploview

We illustrate how Haploview 4.1 can be run on: (1) HapMap Data and (2) non-HapMap Genotype Data.

3.3.1 How to Use HapMap Data in Haploview

1. Load HapMap Data
   Open Haploview. There are six buttons, "Linkage Format," "Haps Format," "HapMap Format," "HapMap PHASE," "HapMap Download," and "PLINK Format" on the left-hand side at the "Welcome to HaploView" window. Click on the "Load HapMap data" button. Load the file "dumped_region_IL10_CEU.txt" the user saved in Sect. 2.1. The user can leave the remaining options as the default.
2. Show LD Plots
   There are four tabs of the Hapoview v4.1, "LD Plot" tab, "Haplotypes" tab, "Check Markers" tab, and "Tagger" tab. The first view of the data is through the "Check Markers" tab. This provides a summary of the marker data, including columns for "#," "Name," "Position," "ObsHET," "PredHET," "HWpval," "%Geno," "FamTrio," "MendErr," "MAF," "Alleles," and "Rating." Among the 19 SNPs in the "dumped_region_IL10_CEU.txt" dataset, 7 SNPs had minor allele frequency (MAF) values shown in red because their values were below the default 0.0010 for the "Minimum minor allele freq." parameter.

   (i) Show a D' Plot: Click on the "LD Plot" tab, which presents a visual display of the LD statistic for the 12 polymorphic IL10 SNPs for the HapMap CEU population. The default block definition is the "Confidence interval (Gabriel et al.)" method [30] in the "Define Blocks" of the "Analysis" menu, which by default ignores marker(s) (in this case, rs3024508) with an MAF <0.05. The MAF and the confidence interval thresholds can be modified by selecting "Customize Block Definition" of the "Analysis" menu. The user will see the D' plot as shown in Fig. 4. Diamonds without numbers represent D' values of 1.0; all numbers represent the D' value expressed as a percentile. Bright red color represents LOD score for LD $\geq$ 2 and $D' = 1$, shades of pink/red

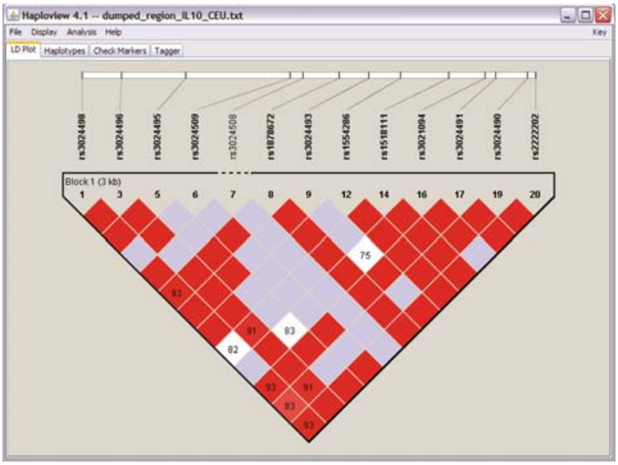


Fig. 4 LD (D') Plot of IL10 Gene SNPs for the HapMap CEU population

represents LOD ≥ 2 and $D' < 1$, blue color represents $D' = 1$ but LOD < 2, and white squares represent LOD < 2 and $D' < 1.0$.

(ii) Show an r^2 Plot: Staying at the "LD Plot" tab, goto "Display" menu, select "LD color scheme," and select "R-squared," and then, in the "Display" menu, select "Show LD values," and similarly select "R-squared," and the user may remove the D'-based haplotype block structure by selecting "Clear all blocks" in the "Analysis" menu. The user will see an r^2 plot as shown in Fig. 5. The r^2 plot is analogous to the D' plot shown in Fig. 4. The r^2 value is shown on a gray scale, where white color represents $r^2 = 0$, shades of gray represent $0 < r^2 < 1$, and black color represents $r^2 = 1$. SNP names and locations are the same as in Fig. 4.

3.3.2 How to Use Non-HapMap Genotype Data in Haploview

1. Prepare Data Files
   Both a "Data File" and a "Locus Information File" are needed for using non-HapMap Genotype data in Haploview.

   (i) Prepare a "Data File" in linkage format
   A "Data File" in linkage format is a data file that contains the genotype information. This data file should be in the Linkage Pedigree (pre MAKEPED) format, with columns of pedigree name, individual ID, father's ID ("0" if unknown), mother's ID ("0" if unknown), gender (1 = Male, 2 = Female), affection status ("0" = unknown, "1" = unaffected, "2" = affected), and genotypes (one for each allele, separated by a space, and coded either "$ACGT$" or

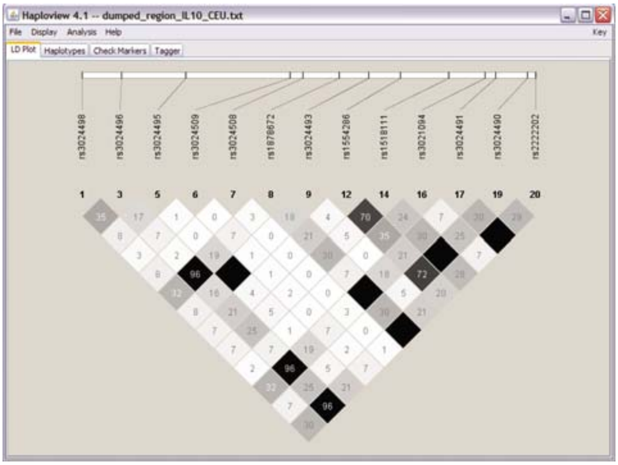


Fig. 5 LD (r^2) Plot of IL10 Gene SNPs for the HapMap CEU population

"1–4" where: "1" $= A$, "2" $= C$, "3" $= G$, "4" $= T$, and "0" indicates missing data). The file should not have a header line.

It should be noted that this "linkage format" data file can easily accommodate nonfamily based data by using a dummy value for the pedigree name and by filling in "0"s for father and mother IDs.

An input example of "Genotype file" named "Haploview_genotype_data.txt" has been generated using the **mksample** program (described in Sect. 2.3) under the neutral model and is shown as follows:

```
 1 1001 0 0 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1
 2 1002 0 0 2 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
 3 1003 0 0 2 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 2
 4 1004 0 0 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1
 5 1005 0 0 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 2
 6 1006 0 0 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
 7 1007 0 0 2 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1
 8 1008 0 0 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
 9 1009 0 0 1 1 1 1 2 2 2 2 2 2 2 2 1 1 1 1 1 1
10 1010 0 0 2 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1
11 1011 0 0 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
12 1012 0 0 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
13 1013 0 0 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1
14 1014 0 0 1 1 1 1 2 2 2 2 2 2 2 2 1 2 1 1 1 1
15 1015 0 0 2 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
16 1016 0 0 2 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
17 1017 0 0 2 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
18 1018 0 0 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1
```

```
19 1019 0 0 2 1 1 2 1 2 1 2 1 2 1 2 1 1 1 2 1 1
20 1020 0 0 2 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1
```

(ii) Prepare a "Locus Information File"
The "Locus Information File" consists of two columns, the first column designates "marker name" and the second column designates "position." The positions can be either absolute genome coordinates or relative positions in bp.
An input example of "Locus Information File" named "Haploview_locus_info.txt" has been generated under the uniform distribution model, and is shown as follows:

```
Marker01 1000
Marker02 1616
Marker03 1926
Marker04 2828
Marker05 3237
Marker06 4003
Marker07 4528
Marker08 5513
```

2. Load Data Files
In the Haploview program, select "Open new data" in the "File" menu.
Click on the "Linkage Format" button, which is the default setting. Next, the user will be presented with a dialog window that allows the user to choose input files and to select some options.
Load "Haploview_genotype_data.txt" file into the "Data File" field by clicking on the "Browse" button on the right-hand side and browse to the location of the file and load the file, and load the "Haploview_locus_info.txt" file into the "Locus Information File" field by clicking on the "Browse" button on the right-hand side and browse to the location of the file and load the file. The user can leave the remaining options as the default.
3. Check Markers
When the input files were successfully loaded, the first tab that appears is the "Check Markers" tab. Then, the user can modify the criteria for the inclusion of markers by changing the corresponding values of the parameters given. The user can choose "0.05" as the threshold value for "Minimum minor allele freq.," and can leave the remaining options as default. When the user can click on the "Rescore Markers" button, the checkmark to Marker06 in the "Rating" column would be dropped because the MAF was below the predefined threshold.
4. Show LD Plots
Click on the "LD Plot" tab, and a visual display of the LD statistic as well as the haplotype block structure for the eight polymorphic SNPs for the 20 subjects appears as shown in Fig. 6. The default definition for "Define Blocks" in the "Analysis" menu is the definition based on confidence intervals given by Gabriel et al. [30]. Similar to what has been described in Fig. 6, the user can generate an r^2 Plot.

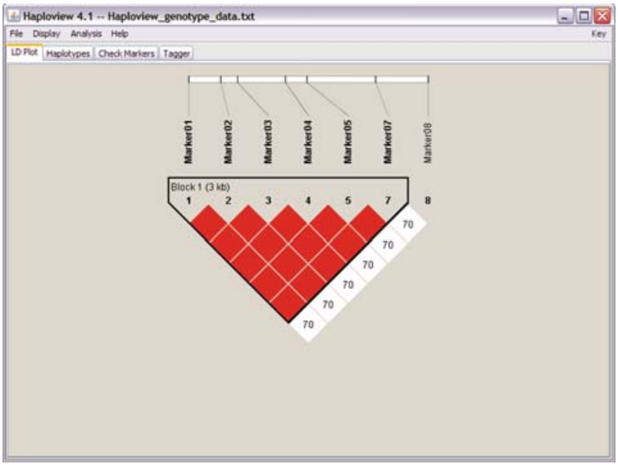
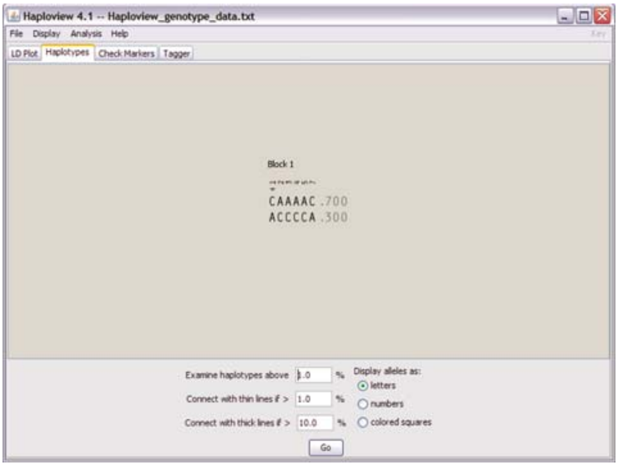


Fig. 6 LD (D') Plot of the genotype data simulated using the mksample program

5. Tag SNP Selection
   Haploview allows the user to select an optimized set of tag SNPs using a variant of the Tagger algorithm. The configuration page is shown by clicking on the "Tagger" tab.
   In the configuration page, the user may specify which alleles in this dataset to tag, which markers to "Force Include," and which markers to "Force Exclude." There are three options for tagging: (i) "pairwise tagging only"; (ii) "aggressive tagging: use 2-marker haplotypes"; and (iii) "aggressive tagging: use 2- and 3-marker haplotypes." The user can use the default setting, and choose the "pairwise tagging only" option. Then, the user can click on the "Run Tagger" button. The program will return a screen summarizing the tagging results.
6. Show Haplotypes
   Click on the "Haplotypes" tab, and goto "Display" menu, and select "Show tags in blocks." The user is presented with two haplotypes for Block 1 with their respective frequencies and the tag SNP (i.e., Marker01) indicated by the gray inverted triangle (Fig. 7).
7. Export Data and Images
   The "File" menu contains two options, "Export current tab to text" and "Export current tab to PNG" for exporting data to both text and PNG formats allowing the user to export the data contained in the current selected tab.

4 Haplotype Inference Methods

Here, we introduce several widely used haplotype inference methods. For each method, we first introduce its algorithm in terms of concepts as well as the underlying mathematical theories. We then describe how to prepare the input data file(s) for



Block 1

CAAAAC .700
ACCCCA .300

Examine haplotypes above 1.0 %
Connect with thin lines if > 1.0 %
Connect with thick lines if > 10.0 %

Display alleles as:
● letters
○ numbers
○ colored squares

Fig. 7 A display of block-based haplotypes, their recombination(s) (when applicable), and their respective frequencies

each program, how to run the program, what command options are available, and how to interpret the output files(s) for each program, through a common dataset A. The dataset A consists of genotypes at five loci for five individuals:

Locus positions (bp):	100 300 500 700 900
Individual 1	$AA\ TT\ GC\ AC\ AC$
Individual 2	$AG\ TA\ GC\ CC\ AC$
Individual 3	$AA\ TT\ CC\ AA\ CC$
Individual 4	$AA\ TA\ CC\ CC\ CC$
Individual 5	$AA\ TT\ GG\ AC\ AA$

We attempt to summarize the advantages and disadvantages associated with each method. All methods described here are available online.

4.1 Clark's Algorithm

Clark proposed the first computational algorithm [3] in 1990 to reconstruct haplotypes using multi-locus genotype data. Clark's approach is to assign the smallest number of haplotypes to explain the genotype data, based on the principle of parsimony.

Given a set of multi-locus genotypes $G = (g_1, \ldots, g_N)$ of N individuals, Clark's algorithm works through a convoluted updating procedure as follows:

1. Search for individuals whose genotypes, say g_i, can be uniquely resolved by a pair of haplotypes. In other words, we identify those individuals who either have

homozygous genotypes at all loci or have a heterozygous genotype at a single locus. Let H_1 denote the obtained pool of haplotypes in the first round.
2. At the kth round, given a haplotype pool H_k, search within the remaining unresolved individuals whose genotypes g_i can be resolved either by two haplotypes in H_k or by one haplotype in H_k plus a new haplotype. The haplotype pool H_{k+1} for the $(k+1)$th round is then constructed including both H_k and the new haplotypes.
3. Repeat step 2 until all individual genotypes are resolved or no more individuals can be resolved.

Albeit simple in nature, Clark's algorithm has been very popular and has generated meritorious results in the delineation of gene-based haplotype variations [6] and of the genome-wide LD in populations with different histories [4]. The disadvantage of Clark's method, however, is that the method may not even start if no individual genotypes can be unambiguously resolved (without manual intervention, H_1 can be empty). In addition, the phasing result depends on the order of individuals searched, which leads to undesired inference uncertainty attributed to the order. Motivated by Clark's method, many haplotype inference algorithms have been developed in the past decades that make use of more sophisticated models and theories.

Advantages:

1. Simple to implement.
2. Produces reasonably accurate results for datasets containing a limited number of common haplotypes.

Disadvantages:

1. The program may not even start when no phase unambiguous individuals are present.
2. The phasing results depend on the order of individuals scanned.
3. The method may not be practical for datasets genotyped at a large number of loci with low LD.

4.1.1 Software Usage

Availability:
http://linkage.rockefeller.edu/soft/list2.html#hapinferx

Command:
./hapinferx < input > output

The source code of the Clark's algorithm is written in FORTRAN 77. Note that the symbols "<" and ">" in the command line are not brackets, but are redirection signs in Linux shell.

Input:

To prepare the input file for **hapinferx**, the user should convert each single genotype vector to two haplotype vectors. For example, if the genotype vector of an individual at five loci is $\{AA\ TT\ GC\ AC\ AC\}$, we can represent this genotype vector by two haplotype vectors "00000" and "00111." Here, "0" and "1" represent the major and the minor alleles, respectively. Since the haplotype information is unknown, the order that at each locus which allele should be placed in which line is arbitrary. In addition, missing alleles can be represented by a character other than the two allele characters, such as "?."

In the input file, each haplotype vector takes one line, and there should be no space separating neighboring characters. In addition, each individual's data need to be preceded by an unique ID. The input file for the common dataset A is:

```
#1
00000
00111
#2
11000
00101
#3
00010
00010
#4
01000
00000
#5
00111
00101
```

The above input file only shows one of several possible representations for the same genotype data.

Output:

The output file consists of the following information:

1. A summary of input genotype data.
2. A list of individuals (and their haplotypes) whose genotypes are homozygotes at all loci.
3. A list of individuals (and their haplotypes) whose genotypes are heterozygous at only one locus.
4. A summary of haplotypes used to explain the genotype data, along with haplotype IDs.
5. Phasing results for all individuals. If an individual's genotype can be resolved by several alternative haplotype pairs using the Clark's algorithm, these possible haplotype pairs are listed.

The output file corresponding to the example input file is shown as follows:

```
#1                                                   ———part (1)
00000
00111
#2
11101
00000
#3
00010
00010
#4
00000
01000
#5
00101
00111

Homozygotes:                                         ———part (2)
#3
00010      1
00010      1
The number of homozygotes = 1
The number of distinct homo = 1

Single–site heterozygotes:                           ———part (3)
#4
00000      2
01000      3
#5
00101      4
00111      5

The number of single hets = 2
Count of unambiguous haplotypes = 5

List of unambiguous haplotypes:                      ———part (4)

. . . . . .

#mega
TITLE: test data
#1
00010
#2
```

```
00000
#3
01000
#4
00101
#5
00111
#6
11101
#1                                                  ———part (5)
00010       1          ——Option 1
00101       4            to Phase Individual 1
00000       2          ——Option 2
00111       5            to Phase Individual 1
  *
#2
00000       2
11101       6
  *
#3
00010       1
00010       1
  *
#4
00000       2
01000       3
  *
#5
00101       4
00111       5
  *
```

4.2 PHASE

According to the coalescent theory, haplotypes in the current generation are correlated with each other by sharing common ancestors in the past. Haplotype structures can therefore be clustered according to their configurations, e.g., one haplotype can be converted to another haplotype through a few steps of mutations and recombinations. Distinct from a parsimonious solution, a coalescence-based haplotype inference algorithm does not attempt to solve the problem using a minimum number of haplotypes, but to infer the most likely solution based on approximations to the coalescence process.

4.2.1 PHASE Algorithm

PHASE [6–8] is the first algorithm to infer haplotypes according to the coalescence process. The algorithm is built upon a Bayesian framework, which can be represented as:

$$P(H|G) = \frac{P(G|H)P(H)}{P(G)} \tag{1}$$

Here, G and H denote observed genotype data and unobserved haplotype data, respectively. By default, we use uppercase letters to represent vectors, i.e., the genotypes or haplotypes of all sampled individuals. The salient feature of the Bayesian inference is that, it provides a distribution of parameters of interest instead of just a point estimate, such that the variation of inference can be properly evaluated. In addition, missing values and model parameters can all be naturally incorporated into a likelihood model and simultaneously inferred.

Ideally, the relationships among haplotypes, according to the coalescent theory, can be built into the prior distribution $P(H)$ of haplotypes. Here, H denotes the parameter of interest, and $P(H)$ is called a prior distribution that captures the pre-knowledge of H before any observations were made. Coupled with the likelihood function $P(G|H)$ of genotypes, haplotype phases can then be reconstructed from the distribution $P(H|G)$, which is called a posterior distribution. The denominator $P(G)$, an unknown normalizing constant, can be ignored here.

PHASE reconstructs haplotypes from $P(H|G)$ using Markov Chain Monte Carlo (MCMC) techniques (see [1, 2]). The first version, PHASE v1.0 [6], only considers mutation events in the coalescence process, while later versions of PHASE (v2.0 [7] and v2.1 [8]) incorporated recombination events, and thus could infer haplotypes for regions with more complicated evolutionary histories across the genome.

Denote the genotype data of n individuals by $G = (g_1, g_2, \ldots, g_n)$ and a haplotype solution by $H = \{(h_{1a}, h_{1b}), (h_{2a}, h_{2b}), \ldots, (h_{na}, h_{nb})\}$. PHASE starts by randomly assigning haplotype pairs to individuals such that for individual i, the assigned pair (h_{ia}, h_{ib}) can explain the observed genotype g_i. The algorithm then iteratively updates $(h_{ia}, h_{ib}) \in H$ for each individual i from an easy-to-compute distribution $P(h_{ia}, h_{ib}|H_{-i}, G)$ conditional on haplotypes of the remaining individuals. After a sufficient number of iterations, samples of H will converge to the distribution $P(H|G)$. The algorithm then outputs the most frequently sampled haplotype pair for each individual, which is the most likely solution based on $P(H|G)$.

The updating scheme described above is called Gibbs sampling, a type of MCMC, given that $P(h_{ia}, h_{ib}|H_{-i}, G)$ is derived from the posterior distribution $P(H|G)$. Instead of employing the canonical version of Gibbs sampling, PHASE works on $P(h_{ia}, h_{ib}|H_{-i}, G)$ directly, which does not correspond to any joint likelihood of H. That is, PHASE defines

$$P(h_{ia}, h_{ib}|H_{-i}, G) = \pi(h_{ia}|H_{-i}, G)\pi(h_{ib}|h_{ia}, H_{-i}, G) \tag{2}$$

where

$$\pi(h|H_{-i}, G) = \sum_{\alpha \in E} \sum_{s=0}^{\infty} \frac{r_\alpha}{r} (\frac{\theta}{r+\theta})^s \frac{r}{r+\theta} (P^s)_{\alpha h} \tag{3}$$

Here, E denotes the set of all possible haplotypes, r_α denotes the number of haplotype α in H_{-i}, $r = \sum_{\alpha \in E} r_\alpha$ denotes the total number of haplotypes in H_{-i}, θ denotes the normalized mutation rates for the region, and P denotes a transition matrix for mutation events.

The underlying idea of the function $P(h_{ia}, h_{ib}|H_{-i}, G)$ is to assume that all individuals except i are drawn from the parental population of i. Based on the coalescent theory, a modern-day haplotype is either a split (coalescence when looked backward in time) or a mutant version of its parental lineage. A detailed explanation of the formula can be found in [5]. Later versions of PHASE modified π to further incorporate recombination events.

PHASE directly defines the function $P(h_{ia}, h_{ib}|H_{-i}, G)$ instead of deriving the function from a joint distribution $P(H, G)$. Although empirically working well, PHASE's performance lacks theoretical justifications underlying standard MCMC methods, such as the chain convergence to a proper posterior distribution. PHASE's updating scheme is therefore called pseudo-Gibbs sampling [6].

Advantages:

1. Produces accurate haplotype inference results, especially for datasets conforming to the coalescent theory.
2. Can infer haplotypes for SNP, microsatellite, and variable number of tandem repeat loci.
3. Can identify recombination hotspots and estimate recombination rates.
4. Can infer missing genotypes.

Disadvantages:

1. Inference may be significantly biased for datasets with complicated population structures and evolutionary histories [10].
2. The earlier versions of PHASE are computationally slow, particularly when dealing with a large number of individuals or loci, and when considering recombinations. Scheet and Stephens [9] recently developed fastPHASE that can handle hundreds of thousands of makers and thousands of individuals, at the cost of slightly reduced inference accuracy.
3. Large-sample theory does not justify the asymptotic consistency of PHASE's results, and the results are not formally interpretable from a Bayesian perspective.

4.2.2 Software Usage

Availability:
http://www.stat.washington.edu/stephens/software.html

Command:

```
./PHASE [options] input output
```

Input:

The **input** file is supplied by the user to specify the number of individuals to be analyzed, the number of loci genotyped, the type of each locus (SNP or microsatellite), and the genotype data for each individual. One can optionally specify the physical position of each locus to obtain an approximate of recombination parameters. Here, we use PHASE v2.1 [8] for illustration. An example input file for the common dataset A is shown as follows:

```
5
5
P 100 300 500 700 900
SSSSS
#1
0 0 0 1 0
0 0 1 0 1
#2
1 1 1 0 0
0 0 0 0 1
#3
0 0 0 1 0
0 0 0 1 0
#4
0 1 0 0 0
0 0 0 0 0
#5
0 0 1 1 1
0 0 1 0 1
```

The first two lines specify that there are five diploid individuals (first line) and five genotyped loci (second line). The positions of loci measured in bp are specified in the third line. The forth line specifies that all loci are SNPs, denoted by “S.” Microsatellite loci are denoted by “M.” The genotype data for each individual start from a label of that individual followed by the individual’s genotype data for all loci. There are two lines per individual for diploid organisms such as *Homo Sapiens*. Each line represents one chromosome copy for each individual. Since haplotype information is unknown, the order that at each locus which allele should be placed in which line is arbitrary. Alleles for both microsatellites and SNPs are represented by integers, e.g., SNP allele only takes two possible values, “0” and “1.” For missing alleles, use “−1” for microsatellite loci and “?” for SNP loci, respectively. Alternative input formats can be found in PHASE’s manual.

Options:

There are three main running options for PHASE v2.1: **-MS, -MR**, and **-MQ**. The **-MS** option only considers mutation events in coalescence. The **-MR** option considers both mutation and recombination events in coalescence. With more specified options, **-MR0, -MR1,...**, one can specify various recombination scenarios assumed in PHASE's model. The **-MQ** option is a hybrid of **-MS** and **-MR** options, which uses the faster **-MS** option for the preliminary computations, and then uses the **-MR** option for the final computations. Empirical results suggested that the **-MQ** option is roughly as accurate as the **-MR** option, but runs more quickly. However, the **-MR** option generally provides the best results and thus is the default method.

Here are a couple of notes, as stated from PHASE's manual:

1. In general, the **-MR** option gives less confident phase calls than those given by the **-MS** option, because of less stringent assumptions for coalescence.
2. For small datasets, say <20 individuals and <10 loci, the **MR** option may provide unreliable estimates for recombination parameters and may lead to unreliable inference results.

Output:

PHASE produces several output files. The reconstructed haplotypes are summarized in the main file, which has the user-specified name. Inference details such as estimated haplotype frequencies and recombination parameters are output to additional files each of which uses the user-specified name as its prefix.

The format of the main output file consists of the following:

1. A header containing the software version and credits to authors;
2. The command line used to run the program;
3. A list of all output files produced;
4. A summary of the input file;
5. A list of haplotypes included in the "best" haplotype reconstruction, with a summary of the frequency with which each haplotype occurred in the "best" reconstruction;
6. A list of the "best" guess of haplotype pair for each individual, where haplotypes are represented by their IDs presented in the previous haplotype summary list;
7. A detailed list of haplotypes for each individual with "()" at positions where the phase was difficult to infer, and "[]" at alleles which were difficult to infer (for missing alleles);
8. The same list of haplotypes for each individual but with those unambiguous loci masked; and
9. A list of the confidence probabilities associated with each phase call.

The main output file corresponding to the input file presented above is shown as follows:

```
**************************************************************
***               Output from PHASE v2.1.1               ****
***   Code by M Stephens, with contributions from N Li   ****
**************************************************************


BEGIN COMMAND_LINE
PHASE input output
END COMMAND_LINE
```

```
BEGIN OUTFILE_LIST
output_freqs : haplotype frequency estimates
output_pairs : most likely haplotype pairs for each individual
output_recom : estimates of recombination parameters
output_monitor : file for monitoring convergence
END OUTFILE_LIST

BEGIN INPUT_SUMMARY
Number of Individuals: 5
Number of Loci: 5
Positions of loci: 100 300 500 700 900
END INPUT_SUMMARY

List of haplotypes found in best reconstruction, with counts. (See file output_freqs
for haplotype population frequency estimates)

BEGIN LIST_SUMMARY
        1 00010 3.000000
        2 00000 1.000000
        3 00111 1.000000
        4 00101 2.000000
        5 01000 2.000000
        6 10101 1.000000
END LIST_SUMMARY

Summary of best reconstruction (numbers refer to the list of haplotypes given above)

BEGIN BESTPAIRS_SUMMARY
#1: (1,4)
#2: (6,5)
#3: (1,1)
#4: (2,5)
#5: (3,4)
END BESTPAIRS_SUMMARY

Haplotype estimates for each individual, with uncertain phases enclosed in "()" and
uncertain genotypes enclosed in "[]":

BEGIN BESTPAIRS1
0 #1
0 0 0 (1) 0
0 0 1 (0) 1
0 #2
(1) (0) 1 0 1
```

```
(0) (1) 0 0 0
0 #3
0 0 0 1 0
0 0 0 1 0
0 #4
0 0 0 0 0
0 1 0 0 0
0 #5
0 0 1 1 1
0 0 1 0 1
END BESTPAIRS1
```

Haplotype estimates for each individual, with uncertain phases enclosed in “()” and uncertain genotypes enclosed in “[]” with phase known positions indicated by “=”

```
BEGIN BESTPAIRS2
0 #1
= = 0 (1) 0
= = 1 (0) 1
0 #2
(1) (0) 1 = 1
(0) (1) 0 = 0
0 #3
= = = = =
= = = = =
0 #4
= 0 = = =
= 1 = = =
0 #5
= = = 1 =
= = = 0 =
END BESTPAIRS2
```

Phase probabilities at each site with phase known positions indicated by “=” and missing data positions indicated by “?”

```
BEGIN PHASEPROBS
= = 0.91 0.72 0.91
0.51 0.85 0.90 = 0.90
= = = = =
= 1.00 = = =
= = = 1.00 =
END PHASEPROBS
```

4.3 *HAPLOTYPER*

HAPLOTYPER [10] employs a full Bayesian model to infer multi-locus haplotypes. Similar to PHASE, HAPLOTYPER utilizes Gibbs sampling to iteratively update haplotypes H from a posterior distribution. Different from PHASE, HAPLOTYPER is built upon a proper joint likelihood function of H and G such that the results obtained by maximizing the posterior distribution $P(H|G)$ are readily interpretable. HAPLOTYPER does not assume any a priori population evolutionary model, such as the coalescent model. As a result, the method is robust when applied to populations with diverse evolutionary histories, such as past gene flows, stratifications, or bottlenecks [10, 50]. To infer haplotypes for a large number of linked loci, HAPLOTYPER introduces a technique called partition–ligation (PL) that effectively speeds up the computation for long-range haplotypes.

4.3.1 Bayesian Model

As before, let $G = (g_1, \ldots, g_n)$ and $H = \{(h_{11}, h_{12}), \ldots, (h_{n1}, h_{n2})\}$ denote the genotypes and haplotypes for n individuals typed at l loci. We use the notation $g = h_a \oplus h_b$ to represent that a pair of haplotypes (h_a, h_b) can explain the genotype g, called "compatible with g." Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_M)$ denote the population haplotype frequencies, where M denotes the number of all possible haplotypes. Suppose the Hardy–Weinberg equilibrium (HWE) holds true such that the population fraction of individuals with the ordered haplotype pairs (h_{ia}, h_{ib}) is $\theta_{h_{ia}}\theta_{h_{ib}}$. Then the likelihood function of G can be expressed as

$$P(G|\Theta) = \prod_{i=1}^{n} P(g_i|\Theta) = \prod_{i=1}^{n} \sum_{(h_{ia},h_{ib}):h_{ia}\oplus h_{ib}=g_i} \theta_{h_{ia}}\theta_{h_{ib}}. \qquad (4)$$

Assuming a Dirichlet prior distribution for Θ with parameters $\beta = (\beta_1, \ldots, \beta_M)$, the joint likelihood of (G, H, Θ) is

$$P(G, H, \Theta) \propto \prod_{i=1}^{n} \theta_{h_{ia}}\theta_{h_{ib}} \prod_{j=1}^{M} \theta_j^{\beta_j - 1} \qquad (5)$$

for a solution H compatible with G, and $P(G, H, \Theta) = 0$ otherwise.

To improve MCMC sampling efficiency, the parameter Θ is integrated out, and the joint likelihood of both G and H takes the form

$$P(G, H) \propto \prod_{i=1}^{M} \Gamma(c_i + \beta_i), \qquad (6)$$

where c_i denotes the count of haplotype $h_i \in H$.

Now, we have derived the full likelihood function $P(G, H)$, and thus the posterior distribution $P(H|G)$ (which only differs from $P(G, H)$ by an unknown normalizing constant). HAPLOTYPER employs the Gibbs sampling method to update H. In brief, the method starts from a random assignment of haplotype pairs for all individuals, and then iteratively updates $(h_{ia}, h_{ib}) \in H$ for each individual i based on genotypes G and the haplotypes of the remaining individuals. The haplotype updating step is based on a proper conditional distribution:

$$P(h_{ia}, h_{ib}|H_{-i}, G) \propto \begin{cases} (c_{h_{ia}} + \beta_{h_{ia}})(c_{h_{ib}} + \beta_{h_{ib}}), & h_{ia} \neq h_{ib} \\ (c_{h_{ia}} + \beta_{h_{ia}})(c_{h_{ib}} + \beta_{h_{ib}} + 1), & h_{ia} = h_{ib}, \end{cases} \tag{7}$$

which is derived from the joint likelihood function $P(G, H)$. Therefore, the haplotype samples output by HAPLOTYPER are guaranteed to converge to $P(H|G)$. The algorithm then outputs the most frequently sampled haplotype pairs for all individuals.

4.3.2 Partition–Ligation (PL)

The complexity of haplotype inference increases exponentially with respect to the number of heterozygous loci observed for each individual. To see this, consider a genotype at five SNP loci as $(A/A, B/b, C/c, D/D, E/e)$. Here, uppercase letters represent major alleles and lowercase letters represent minor alleles. There are three heterozygous loci: B/b, C/c, and E/e. Without additional information, four haplotype solutions are equally likely to explain the observed genotype: ($ABCDE, AbcDe$), ($ABCDe, AbcDE$), ($ABcDE, AbCDe$), and ($AbCDE$, $ABcDe$). In other words, the number of possible solutions is 2^{k-1} for genotypes with k heterozygous loci. The exponential complexity poses a great challenge in reconstructing haplotypes for a large number of linked loci. In particular, implementing a full Bayesian model accounting for all loci simultaneously can be difficult.

HAPLOTYPER proposed the PL technique to significantly reduce the computation complexity for inferring long-range haplotypes without sacrificing much of the inference accuracy. The PL technique works as follows:

1. Suppose a sequence of l loci are genotyped, choose an "atomistic unit" size d (typically $d \leq 8$), and partition both the genotype data G, and haplotypes H into $\lceil \frac{l}{d} \rceil$ subsets, each of size d. This is the partition step.
2. In the ligation step, first, reconstruct haplotypes for each "atomistic unit," and then ligate the solutions for sub-haplotypes for two adjacent units together. This is done by concatenating short sub-haplotypes comprising d loci for each atomistic unit into longer sub-haplotypes covering $2d$ loci. The M most probable sub-haplotypes are retained as candidates for the next inference step. The selected M sub-haplotype candidates are required to be self-sufficient to solve G.

3. The ligation step results in a new "atomistic unit" of size $2d$ and a set of M sub-haplotype candidates. Repeat step 2 until all loci of the region are ligated together.

An illustration of the PL algorithm is shown in Fig. 8. The underlying assumption is that sub-haplotypes reconstructed locally will be sufficiently close to the "true" haplotypes, and thus in each ligation step, the "true" answer is most likely included among the top M candidates. The PL approach can be implemented either hierarchically or progressively.

The computation time of the PL method is linearly related to the total number of loci. Coupled with this strategy, HAPLOTYPER is able to reconstruct haplotypes for a large number of loci. For example, the software can handle 100 individuals at 256 SNPs. The PL method was also adopted by many other haplotype inference algorithms, such as PHASE v2.1 [8] and CHB [11] (described in Sect. 4.4), and has played a pivotal role in large-scale haplotype reconstructions [51].

Advantages:

1. Robust to populations with gene flow, stratifications, and bottleneck effects.
2. The PL strategy enables rapid haplotype reconstruction for a large number of SNPs.
3. Sampling is based on a full Bayesian likelihood model such that asymptotic properties of the performance are guaranteed.
4. Can infer missing genotypes.

Disadvantages:

1. Handles SNP data only.
2. Does not take into account of any evolution events, such as mutation and recombination events.

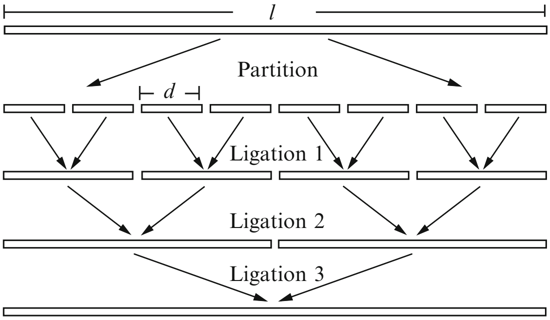


Fig. 8 A schematic illustration of the partition–ligation procedure. Parameter l denotes the total number of SNPs in the region and d denotes the number of SNPs in the initial atomistic unit. Modified from Fig. 1 in Niu et al. [10]

4.3.3 Software Usage

Availability:
http://www.people.fas.harvard.edu/∼junliu/Haplo/docMain.htm

Command:
./htyper input output locs inds iters

There are two programs for HAPLOTYPER: (1) **htyper** (the maximum number of SNPs allowed is 256, and the maximum number of individuals allowed is 100) and (2) **htyperv2** (the maximum number of SNPs allowed is 100, and the maximum number of individuals allowed is 500). Here, **htyper** is used for illustration of the usage of HAPLOTYPER.

Input:
The program will take the input genotype data from the **input** file, and save the inferred haplotypes in the **output** file. The user also needs to specify the number of loci typed (**locs**), the number of individuals typed (**inds**), and the total number of MCMC sampling steps (**iters**) required to run the program.

The format of the **input** file is shown as follows:

1. One line for each individual.
2. Each line contains l single digits coding for genotypes of l loci in their physical order.

The software uses the following coding scheme for all possible genotypes at a locus: $0 =$ "A/a," $1 =$ "A/A," $2 =$ "a/a," $3 =$ "?/?," $4 =$ "$A/?$," and $5 =$ "$a/?$." Here, "A" and "a" represent major and minor alleles, and "?" represents missing alleles. An example input file for the common dataset A is shown as follows:

```
11000
00010
11121
10111
11202
```

To infer haplotypes of this data, type "**./htyper input output 5 5 100**." We suggest to run the program for at least 100 iterations for small datasets, and more iterations for large datasets (e.g., datasets containing 100 individuals typed at >50 loci).

Output:
HAPLOTYPER outputs the inference result to the user-specified **output** file. The output file consists of two parts:

1. The "best" haplotype reconstruction. A pair of haplotypes for each individual is listed, along with their IDs indicated in the haplotypesummary list. The posterior

probability of each inferred haplotype pair is also listed, which can be used as a measure of the confidence of phase calls.

2. A summary of haplotypes based on the "best" reconstruction. The frequency and percentage of each haplotype contained in the "best" haplotype reconstruction are listed. If the number of loci is less than 20, a unique coding for each haplotype is assigned to help identify haplotypes of the same length. This feature can be helpful when one is doing a cross-platform comparison.

The output file corresponding to the above input data is shown as follows:

```
*************************************
*                                   *
*          Haplotyper Result        *
*                                   *
*************************************

0, 0.90000 — 1, 2
00010
00101
1, 1.00000 — 2, 3
00101
11000
2, 1.00000 — 1, 1
00010
00010
3, 1.00000 — 4, 5
00000
01000
4, 1.00000 — 2, 6
00101
00111

ID    Frequency    %          Haplotype
1      3          30.00000    00010 (2)
2      3          30.00000    00101 (5)
3      1          10.00000    11000 (24)
4      1          10.00000    00000 (0)
5      1          10.00000    01000 (8)
6      1          10.00000    00111 (7)
```

Notice that an unique code for each haplotype is included in "()," which is a decimal equivalent for the binary string. For example, The decimal equivalent of the binary string 11101 is $1 \times 16 + 1 \times 8 + 1 \times 4 + 0 \times 2 + 1 \times 1 = 29$.

4.4 CHB

CHB [11] is a haplotype inference method based on a Coalescence-guided Hierarchical Bayes (acronymed as CHB) model. In this model, a hierarchical structure is imposed on the prior haplotype frequency distributions so as to capture the similarities of modern-day haplotype configurations attributable to their common ancestry. Empirical results of CHB compared favorably to those of PHASE and HAPLOTYPER for both coalescence-simulated and empirical datasets with disparate evolutionary histories, with or without missing genotypes.

Ostensibly unrelated modern-day haplotypes are correlated through a coalescence process, because both mutation and recombination events can diversify the configurations of their shared ancestral haplotypes. CHB utilizes a hierarchical model in the prior distribution of modern-day haplotypes, $P(H)$, to capture the effect of the coalescence process. As demonstrated in several studies [6, 8, 11], the haplotype phasing accuracy can be substantially improved by incorporating coalescence. The model allows distinct modern-day haplotypes to have different a priori probabilities according to an inferred hierarchical structure, which results in a proper joint posterior distribution for all the parameters of interest.

4.4.1 CHB Model

As described in Sect. 4.3, when HWE holds true, the probability of a haplotype solution H compatible with genotypes G given the haplotype frequencies Θ can be written as:

$$P(G,H|\Theta) = \prod_{i=1}^{n} P(g_i = h_{ia} \oplus h_{ib}|\Theta) = \prod_{i=1}^{n} \theta_{ia}\theta_{ib}. \tag{8}$$

Here, $g_i = h_{ia} \oplus h_{ib}$ indicates that genotype g_i of individual i can be resolved by the haplotype pair (h_{ia}, h_{ib}), and n is the number of sampled individuals. Let c_j denote the number of haplotype h_j contained in H and assume a Dirichlet prior distribution of Θ with parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$, where M denotes the total number of possible haplotypes. We can integrate Θ out and obtain the joint distribution of $P(G,H)$ as:

$$P(G,H) = \frac{\prod_{j=1}^{M} \Gamma(c_j + \beta_j)}{\Gamma(\sum_{j=1}^{M} c_j + \beta_j)}. \tag{9}$$

The choice of $\boldsymbol{\beta}$ reflects our preknowledge about the frequency distribution of haplotypes in the modern-day haplotype pool. Without any specific information about these haplotypes, one (e.g., HAPLOTYPER) can let each parameter $\beta_j, \forall j \in [1, M]$ be equal to a constant. That is, each haplotype h_j is assumed to be equally likely to occur, a priori.

In CHB, however, modern-day haplotypes are assumed to be resulted from a coalescence process. CHB assigns different values to $\beta_j s, \forall j \in [1, M]$, corresponding to inferred ancestral haplotype frequencies. Thus, each haplotype h_j has a different a priori probability of occurrence. To illustrate the intuition of CHB, assuming that the modern-day haplotypes are descendants of an ancestral haplotype h_A, say, 100 generations back in time, then modern-day haplotypes would resemble h_A, i.e., differing only at a few loci. If we observe haplotype $h_1 = 0000$ in a majority of individuals, we would guess that this is the ancestral haplotype and that the probability of observing $h_2 = 0010$ in a future individual is greater than that of observing $h_3 = 0111$. Here, “0” and “1” denote the major and minor alleles at each SNP locus, respectively.

To account for the coalescence effect, let hyper-parameter $\Theta^* = (\theta_1^*, \cdots, \theta_M^*)$ denote the haplotype frequencies in a hypothetical ancestral population, from which haplotypes of currently sampled individuals are derived. Based on Θ^*, CHB first computes the expected frequencies $E(\Theta|\Theta^*)$ of modern-day haplotypes accounting for both mutation and recombination events. Then, let $\boldsymbol{\beta} = cE(\Theta|\Theta^*)$, which denote the parameters in the Dirichlet prior distribution for Θ. Here, c denotes a scaling constant, with a large c indicating more “confidence,” a priori, of the configuration of haplotypes. The hierarchical structure of the CHB model can be depicted as

$$\Theta^* \to E(\Theta|\Theta^*) \to \boldsymbol{\beta} \to \Theta.$$

A schematic diagram of the CHB model is given in Fig. 9.

The joint likelihood model of CHB is therefore:

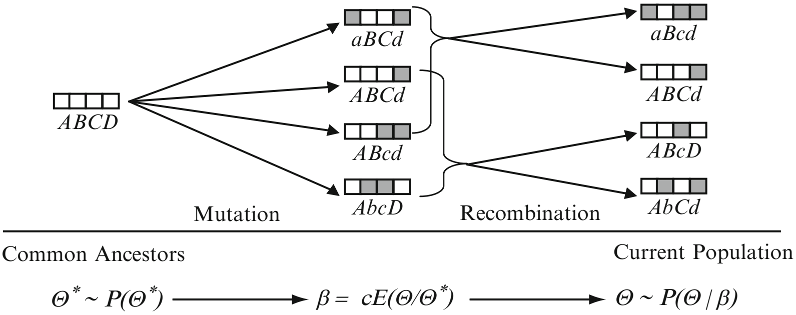


Fig. 9 A schematic diagram of CHB. Hyper-parameter Θ^* represents the frequencies of ancestral haplotypes from which the current individual samples are descended from. Assuming a robust “star-like” topology, a prior expectation of the modern-day haplotype frequencies, $E(\Theta|\Theta^*)$ is computed, taking into consideration of both mutation and recombination events. Each haplotype consists of four SNPs with uppercase letters (*white boxes*) indicating wild-type alleles, and lowercase letters (*gray boxes*), indicating mutant alleles. Modified from Fig. 1 in Zhang et al. [11]

$$
\begin{aligned}
P(G,H,\Theta^*,\gamma) = &\ P(G,H|\Theta^*,\gamma)P(\Theta^*)P(\gamma), \\
= &\ \frac{\prod_{j=1}^{M}\Gamma(c_j+\beta_j)}{\Gamma(\sum_{j=1}^{M}c_j+\beta_j)}P(\Theta^*)P(\gamma),
\end{aligned} \tag{10}
$$

where $\gamma = (\gamma_1, \ldots, \gamma_{l-1})$ denotes the vector for recombination probabilities for all pairs of adjacent loci.

4.4.2 MCMC Sampling and Convergence

By imposing a hierarchical structure on the prior distribution $P(H)$, we obtain a joint likelihood function which is very similar to the one used in HAPLOTYPER [10], with two additional parameters Θ^* and γ. Starting from random assignments of parameters H, Θ^*, and γ, CHB first updates the haplotype pair (h_{ia}, h_{ib}) for each individual i from the following conditional distribution based on the rest of parameters:

$$
P(h_{ia}, h_{ib}|H_{-i}, G) \propto \begin{cases} (c_{h_{ia}}+\beta_{h_{ia}})(c_{h_{ib}}+\beta_{h_{ib}}), & h_{ia} \neq h_{ib}, \\ (c_{h_{ia}}+\beta_{h_{ia}})(c_{h_{ib}}+\beta_{h_{ib}}+1), & h_{ia} = h_{ib}, \end{cases} \tag{11}
$$

Assuming H is given, CHB updates Θ^* and γ using the Metropolis–Hastings recipe [2]. Through iterations of this procedure, the posterior distributions of H, Θ^* and γ can be learned through iterative updating.

An important step when using MCMC methods for Bayesian inference is to check for the convergence of the MCMC sampling. One approach is to compare samples collected from several parallel chains [12]. CHB runs two chains in parallel from different starting points. Along with iterations, CHB monitors the ratio of the within-chain variation over the between-chain variation of log-likelihoods. When the ratio moves above a predefined threshold, it indicates that two chains may have converged to a common mode, and would only upon reaching that point, CHB starts to collect posterior samples.

In addition, CHB employs the PL technique described in Sect. 4.3 to handle those haplotypes for a large number of loci.

Advantages:

1. Can accurately and robustly infer haplotypes.
2. Can estimate recombination probabilities between adjacent loci.
3. Employs a proper Bayesian model such that asymptotic properties of the performance are guaranteed.
4. Can infer missing genotypes as well as frequencies of ancestral haplotypes.

Disadvantages:

1. Handles SNP loci only.
2. Computationally intensive when considering recombinations for a large number of loci.

4.4.3 Software Usage

Availability:
http://www.people.fas.harvard.edu/∼junliu/chb/index.htm

Command:
./CHB [options] input

Input:
The user needs to supply an **input** file that contains the genotype data for the program. The current version of CHB only supports datasets containing exclusively bi-allelic SNP data.

Let "A," "a," and "?" denote the major, minor, and missing alleles, respectively. CHB accepts two alternative input formats for genotype data:

1. For each individual, the genotype data are represented in a single line. Within each line, each single digit represents the genotype at each locus. No space separating neighboring digits is allowed. The coding scheme for each possible genotype is shown as follows:

Genotype	A/A	a/a	A/a	$A/?$	$a/?$	$?/?$
Code	0	1	2	3	4	5

An example input file for the common dataset A is shown as follows:

```
00222
22202
00010
02000
00121
```

2. For each individual, the genotype data are represented in two consecutive lines. In each line, each single digit represents the allele at each locus without any spaces separating the neighboring digits. The genotype at each locus is represented at the same position in the two consecutive lines. Since haplotype information is unknown, the order that at each locus which allele should be placed in which line is arbitrary. The same genotype data can then be presented as:

```
00010
00101
11100
00001
00010
00010
01000
```

```
00000
00111
00101
```

CHB can automatically distinguish between the above two input formats, given that there exists at least one heterozygous locus or one missing genotype in the input dataset.

Options:
The default CHB model only incorporates mutation events into the coalescence process, while recombinations are not accounted for. As a result, the default CHB is proper to infer haplotypes for a set of tightly linked loci where no recombination hotspots exist. Alternatively, the user can specify the **-r** option to let the program incorporate recombination events. When **-r** is specified, CHB simultaneously estimates the recombination probability for each pair of adjacent loci. The average performance of CHB under the **-r** option is better than the default when there are recombination hotspots present in the dataset. Without prior knowledge regarding the recombination hotspots *a priori*, the user can use the **-r** option to run CHB first, and check the output posterior means of recombination probabilities for all pairs of adjacent loci, of which values larger than 0.1 are often indicative of the presence of recombination hotspots. Note that the running time of CHB under the **-r** option is longer than that under the default setting.

Output:
The user can use the **-o output** option to specify the file name, where the inferred haplotypes will be output into. Under the default setting of CHB, the output file comprises two parts:

1. The "best" and alternative haplotype reconstructions, with three consecutive lines per individual. The first line starts with the individual ID, followed by the genotype data, and then the "best" haplotype pair. The posterior probability of the "best" haplotype pair is also provided, which can be used as a measurement of the "confidence" of phase calls. The second and the third lines consist of the inferred haplotype pair, with "0" and "1" coding for the major and minor alleles at each locus, respectively.
   In addition to the "best" haplotype calls, CHB also outputs alternative solutions if the posterior probabilities of alternatives are at least half of those of the best call. The "best" pair is enclosed in "[,]" and the alternative pair(s) are enclosed in "(,)."
2. A summary of haplotypes in the "best" or alternative reconstructions. Both the frequency and the percentage of each haplotype contained in the "best" reconstruction are listed. The posterior distributions of individual haplotypes are also included.

If the **-r** option is specified, CHB also outputs the posterior means and variances of recombination probabilities for all pairs of adjacent loci at the end of the output file.

As an example, the output file by typing in the command line **./CHB -r input** for the above input data is shown as follows:

```
CHB result

0:           00222 — [0,1]=0.631, (2,3)=0.360
             00010
             00101
1:           22202 — [4,5]=0.414, (1,6)=0.350
             01000
             10101
2:           00010 — [0,0]=1.000
             00010
             00010
3:           02000 — [2,4]=1.000
             00000
             01000
4:           00121 — [1,3]=1.000
             00101
             00111

ID    Count    %       Hap      Posterior
0:      3     30.0    00010       0.263
1:      2     20.0    00101       0.198
2:      1     10.0    00000       0.148
3:      1     10.0    00111       0.136
4:      2     20.0    01000       0.141
5:      1     10.0    10101       0.041
6:                    11000       0.035

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Recombination Frequencies:

Interval          Mean          Var
[1, 2]:          0.0090        0.0004
[2, 3]:          0.0235        0.0086
[3, 4]:          0.0452        0.0067
[4, 5]:          0.0091        0.0004
```

Notice that multiple phase calls for Individuals 1 and 2 are reported along with their posterior probabilities.

4.5 Comparison of Phasing Results

Because we used a common dataset for Clark's algorithm, PHASE, HAPLOTYPER, and CHB, we can compare the results across these methods. We observed that the inferred haplotype pairs for Individuals 3, 4, and 5 were the same for all four methods, because for these individuals, either all loci had homozygous genotypes or only one locus had heterozygous genotype. For Individual 1, Clark's algorithm and CHB provided the same two alternative phase calls, whereas PHASE and HAPLOTYPER only provided one phase call. In addition, for each individual, PHASE provided locus-specific posterior probabilities, while HAPLOTYPER and CHB provided posterior probabilities for haplotype pairs. The sole phase calls yielded by PHASE and HAPLOTYPER agreed to the "best" phase call by CHB. For Individual 2, Clark's algorithm provided a different phase call compared to the results of all other methods. The sole phase call by PHASE agreed with the "best" phase call by CHB, while HAPLOTYPER provided a different phase call which agreed to the runner-up yielded by CHB. The difference is attributable to the fact that both PHASE and CHB share the spirit of coalescence in their models, while HAPLOTYPER ignores any population structures.

5 Estimation of Recombination Rate

Recombination rate, measured by the expected number of recombination events occurring per unit length per meiosis, varies considerably across the human genome [4]. Inference on recombination rates can be achieved using summary statistics. For example, estimation methods based on the number of pairwise differences [13] or the minimum number of recombination events required [14] have been developed. Summary statistics, however, can be inefficient as they ignore some information contained in the data. Due to strong correlations across closely linked loci, merely increasing the sample size may not provide substantially more information about recombination. It is therefore imperative to use as much information contained in the data as possible. Here, we introduce two statistical methods that can estimate recombination rates and pinpoint recombination hotspots/coldspots.

A common dataset B containing 19 haplotypes at eight loci is depicted as follows:

Locus Positions (bp): 148 372 432 509 660 775 809 950

Freq.	haplotype configuration
1	11010000
1	01010111
1	11110100
14	00000111
2	11110111

5.1 LDhat

LDhat [20] estimates recombination rates using a composite likelihood approach. The method is an extension of the Hudson's approximate-likelihood method [21] based on the coalescent theory. Hudson's method only considers the two-locus scenario assuming an infinite-sites model of mutation. By contrast, LDhat considers multiple loci simultaneously assuming a finite-sites model of mutation, where the rate of recurrent mutation may be high. In addition, LDhat tests for the presence of recombination events based on population samples through permutations.

Under the Wright–Fisher model, the coalescent theory provides a statistical framework for modeling the genealogical history of sequences sampled from a large population. Within the framework, the effect of recombination is a function of the product of the recombination rate per unit length per generation, r, and the effective population size, N [22]. Without knowing one of parameters r or N, it is only possible to estimate $\rho = 4Nr$, known as the population recombination rate.

5.1.1 Composite Likelihood Estimation of ρ

LDhat assumes a simple model such that all sites in a sequence has two alternative alleles under a reversible and symmetric mutation model. According to McVean et al. [20], the estimation of ρ consists of four steps:

1. Estimate the population mutation rate per site, $\theta = 4N\mu$, from an approximate finite-sites version of the Watterson estimate:

$$\hat{\theta} = \left(\sum_{k=1}^{n-1} \frac{1}{k} \right)^{-1} \ln \left(\frac{L}{L-S} \right), \qquad (12)$$

   where n denotes the number of sampled sequences, L denotes the length of sequence analyzed, and S denotes the number of segregating sites.
2. Classify all pairs of segregating sites in the data into equivalent sets. For example, consider a sample of $n = 4$ sequences of length $L = 5$:

   $ACGAC$
   $ACTAC$
   $CCTAC$
   $ACTTC.$

   There are $S = 3$ segregating sites at positions 1, 3, and 4, respectively. The pair of sites 1 and 3 consists of the ordered allele combination set $\{AG, AT, CT, AT\}$ and the pair of sites 3 and 4 consists of the ordered allele combination set $\{GA, TA, TA, TT\}$. Using "0" and "1" to code for major and minor alleles, respectively, the two pairs of segregating sites (1,3) and (3,4) can be represented as $\{01, 00, 10, 00\}$ and $\{10, 00, 00, 01\}$, respectively. These pairs of sites are called "equivalent" because they all consist of two copies of "00," one "01," and one “10,” regardless of the order. The frequencies of distinct allele combinations at each pair of two sites in the sample are the only information used when computing composite likelihoods. By the same taken, the pair of sites (1,4) is also equivalent to the pairs of sites (1,3) and (3,4).

3. Estimate the likelihood of each equivalent set (of pairs of sites) given the estimated $\hat{\theta}$, under a finite-site, symmetric, reversible mutation model, for a range of recombination rates (by default, for $\rho = 0, 1, \ldots, 100$).

   The likelihood is calculated using the importance sampling method of Fearnhead and Donnelly [23], which attempts to sample the genealogical tree for a sample of sequences backward in time towards the MRCA of all sequences.

   Although the method of Fearnhead and Donnelly [23] is self-sufficient to estimate recombination rates, it is computationally intensive, even for only a moderate number of samples and multiple segregating sites. As a result, LDhat only uses their method to compute likelihoods for two-site cases. To improve computation efficiency, LDhat first precomputes a likelihood table for each possible equivalent set, and then uses the precomputated values to estimate recombination rates for various datasets of the same sample size. Theoretically speaking, LDhat is computationally feasible to estimate the recombination rates across the entire human genome.
4. A point estimate of the population recombination rate for the sampled sequences is obtained by combining the likelihoods of all pairs of segregating sites. Let $l(X_{ij}|4Nr_{ij})$ (introduced in [23]) denote the likelihood of the site pair (i,j). The composite likelihood of the entire sample is given by

   $$l_{\mathrm{c}}(4Nr) = \sum_{i,j} l(X_{ij}|4Nr_{ij}). \tag{13}$$

   Here, r_{ij} denotes the recombination rate of the segment between sites i and j, defined as

   $$r_{ij} = \frac{rd_{ij}}{L-1}, \tag{14}$$

   where d_{ij} is the physical distance separating sites i and j.

5.1.2 Likelihood Permutation Test

LDhat tests the presence of recombination events in a population sample using a permutation approach. The intuition of the test is that, under a model of no recombination and assuming a uniform mutation rate, sites are exchangeable. When recombination occurs, however, the physical ordering of sites matters. As a result, the likelihood of observing the sequences is dependent on the physical ordering of sites when recombination events are present in the data.

Based on this principle, the likelihood permutation test for recombination works as follows:

1. Compute the maximum composite likelihood as defined by Equation (13) for a sample of sequences. This leads to the estimation of $\rho = 4Nr$.

2. Permute segregating sites by their physical locations, and for each permutated dataset, compute the maximum composite likelihood.
3. The proportion of permutated datasets with a composite likelihood greater than or equal to that of the original dataset gives the estimated p-value. If the p-value is smaller than a predefined threshold, we conclude that there is statistical evidence for recombination in the sampled sequences.

Advantages:

1. Can efficiently estimate recombination rates if the lookup table of two-site likelihoods is precomputed.
2. Can handle either phased haplotype data or unphased genotype data, and allows for more than two alleles per site.

Disadvantages:

1. Estimation is often biased and only takes on a predefined set of discrete values.
2. Due to the usage of the composite likelihood, there is no standard statistical interpretation of the results.

5.1.3 Software Usage

Availability:
http://www.stats.ox.ac.uk/~mcvean/LDhat/

Command:
./program_name input locs [lookup_table]

The LDhat package is composed of several programs, two of which are for estimating recombination rates:

1. **pairwise**: estimates a constant recombination rate over a region. Either a crossing-over or a gene conversion model can be specified. Two files are required as the input, **sites** that contains the sequence data, and **locs** that contains the physical locations of SNPs. A lookup table of precomputed two-locus likelihoods can be supplied to speed up the program.
2. **interval**: estimates variable recombination rates over a region, using a Bayesian reversible jump MCMC scheme under the crossing-over model only. Due to the usage of the composite likelihood, the results cannot be interpreted from a Bayesian perspective. The same inputs are also required for the program **pairwise**, except that a lookup table is obligatory instead of optional.

Other programs contained in the LDhat package include **convert**, which converts sequence alignment data into appropriate formats for both **pairwise** and **interval**; **stat**, which summarizes the output of **interval**; **complete**, which generates lookup tables; and **lkgen**, which generates lookup tables from existing tables. These programs are optional in estimating recombination rates, and thus we refer the user to the LDhat manual for more details. In the following, we focus on the usages of **pairwise** and **interval**.

Input:

The user needs to supply the input data consisting of either haplotypes or genotypes in two separate files: **sites** and **locs**. The **sites** file contains sequence data at segregating sites in the FASTA format, except for the first line specifying the number of sequences, the number of sites and a flag, 1 or 2, indicating whether the input data consist of haplotypes or genotypes, respectively. The **locs** file contains information of the physical locations of segregating sites. The first line specifies the number of sites, the total length of the region analyzed, and a flag, "L" or "C," indicating whether to fit a crossing-over model or to fit a gene conversion model in the composite likelihood, respectively. The rest of the file contains the physical locations of sites in an ascending order. Examples of **sites** and **locs** files for the common dataset B are shown as follows:

sites file

```
19 8 1
> hap1
11010000
> hap2
01010111
> hap3
11110100
> hap4
11110111
> hap5
11110111
> hap6
00000111
......
```

(and other 13 haplotypes of the same type as "hap6")

locs file

```
8 1000 L
148 372 432 509 660 775 809 950
```

Haplotype data can be coded by either DNA letters "$A/C/T/G$" or numerals "$0/1/2/3$," with the ambiguous nucleotide, missing value, and gap coded by "N," "?," and "–," respectively. In the earlier example, we only used "0" and "1" to describe each haplotype, although "2" and "3" can be used as well. For genotype data, the convention is to use "0" and "1" to denote the two homozygotes for the major and minor alleles, respectively, "2" for heterozygote, and "?" for missing value, respectively.

Options:
By typing in the command line "**./pairwise sites locs [lookup_table]**," the user will be prompted for several options, a few of which are listed below:

1. *Use an existing likelihood file*: If the lookup table is not specified in the command line, the user will be asked whether or not to use an existing likelihood file to speed up the computation. Such likelihood files can be generated by programs **complete** or **lkgen**, and results are output into a file called **new_lk.txt**. Since generating the likelihood file is very time-consuming for large samples, some precomputed likelihood files are available from LDhat's web site.
2. *Sliding window analysis*: An estimation procedure can be carried out in a sliding-window fashion to estimate variable recombination rates across a region (although this option is largely superseded by the **interval** program). Results are output into a file called **window_out.txt**.
3. *Test for recombination*: Nonparametric permutation tests for the presence of recombination can be performed. Currently, the tests are only available for phased haplotype data. Results are output into a file called **rdist.txt**.

The **interval** program estimates variable recombination rates using a penalized likelihood within a Bayesian reversible jump MCMC scheme. By typing in the command line "**./interval sites locs lookup_table**," the user will again be prompted with several options, including the following:

1. *Block penalty*: The method works by fitting piece-wise constant recombination rates to the data, where a penalty is applied to the number of change-points. The user should try a series of different penalties ranging from 0 to 50.
2. *Number of updates for MCMC*: To ensure the convergence of MCMC sampling, the number of updates should be sufficiently large (e.g., one million). The user should also run the program using multiple starting points to check for convergence.
3. *Number of updates between samples*: This is the "thinning" parameter in MCMC that specifies how frequently we keep an update as one posterior sample. It is recommended by the authors of LDhat to sample for every 2,000–5,000 MCMC iterations.

Output:
The **pairwise** program generates several output files. We only discuss relevant output files.

1. **outfile.txt**: contains a point estimate of the constant recombination rate for the region and the composite likelihood value. An output file corresponding to the above input example is shown below. Four recombination rates ranging from 0 to 20 are tested, and the mutation rate is fixed at 0.01. The maximum likelihood is achieved at $\rho = 4Nr =$ 20.

```
Lk surface
Theta = 0.01000
Maximum at 4Nr = 5.000: Lk = −497.629
```

```
4Nr          Pairwise Lk
0.00          −504.130
5.00          −497.629
10.00        −498.652
15.00        −500.092
......
```

2. **window_out.txt**: is generated when the sliding window option is used. "SNP_L": the physical positions (in bp or in kb) of the leftmost site in the window; "SNP_R": the position of the rightmost site in the window; "#SNPs": the number of sites in the window; "4Nr/bp/kb": estimated recombination rate per bp or kb (depending on the unit assumed in the **locs** file) in the window; "CLR": log composite likelihood ratio using the estimated rate within the window and the average rate over the entire region; "Tajima's D": Tajima's D statistics for the window. An example output file for the same input data is shown as follows:

```
Sliding windows analysis – rho for total gene = 0.0500 per bp/kb

SNP_L     SNP_R     #SNPs    4Nr/bp/kb      CLR     Tajima's D
372.00    509.00      3        0.00000       3.28       0.764
509.00    660.00      2        0.02649       0.07      −0.680
775.00    950.00      3        0.00571       0.59      −1.126
```

The **interval** program generates the following two output files:

1. **rates.txt**: contains the posterior samples of recombination rates for all pairs of adjacent sites. The first line specifies the number of posterior samples recorded and the number of sites in the data. Following the first line are the posterior samples of recombination rates and the corresponding composite likelihoods, with one sample per line. An example is shown below, where the first column is the composite likelihoods, followed by recombination rates for 9 intervals for 10 linked sites.

```
50 8
380.254 0.014 0.031 0.043 0.473 1.445 2.843 0.267
30.330 0.003 0.005 0.022 0.065 0.112 0.091 0.014
......
```

2. **bounds.txt**: contains values of "1" and "0" indicating whether or not the recombination rate at each interval is different from its neighboring interval to its left. The format is the same as that of **rates.txt**, where the total number of rate changes for one sample is listed at the beginning of each line.

```
50 8
7      1 1 1 1 1 1 1
7      1 1 1 1 1 1 1
......
```

The results of **interval** can be further analyzed by the **stat** program, available from the LDhat package. The **stat** program summarizes the average, median, 2.5th and 97.5th percentiles of the estimated recombination rate for each pair of adjacent sites.

5.2 HOTSPOTTER

HOTSPOTTER [15] utilizes a statistical method to estimate recombination rates from population samples. The method is based on the coalescent theory that directly relates the patterns of LD of a population sample to the underlying recombination process. Recombination rates for all pairs of adjacent loci are estimated simultaneously, and the method can handle large genomic regions.

Recall that PHASE [6] reconstructs haplotypes from genotype data via conditional likelihoods, in which the effects of both mutation and recombination events are modeled (see Sect. 4.2). A similar approach is used in HOTSPOTTER, but the haplotypes are given as the input and the interest is to estimate the population recombination rate $\rho = 4Nr$ for each pair of adjacent loci. Here, N denotes the effective diploid population size and r denotes the recombination rate per unit length per meiosis. According to the coalescent theory, population samples contain information on the value of the product between N and r but not separately.

5.2.1 PAC Model

HOTSPOTTER relates the distribution of sampled haplotypes $(h_1, h_2, \ldots, h_n)$ to the underlying population recombination rate ρ by the following equation:

$$P(h_1, \ldots, h_n|\rho) = P(h_1|\rho)P(h_2|h_1, \rho) \cdots P(h_n|h_1, \ldots, h_{n-1}, \rho), \qquad (15)$$

where ρ denotes a vector of parameters because recombination rates can vary substantially along the genome [15]. The conditional probabilities on the right-hand side can be approximated by a series of conditionals (i.e., $\hat{\pi}$) such that an approximation of the joint distribution of haplotypes is given by

$$P(h_1, \ldots, h_n|\rho) \approx \hat{\pi}(h_1|\rho)\hat{\pi}(h_2|h_1, \rho) \cdots \hat{\pi}(h_n|h_1, \ldots, h_{n-1}, \rho). \qquad (16)$$

HOTSPOTTER refers this model as a “Product of Approximate Conditionals (PAC)” model, and denotes the right-hand side of (16) as the PAC likelihood, $L_{\text{PAC}}(\rho)$ [15]. HOTSPOTTER estimates ρ via maximum likelihood estimation (MLE), i.e., the estimator $\hat{\rho}_{\text{PAC}}$ takes a value that maximizes $L_{\text{PAC}}(\rho)$:

$$\begin{aligned} \hat{\rho}_{\text{PAC}} = & \ arcmax_{\rho} L_{\text{PAC}}(\rho) \qquad (17) \\ = & \ arcmax_{\rho}\{\hat{\pi}(h_1|\rho) \cdots \hat{\pi}(h_n|h_1, \ldots, h_{n-1}, \rho)\}. \end{aligned}$$

The validity of the PAC model critically depends on the choice of the approximate conditional $\hat{\pi}$. In a random sample of k haplotypes estimated from a population, $\hat{\pi}(h_k|h_1, \ldots, h_{k-1}, \rho)$ denotes the probability of the next sampled haplotype h_k conditional on the $k-1$ previously observed haplotypes $h_1, \ldots, h_{k-1}$, and the recombination rate ρ. According to the coalescence process, $\hat{\pi}$ should capture the following properties:

1. The next haplotype is more likely to match to a frequently observed haplotype than a rarely observed haplotype.
2. The probability of observing a novel haplotype decreases as the number of observed haplotypes increases.
3. The probability of observing a novel haplotype increases as the mutation rate increases.
4. The next haplotype tends to resemble the patterns of the observed haplotypes.

Recombination rates are determined by the patterns of haplotypes through property 4. In particular, the next haplotype either matches to one of the observed haplotypes, or takes a mosaic form of the observed haplotypes with a small number of mutations. The mosaic structure of the next haplotype is a result of recombination, such that the size of each small segment of the next haplotype is determined by the recombination rates of the region. The mosaic haplotype can be imperfect due to mutations. An illustration of the mosaic haplotype structure is shown in Fig. 10.

Since the mosaic structures of haplotypes are unknown, one needs to sum over all possible mosaic structures to compute the probability of observing h_k conditional on $h_1, \ldots, h_{k-1}$. HOTSPOTTER uses a hidden Markov model (HMM) to efficiently calculate $\hat{\pi}(h_k|h_1, \ldots, h_{k-1}, \rho)$.

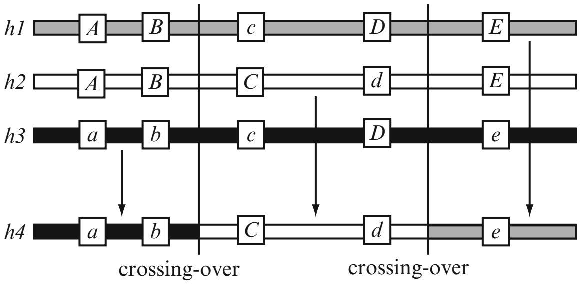


Fig. 10 Illustration of how $\hat{\pi}(h_k|h_1, \ldots, h_{k-1}, \rho)$ builds h_k as an imperfect mosaic of $h_1, \ldots, h_{k-1}$. Here, $k = 4$ and h_4 can be deemed as being created by "copying" from segments of h_1, h_2, and h_3. The arrow signs indicate the origins of segments. Each column of boxes contains letters representing alleles at that SNP locus, where uppercase and lowercase letters represent the two alternative alleles, respectively. The imperfect nature of the mosaic structure can be observed at the fifth locus, where the allele at the locus is mutated although the segment is "copied" from h_1. Modified from Fig. 2 in Li and Stephens [15]

5.2.2 Computing the Conditional Distribution $\hat{\pi}$

Let $h_1, \ldots, h_n$ denote the n sampled haplotypes typed at l SNP loci. Assume that the distribution of the first haplotype conforms to a uniform distribution, i.e., all 2^l possible haplotypes are equally likely, then $\hat{\pi}(h_1) = 1/2^l$. Consider now the distribution of h_k given $h_1, \ldots, h_{k-1}$ and ρ. As illustrated in Fig. 10, at each locus, h_k consists of a copy of allele from one of $h_1, \ldots, h_{k-1}$, with possible mutations. Let X_j denote the haplotype from which h_k copies the allele at locus j. Thus, X_j is the hidden state and $X_j \in \{1, 2, \ldots, k-1\}$. For the example shown in Fig. 10, we have $(X_1, X_2, X_3, X_4, X_5) = (3, 3, 2, 2, 1)$. To mimic recombination events, the distribution of $\{X_j\}$ can be described by a Markov chain with $P(X_1 = x) = 1/(k-1)$ and

$$P(X_{j+1} = x' | X_j = x) = \begin{cases} \mathrm{e}^{-\rho_j d_j/(k-1)} + \frac{1-\mathrm{e}^{-\rho_j d_j/(k-1)}}{k-1}, & x' = x, \\ \frac{1-\mathrm{e}^{-\rho_j d_j/(k-1)}}{k-1}, & x' \neq x. \end{cases} \tag{18}$$

Here, d_j denotes the physical distance between sites j and $j+1$, and $\rho_j = 4Nr_j$ where r_j denotes the average rate of crossing-over per unit physical length per meiosis between the pair of adjacent loci j and $j+1$. The transition probabilities capture the idea that, with small recombination rates (for a shorter distance or for a small value of ρ_j) between loci j and $j+1$, X_{j+1} is very likely to be the same as X_j.

HOTSPOTTER considers three kinds of recombination models:

1. Constant recombination rate model: $\rho_j = \alpha$ for all j
2. Single recombination hotspot model: $\rho_j = \gamma\alpha$ if the region between loci j and $j+1$ is the hotspot region, and $\rho_j = \alpha$ otherwise and
3. General variable recombination rate model: $\rho_j = \gamma_j \alpha$

Note that $P(X_{j+1}|X_j)$ designates the probability of the hidden state from which the segment of h_k at locus j is “copied” from, but not the probability of the observed alleles. Due to mutations, the probability of observing a particular allele a in h_k at locus j, called the emission probability, can be written as

$$P(h_{k,j} = a | X_j = x, h_1, \ldots, h_{k-1}) = \begin{cases} \frac{k-1}{k-1+\lambda} + \frac{1}{2}\frac{\lambda}{k-1+\lambda}, & h_{k,j} = a \\ \frac{1}{2}\frac{\lambda}{k-1+\lambda}, & h_{k,j} \neq a \end{cases} \tag{19}$$

where λ denotes the prespecified mutation rate per site normalized by the effective population size N.

Given both the transition probabilities (18) between states and the emission probabilities (19) of alleles, HOTSPOTTER can efficiently compute the conditional likelihood function $\hat{\pi}(h_{k+1}|h_1, \ldots, h_k, \rho)$ and thus the PAC likelihood $L_{PAC}(\rho)$ using the well-known Forward–Backward algorithm for Markov models. For an introduction to HMM, see [16].

Advantages:

1. Can estimate recombination rates based on the pattern of haplotypes via the PAC likelihood model.
2. Accommodates three different recombination models.

Disadvantages:

1. Estimation is biased and the results are empirically adjusted.
2. The PAC likelihood depends on the physical ordering of input haplotypes, and thus, the estimation may vary if haplotypes are given in different orders.
3. The input data must be phased haplotypes.

5.2.3 Software Usage

Availability:
http://www.biostat.umn.edu/∼nali/SoftwareListing.html

Command:

```
./program_name [options] input
```

There are three programs in the HOTSPOTTER package. They correspond to the three models for recombination rates described above, and thus the **program_name** can be one of the following:

1. **rholike**: constant recombination rate
2. **hotspot**: single hotspot and
3. **fullopt**: general variable recombination rate

To run HOTSPOTTER, two external libraries for C++ must be installed in advance:

1. **Boost**: a collection of C++ libraries, available from **http://www.boost.org**.
2. **GSL**: GNU Scientific Library, available from **http://www.gnu.org/software/gsl/**.

If the user does not have these packages preinstalled, these packages should be downloaded and installed before running HOTSPOTTER.

Input:
HOTSPOTTER requires the user to supply an **input** file that contains information on the number of haplotypes, the number of loci, the physical position for each locus, haplotype frequencies, and the configuration of each haplotype observed.

There are two possible input formats taken by HOTSPOTTER. An example input file for the common dataset B, in the first input format (i.e., the default format), is shown as follows:

```
5 8
0.148 0.372 0.432 0.509 0.660 0.775 0.809 0.950
```

```
1   11010000
1   01010111
1   11110100
14  00000111
2   11110111
```

The first line specifies that there are five different haplotypes typed at eight loci. The second line specifies the physical positions of the loci. The rest of the file contains the detailed haplotype information, with each line specifying the frequency and the configuration of each unique haplotype. The configuration of each haplotype is represented by a binary string 0's and 1's, with "0" and "1" denoting the major and minor alleles, respectively. HOTSPOTTER can only handle bi-allelic data and no missing data are allowed.

The second input format is the same format as used in Hudson's **mksample** program [17]. This can be handy when the user uses **mksample** to generate simulated haplotype data.

Options:
Three running options of HOTSPOTTER are often used (applicable to all programs): (1) the **-o outputfile** option: to specify where the result is saved into, otherwise the result will not be saved; (2) the **-f3** option: to specify whether the input file is in the second format (e.g., produced by **mksample**); and (3) the **-a** option: to specify whether to disable the bias correction step; by default, the bias in estimation is automatically "corrected." Additional options for each of the three programs are available, the details of which are skipped here.

Output:
For each program, as shown later, the output format slightly differs from each other.

1. **rholike**: estimates the background recombination rate. The output file corresponding to the earlier input example is shown as follows:

```
pop    rhobar     var       logLhat    CIlow   CIupp
 1     4.53531    1.76363   −50.716     −1      −1
```

   Here, "pop" denotes the population ID to distinguish the results if multiple population data are included in the input file; "rhobar" denotes the estimated background recombination rate $\hat{\rho}$; "var" denotes the estimated variance of $log\hat{\rho}$; "logLhat" denotes the log PAC likelihood at MLE; and "CIlow" and "CIupp" denote lower and upper bounds of the 95% confidence interval (CI), respectively.
2. **hotspot**: estimates both the background and the hotspot recombination rates. Thus, this program can be used to test for the existence of recombination hotspots, and the output file corresponding to the same input example is shown as follows:

```
pop  rhobar  maxLbar  left  right  hhat  vloghat   h.low  h.upp  rhat  vlogrhat  maxLhot
 1    4.54   −50.72    0     0.1   2.72  36227.7    −1     −1    4.54    1.76    −50.72
 1    4.54   −50.72   0.1    0.2   0.00  2524.71    −1     −1    4.85    1.81    −50.69
```

```
1   4.54   −50.72  0.2  0.3    0.00     34630.3   −1      −1    5.11   1.89   −50.68
1   4.54   −50.72  0.3  0.4    0.00     16051.3   −1      −1    5.50   1.79   −50.60
1   4.54   −50.72  0.4  0.5    0.00   −18039.6    −1      −1    6.92   1.66   −50.45
1   4.54   −50.72  0.5  0.6  575.61       2.54    4.04  7731.8  0.13   2.51   −50.04
1   4.54   −50.72  0.6  0.7  101.65       3.36    −1      −1    1.12   2.90   −49.83
1   4.54   −50.72  0.7  0.8  183.49       3.96    0.34    −1    0.93   3.39   −49.47
1   4.54   −50.72  0.8  0.9    0.00    4185.43    −1      −1    5.44   1.71   −50.61
1   4.54   −50.72  0.9  1      0.00     11288.6   −1      −1    4.96   1.73   −50.64
```

Here, "pop" denotes the population ID; "rhobar" denotes the estimated background recombination rate assuming no recombination hotspots; "maxLbar" denotes the log PAC likelihood using the value of "rhobar" assuming a constant recombination rate; "left" and "right" specify the physical boundaries of the region under consideration; "hhat" denotes the estimated "intensity" of the hotspot; "vloghhat" denotes the estimated variance of $log(hhat)$; "h.low" and "h.upp" denote lower and upper bounds of the 95% CI for "hhat," respectively; "rhat" denotes the estimated recombination rate assuming a hotspot; "vlogrhat" denotes the variance of "rhat"; and "maxLhot" denotes the log PAC likelihood using the value of "rhat" assuming the presence of a hotspot.

3. **fullopt**: estimates recombination rates assuming a general recombination model such that the recombination rate can vary. The output file corresponding to the same example is shown as follows:

```
i     position     rhobar     rhohat
0      0.148       4.5353      4.1299
1      0.372       4.5353      3.7584
2      0.432       4.5353      3.9613
3      0.509       4.5353      6.1379
4      0.660       4.5353      5.7190
5      0.775       4.5353      5.4367
6      0.809       4.5353      3.6570
7      0.950       4.5353      3.6570
```

Here, "i" denotes the ith inter-locus interval (from 0 to $l-1$, where l denotes the number of loci typed); "position" denotes the position of the ith locus; "rhobar" denotes the average recombination rate across all intervals; and "rhohat" denotes the estimated recombination rate for the ith interval.

Overall, both LDhat and HOTSPOTTER identified approximately the same recombination hotspot region between 0.509 and 0.660 kb, although the scales of estimated recombination rates were different.

Summary

Spurred by the International HapMap Project, interest in the assignment and analysis of haplotypes has increased immensely. In this chapter, we have summarized the key features of some widely used haplotype analysis methods. We have also provided terse guides regarding how the user can apply these software programs in their studies. We think these software programs will add tremendous value to the user for either disease-gene mapping or molecular evolution studies.

Web Resources

Clark's algorithm: http://linkage.rockefeller.edu/soft/list2.html#hapinferx
CHB: http://www.people.fas.harvard.edu/∼junliu/chb/index.htm
HAPLOTYPER: http://www.people.fas.harvard.edu/∼junliu/Haplo/docMain.htm
HAPLOVIEW: http://www.broadinstitute.org/haploview/haploview-downloads
HOTSPOTTER: http://www.biostat.umn.edu/∼nali/SoftwareListing.html
LDHAT: http://www.stats.ox.ac.uk/∼mcvean/LDhat/
MKSAMPLE: http://home.uchicago.edu/∼rhudson1/source/mksamples.html
PHASE: http://www.stat.washington.edu/stephens/software.html
The HapMap ENCODE website: http://www.hapmap.org/downloads/encode1.html.en
The International HapMap Project website: http://www.hapmap.org

Acknowledgements We thank Professor Jun Liu (Department of Statistics, Harvard University) for his encouragement and valuable discussions. This work was supported in part by National Institutes of Health grants R01 HG002518.

References

1. Gilks WR, Richardson S, Spiegelhalter DJ (eds) (1996) Markov chain Monte Carlo in practice. Chapman & Hall, London
2. Liu JS (2001) Monte Carlo strategies in scientific computing. Springer, New York
3. Clark A (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122
4. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204
5. Stephens M, Donnelly P (2000) Inference in molecular population genetics. J R Stat Soc B 62:605–655
6. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
7. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169
8. 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
9. 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
10. Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169
11. Zhang Y, Niu T, Liu JS (2006) A coalescence-guided hierarchical Bayesian method for haplotype inference. Am J Hum Genet 79:313–322

12. Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–472
13. Hudson RR (1987) Estimating the recombination parameter of a finite population without selection. Genet Res 50:245–250
14. Hudson RR, Kaplan R (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164
15. Li N, Stephens M (2003) Modeling linkage disequilibrium, and identifying recombination hotspots using SNP data. Genetics 165:2213–2233
16. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77:257–286
17. Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18:337–338
18. Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford.
19. Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159
20. McVean G, Awadalla P, Fearnhead P (2002) A coalescence-based method for detecting and estimating recombination from gene sequences. Genet 160:1231–1241
21. Hudson RR (2001) Two-locus sampling distributions and their application. Genet 159:1805–1817
22. Griffiths RC, Marjoram P (1996) An ancestral recombination graph. In: Donnely PJ, Tavare S (eds) IMA volume on mathematical population genetics. Springer-Verlag, Berlin, pp. 257–270
23. Fearnhead P, Donnelly PJ (2001) Estimating recombination rates from population genetic data. Genet 159:1299–1318
24. Romero R, Kuivaniemi H, Tromp G, Olson J (2002) The design, execution, and interpretation of genetic association studies to decipher complex diseases. Am J Obstet Gynecol 187:1299–1312
25. The International HapMap Consortium (2003) The international hapMap project. Nature 426:789–796
26. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
27. Arnheim N, Calabrese P, Nordborg M (2003) Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am J Hum Genet. 73:5–16
28. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232
29. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709
30. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229
31. Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–111
32. Greenspan G, Geiger D (2004) High density linkage disequilibrium mapping using models of haplotype block variation. Bioinformatics 20(Suppl 1):I137–I144
33. Hein J, Schierup MH, Wiuf C (2005) Gene genealogies, variation and evolution. Oxford University Press, London
34. Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222
35. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto–Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237
36. Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903
37. Kingman JFC (1982a) The coalescent. Stochastic Processes Applications 13:235–248
38. Kingman JFC (1982b) On the genealogy of large populations. J Appl Probab 19A:27–43

39. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4:981–994
40. Neuhauser C, Krone SM (1997) The genealogy of samples in models with selection. Genetics 145:519–534
41. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high–resolution scanning of human chromosome 21. Science 294:1719–1723
42. Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3:380–390
43. Schneider JA, Peto TE, Boone RA, Boyce AJ, Clegg JB (2002) Direct measurement of the male recombination fraction in the human beta-globin hot spot. Hum Mol Genet 11:207–215
44. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595
45. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71:1227–1234
46. Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. Bioinformatics 19: 1300–1301
47. Zhang K, Qin Z, Chen T, Liu JS, Waterman MS, Sun F (2005) HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. Bioinformatics 21:131–134
48. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610
49. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265
50. Niu T (2004) Algorithms for inferring haplotypes. Genet Epidemiol 27:334–347
51. Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasin GR, Donnelly P, International HapMap Consortium (2006) A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78:437–450

Linkage Analysis of Qualitative Traits

Mingyao Li and Gonçalo R. Abecasis

Abstract Linkage analysis of pedigree data is a powerful tool for mapping genomic regions that are likely to contain genes influencing human diseases. In this chapter, we will first introduce concepts and rationale of linkage analysis. Following this, we will then describe in detail two major types of linkage analysis strategies: model-based and model-free linkage analysis methods for qualitative traits. We will illustrate practical issues with linkage analysis by analysis of a real dataset collected from an age-related macular degeneration study. We will also describe how to identify the single nucleotide polymorphisms (SNPs) that account for linkage signal after linkage analysis is conducted. Finally, we will compare model-based and model-free linkage analysis methods and various software packages.

1 Introduction

Linkage analysis is an important step for initial localization of genetic variants that influence a trait of interest. Linkage refers to the phenomenon where two genetic loci cosegregate within families. Two loci are called genetically linked if the recombination fraction between them is less than $1/2$. The objective of linkage analysis is to estimate the recombination fraction and to test if it is less than $1/2$. In gene mapping studies of human diseases, the goal is to map a disease locus to a genetic locus, as represented by one or several genetic markers. Linkage analysis for qualitative traits requires (1) pedigrees – sets of individuals of known relationship, (2) marker genotypes – microsatellites or single nucleotide polymorphisms (SNPs), and (3) phenotypes – disease affection status.

As an initial step in positional cloning, linkage analysis is solely based on the known position of genetic markers. It has the advantage that no knowledge of the

M. Li (🖂)
Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA,
e-mail: mingyao@mail.med.upenn.edu

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_3, © Springer-Verlag Berlin Heidelberg 2010

gene function of the disease genes is required. With the availability of a large number of genetic markers throughout the human genome, linkage analysis has been successfully applied in identifying genes responsible for many human diseases, including Huntington's disease [1], Duchenne muscular dystrophy [2], and cystic fibrosis [3–5].

There are two major types of linkage analysis strategies: model-based linkage analysis and model-free linkage analysis. In this chapter, we will discuss each of these two types of linkage analysis methods in detail. We will also discuss how to identify which genetic marker explains the linkage signal after linkage analysis is conducted. Finally, we will present a real data example and illustrate the differences between model-based and model-free linkage analysis methods.

2 Model-Based Linkage Analysis

Model-based linkage analysis requires specification of the genetic model, as represented by the pattern of penetrances, i.e., the probability of developing the disease given genotype at the disease locus. Commonly assumed genetic models include multiplicative, additive, dominant and recessive models. To illustrate the basic procedures in model-based linkage analysis, we will first consider phase-known pedigrees. We will then consider a general case in which the phase may be unknown. We will also describe an efficient pedigree likelihood calculation algorithm, the Elston–Stewart algorithm (1971).

2.1 Phase-Known Pedigrees

Linkage phase is a key element in model-based linkage analysis. Linkage phase refers to the arrangement of alleles of linked loci on the same chromosome. It reveals which parental gamete is transmitted to the offspring and whether it is a recombinant or a nonrecombinant. For phase-known pedigrees, linkage analysis can be carried out by simply counting the number of recombinant and nonrecombinant gametes. If the majority of the gametes are nonrecombinant, then the two loci are probably linked.

Consider a simple case where the phase is known (Fig. 1). Individuals in this two-generation family are genotyped at two genetic markers with alleles A and a at the first locus and alleles B and b at the second locus. Since the linkage phase is known, we can easily count the number of recombinant and nonrecombinant gametes.

In this family, the father is doubly heterozygous (i.e., heterozygous at both loci), whereas the mother is doubly homozygous (i.e., homozygous at both loci). Among the four children, it is clear that the first two have inherited nonrecombinant gametes from the father, whereas the other two have inherited recombinant gametes from the father. Since the mother is doubly homozygous, we cannot tell whether the children

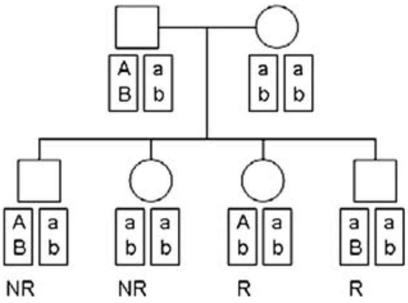


Fig. 1 Two-generation pedigree to illustrate linkage phase. All individuals are genotyped at the marker loci with alleles A and a at the first locus and alleles B and b at the second locus. NR means that the gamete is a nonrecombinant, and R means that the gamete is a recombinant

have inherited recombinant or nonrecombinant gametes from her. Therefore, in this example, the father is informative for linkage but the mother is not.

In general, let n denote the number of gametes that can be determined as recombinants or nonrecombinants, and r denote the number of recombinants. Then the likelihood function can be written based on the binomial distribution:

$$L(\theta) = \binom{n}{r} \theta^r (1-\theta)^{n-r}, \tag{1}$$

where θ is the recombination fraction between the two markers. The maximum likelihood estimate of the recombination fraction is $\hat{\theta} = r/n$.

After the recombination fraction is estimated, the next step is to test for linkage. The two marker loci are linked if the recombination fraction between them, θ, is less than $1/2$. Therefore, the hypothesis to be tested is

$$\mathrm{H}_0 : \theta = 1/2 \text{ vs. } \mathrm{H}_1 : \theta < 1/2.$$

This hypothesis can be tested by a likelihood ratio test. The likelihood ratio is defined as

$$LR(\theta) = \frac{L(\theta)}{L(\theta)\,|_{\theta=0.5}} = \frac{\theta^r (1-\theta)^{n-r}}{0.5^n}. \tag{2}$$

The maximized likelihood ratio statistic is $\mathrm{LRT} = \max_{\{0 \le \theta \le 1/2\}} 2 \log LR(\theta)$. Since the parameter θ hits the boundary of the parameter space under the null hypothesis of no linkage, thus the LRT statistic is asymptotically distributed as a 50:50 mixture of a chi-squared distribution with one degree of freedom and a point mass at 0. The corresponding p-value can be obtained by dividing the p-value from a full χ_1^2 distribution by 2.

For historical reasons, the maximum of the LOD (logarithm of odds) score,

$$\mathrm{LOD} = \max_{\{0 \le \theta \le 1/2\}} \log_{10} LR(\theta), \tag{3}$$

is typically reported over the likelihood ratio statistic. The maximum of the LOD score is often simply called the LOD score. There is a simple relationship between the LRT and the LOD score: $\mathrm{LRT} = 2\log(10)\,\mathrm{LOD} \approx 4.605\,\mathrm{LOD}$ and $\mathrm{LOD} = \mathrm{LRT}/[2\log(10)] \approx \mathrm{LRT}/4.605$.

So far, we have only considered linkage analysis between two genetic markers. In gene mapping studies of human diseases, the interest lies in identifying linkage between a disease locus and a genetic marker. To conduct linkage analysis in this case, the disease needs to be fully described with regard to the underlying genetic model, as specified by penetrances and allele frequencies at the disease locus. Unlike the analysis for two genetic markers, the genotypes for the disease locus are unobserved. Therefore, one has to test for linkage between the disease and the marker by reconceptualizing the disease affection status into a hidden genotype.

Suppose the unknown disease locus has disease allele D and normal allele d, and the marker locus has alleles M and m. Suppose the disease is autosomal dominant with complete penetrance, that is, an individual develops the disease given the disease genotype regardless of genotypes in other genes and environmental factors. For the three-generation pedigree in Fig. 2, we can unambiguously determine the genotypes of all individuals at the disease locus and figure out the linkage phase for the disease and marker loci.

In this family, all individuals are genotyped at the marker locus. To conduct linkage analysis between the unobserved disease locus and the marker locus, we need to determine the linkage phase first. Since the disease is autosomal dominant with complete penentrance, and the third generation includes both affected and unaffected individuals, we can tell that the father's disease locus genotype must be *Dd*.

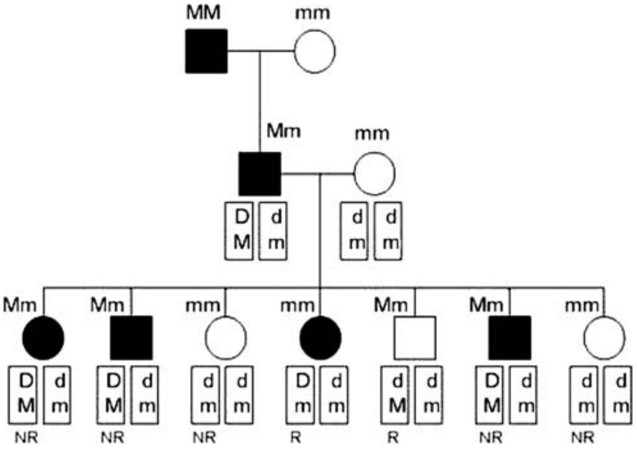


Fig. 2 Illustration of linkage analysis between a genetic marker and an autosomal dominant disease with complete penetrance in a phase-known pedigree. All individuals are genotyped at the marker locus with alleles M and m. The unobserved disease locus has disease allele D and normal allele d. NR means that the gamete is a nonrecombinant, and R means that the gamete is a recombinant

In addition, the father has inherited the *dm* gamete from the grandmother, and thus the *DM* gamete must have been inherited from the grandfather. Once the father's linkage phase is known, the offspring's linkage phases can be easily determined. Comparing the linkage phases of the offspring and the father (note that the mother is doubly homozygous and is not informative), we observe that among the seven offspring, two of them have inherited recombinant gametes from the father, whereas the other five have inherited nonrecombinant gametes from the father.

Let θ denote the recombination fraction between the disease and marker loci, then the likelihood function for this pedigree is

$$L(\theta) = \binom{7}{2} \theta^2 (1-\theta)^5 . \tag{4}$$

The maximum likelihood estimate for the recombination fraction is $\hat{\theta} = 2/7$. The corresponding LOD score is

$$\mathrm{LOD} = \log_{10} \left[\left(\tfrac{2}{7}\right)^2 \left(1 - \tfrac{2}{7}\right)^5 \right] \Big/ 0.5^7 = 0.288, \tag{5}$$

so there is not much evidence of linkage. One important property of the LOD score is that it can be added across families. This is because LOD scores are logs of ratios of two likelihoods, and the contributions from different families to the likelihoods are independent. Suppose we observe 12 such families, then the LOD score is $0.288 \times 12 = 3.46$, which suggests strong evidence of linkage.

The earlier procedure illustrates how one can test for linkage between one genetic marker and a disease locus when the disease model is known. In real gene mapping studies, typically a few hundred markers are genotyped and each of them will be tested for linkage with the disease locus. Since the entire genome is subject to analysis, we need to appropriately correct for multiple testing. Based on a sequential design argument from [6], Newton Morton [7] suggested an LOD score of 3 for genome-wide significance, which is corrected for multiple testing. He showed that an LOD score of 3 corresponds to significance level 0.0001 for a single marker. Although this criterion was originally derived when the analysis is sequential, in practice, it is often used even when the analysis is not sequential.

2.2 Phase-Unknown Pedigrees

Linkage phase cannot always be determined with certainty. In this section, we consider the situation where linkage phase is unknown.

Consider the three-generation pedigree shown in Fig. 2 again. Now, we assume that the grandparental genotypes at the marker locus are missing. In this case, we know that the father's genotype is *DdMm*, however, this corresponds to two possible linkage phases: $DM \,||\, dm$ and $Dm \,||\, dM$, each with equal probability. Figure 3

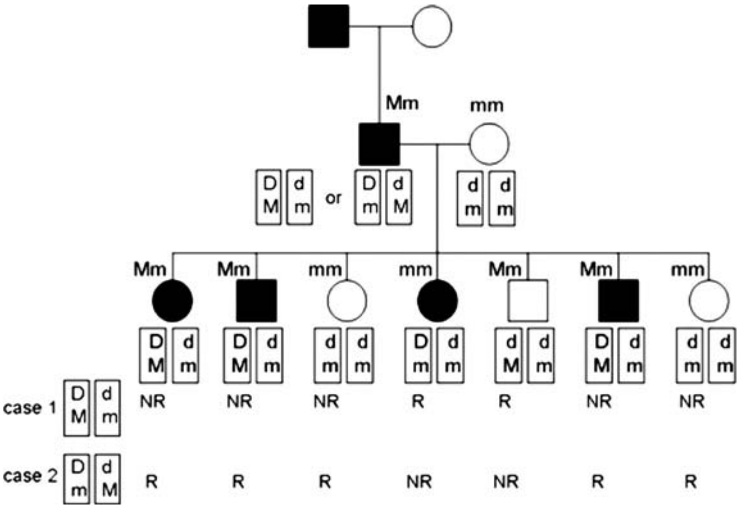


Fig. 3 Illustration of linkage analysis between a genetic marker and an autosomal dominant disease with complete penetrance in a phase-unknown pedigree. All individuals, except for the grandparents, are genotyped at the marker locus with alleles M and m. The unobserved disease locus has disease allele D and normal allele d. NR means that the gamete is a nonrecombinant, and R means that the gamete is a recombinant. Illustrated are two possible phases for the father

displays the recombinants and nonrecombinants in the offspring generation given different paternal linkage phases.

When the paternal linkage phase is $DM \,||\, dm$, the likelihood function is

$$L_1(\theta) = \binom{7}{2} \theta^2 (1-\theta)^5, \tag{6}$$

when the paternal linkage phase is $Dm \,||\, dM$, the likelihood function is.

$$L_2(\theta) = \binom{7}{5} \theta^5 (1-\theta)^2. \tag{7}$$

Since the two linkage phases are equally likely, the overall likelihood for this pedigree is

$$L(\theta) = 0.5L_1(\theta) + 0.5L_2(\theta) = 0.5 \binom{7}{2} \theta^2 (1-\theta)^5 + 0.5 \binom{7}{5} \theta^5 (1-\theta)^2. \tag{8}$$

The maximum likelihood estimate of the recombination fraction is $\hat{\theta} = 0.333$. The corresponding LOD score is

$$\mathrm{LOD} = \log_{10} \frac{0.5\left[0.334^2\left(1-0.334\right)^5 + \left(1-0.334\right)^5 0.334^2\right]}{0.5^7} = 0.023. \quad (9)$$

This LOD score is smaller than the LOD score we calculated earlier when the father's linkage phase is known. The decrease of LOD score is due to the loss of information in the father's linkage phase. This example suggests that knowing other family members' genotypes, such as the grandparental genotypes, can help to infer the linkage phase. In general, large and extended pedigrees are more informative for linkage analysis than small pedigrees because additional family members can help infer linkage phase.

2.3 Linkage Analysis in General Case

In the previous sections, we only considered simple genetic diseases following an autosomal mode of inheritance with complete penetrance.

However, for many diseases, the penetrance may be incomplete, and there might be phenocopies, that is, a person develops the disease even if he/she does not carry the disease genotype. In addition, many families encountered in human genetics studies may vary in size and family structure. Due to this variability, the likelihood approach is attractive since the overall likelihood is simply the product of the likelihood functions across different families. Below we describe the general form of likelihood calculation for pedigree data.

Consider a pedigree with n individuals. Let $Y = (Y_1, \ldots, Y_n)$ denote the phenotype vector (Y_i is 1 if individual i is affected, and is 0 if individual i is unaffected). The likelihood for the analysis of pedigree data involves iterating the overall possible haplo-genotypes, H_i, at the disease and marker loci, for all family members. The overall pedigree likelihood is

$$L\left(Y, G\right) = \sum_{H_1} \cdots \sum_{H_n} \prod_{i=1}^{n} \mathrm{Pen}\left(Y_i | H_i\right) \prod_{j \in O} \mathrm{Prior}\left(H_j\right) \prod_{k \in D} \mathrm{Trans}\left(H_k | H_{k_f}, H_{k_m}\right), \quad (10)$$

where k_f represents the father of individual k, k_m represents the mother of individual k, O represents all originals or founders in the pedigree, and D represents all descendents in the pedigree.

There are three components in the pedigree likelihood in (10): (1) the relationship between phenotype and genotype, $\mathrm{Pen}\left(Y_i \mid H_i\right)$, which is determined by the penetrances; (2) genotype distribution for the founders, $\mathrm{Prior}\left(H_i\right)$, which are determined by allele frequencies at the disease and marker loci; and (3) transmission probabilities of genotypes from parents to offspring, which are determined by Mendelian inheritance. In this pedigree likelihood, we assume that the phenotype of an individual is conditionally independent of the genotypes of other pedigree members, given the individual's own genotype. The conditional independence assumption implies that, given the genotypes of all the individuals in a pedigree, the probability of the joint phenotypes is simply the product of the conditional probabilities of phenotypes of the individuals. This assumption is reasonable when there is no residual correlation induced by other genetic or environmental factors.

The computation time of the pedigree likelihood in (10) increases exponentially with the number of individuals in the pedigree.

Therefore, summing the overall possible haplo-genotypes for each individual can be very time consuming. In Sect. 2.4, we will discuss an efficient computation algorithm proposed by Elston and Stewart [8], which can significantly improve the computation speed for pedigree likelihood calculation based on pedigree peeling.

In model-based linkage analysis, we assume that the model describing the disease locus is correctly specified. For simple Mendelian diseases, one might assume the disease is due to a single major gene effect with complete penetrance. For diseases with incomplete penetrance, the genetic effect is difficult to specify with accuracy. In addition, an individual may develop the disease due to phenocopies. For model-based linkage analysis, misspecification of penetrances can lead to a reduction in power for detecting linkage [9]. Methods that can be used as an alternative to model-based linkage analysis are model-free linkage analysis methods, which will be discussed in Sect. 3.

2.4 Elston–Stewart Algorithm

Pedigree likelihood calculation using (10) can be time-consuming if the pedigree is large. For example, for a pedigree with 20 individuals, even for an autosomal locus with two alleles and three genotypes, the number of terms involved in the summation is $3^{20} = 3{,}486{,}784{,}401$. Therefore, caution is needed when calculating the likelihood. To save computing time, we can eliminate those impossible genotypes, and only sum over possible genotypes for each individual. For example, we can eliminate genotypes that are not consistent with the disease phenotype and offspring genotypes that are not consistent with Mendelian inheritance. However, in certain situations, even after elimination of impossible genotypes, the number of summations involved in the calculation might still be large.

Elston and Stewart [8] developed an efficient algorithm for rapid pedigree likelihood calculation. The basic idea is to calculate the likelihood iteratively with one nuclear family at a time. At each step, choose a nuclear family connected to the rest of the pedigree by only one parent and calculate its contribution to the overall likelihood by considering all family members except for the connecting parent. Repeat this procedure until all nuclear families are considered. In the final step, the overall likelihood is obtained by combining results from all nuclear families together. Figure 4 illustrates the basic procedure of the Elston–Stewart algorithm.

To describe how the Elston–Stewart algorithm operates mathematically, let's consider a nuclear family connected to the rest of the pedigree only through the father.

Let $R = \{r_0, r_1, \ldots, r_s\}$ denote the remainder of the nuclear family (where r_0 represents the mother and $r_1, \ldots, r_s$ represent the offspring), and E denote everyone else in the entire pedigree (including the father).

The likelihood in (10) can be rewritten as

$$\begin{aligned} L &= \sum_{H_E} \prod_{E} \text{Pen} \prod_{E \cap O} \text{Prior} \prod_{E \cap D} \text{Trans} \sum_{H_R} \text{Prior}\,(H_{r_0}) \prod_{j \in R} \text{Pen}\,(Y_j | H_j) \\ &\quad \times \prod_{k=1}^{s} \text{Trans}\,(H_{r_k} | H_{r_0}, H_f) \\ &= \sum_{H_E} \prod_{E} \text{Pen} \prod_{E \cap O} \text{Prior} \prod_{E \cap D} \text{Trans} \sum_{H_R} W\,(H_f, H_{r_0}, \ldots, H_{r_s}), \end{aligned} \tag{11}$$

where

$$W\,(H_f, H_{r_0}, \ldots, H_{r_s}) = \text{Prior}\,(H_{r_0}) \prod_{j \in R} \text{Pen}\,(Y_j | H_j) \prod_{k=1}^{s} \text{Trans}\,(H_{r_k} | H_{r_0}, H_f), \tag{12}$$

is a function that depends only on the haplo-genotype of the father H_f. Equation (11) indicates that at each step, we only need to calculate the likelihood for a nuclear family. After it is calculated, we then choose another such family, and repeat this procedure until all nuclear families are considered. Using this iterated approach, there are fewer items involved in the sum, and can therefore dramatically reduce the amount of computing time. For example, for a marker with three genotypes, if we do not use this family-by-family iterated approach, and instead calculating the likelihood term by term, then there are 3^n terms. However, with the iterated approach, the number of terms need to be evaluated is $3^{s_1+1} + 3^{s_2+1} + \cdots + 3^{s_q+1}$, where q is the number of nuclear families and s_j $(j = 1, \ldots, q)$ is the number of offspring in each family. Consider the three-generation pedigree in Fig. 4. Calculation using (10) involves $3^{14} = 4{,}782{,}969$ summations, whereas the calculation using the Elston–Stewart algorithm only involves $3^{3+1} + 3^{2+1} + 3^{1+1} + 3^{3+1} + 3^{1} = 201$ summations, which is much smaller than the naïve calculation.

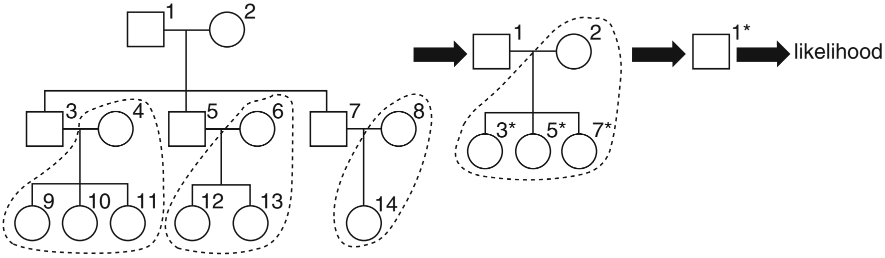


Fig. 4 Illustration of the Elston-Stewart algorithm

The Elston–Stewart algorithm has been proven effective in practical applications. In terms of computation time, it increases linearly with the number of people but exponentially with the number of markers. The algorithm can handle large pedigrees with simple structures, that is, there are no consanguineous marriages. Since the original algorithm was developed in 1971, there have been several enhancements. For example, Ott [10] extended the algorithm to nonlooped complex pedigrees, Lange and Elston [11] extended the algorithm to general complex pedigrees, Lange and Boehnke [12] proposed to update the likelihood one individual rather than one family at a time to save computing time and memory requirements, Lange and Goradia [13] and O'Connell and Weeks [63] proposed efficient algorithms for generating lists of potential haplo-genotypes. The first widely used software package that implemented the Elston–Stewart algorithm is LIPED [14]. With the development of enhanced algorithms, more sophisticated computer implementations of the algorithms have also been developed, including LINKAGE [15], FASTLINK [16], and VITESSE [17]. These sophisticated computer programs allow the analysis of multiple markers, and they have played a crucial role in gene mapping of many Mendelian diseases.

3 Model-Free Linkage Analysis

Although the model-based linkage analysis has been proven successful for the studying of Mendelian diseases, such analysis is less appropriate for complex and non-Mendelian diseases in which the mode of inheritance is unclear. Several studies have shown that misspecification of genetic model parameters can lead to underestimation of the evidence for linkage, as well as to a biased estimate of the disease gene location [9].

For complex diseases, many variations could potentially be involved. However, there are millions of variations in the genome, and we do not know which variations are involved. It is costly to investigate each variation individually. Unlike model-based linkage analysis, model-free linkage analysis does not require a detailed specification for the mode of inheritance of the disease. Therefore, it is more robust and realistic for complex diseases. In this section, we will describe model-free linkage analysis for affected sib pairs (ASPs) and general pedigrees. We will also discuss an efficient pedigree likelihood calculation algorithm, the Lander–Green algorithm.

3.1 Fundamental Principle of Model-Free Linkage Analysis

Model-free linkage analysis was first proposed by Penrose for sib pairs (1935) [18]. The fundamental principle is that individuals with phenotypic similarity should also show genotypic similarity. In other words, individuals who are both affected with the disease should also show a greater similarity than expected at the marker locus if the disease and marker loci are linked. This is because if two individuals are both affected with the disease, then they are more likely to have inherited the same disease allele from a common ancestor. Furthermore, if the marker locus is linked to the disease locus, then they are also likely to have inherited the same allele at the marker locus. Given the relationship between a pair of individuals, the expected degree of genetic similarity between them can be easily calculated when the marker locus is unlinked to the disease locus. This suggests that we can detect linkage by assessing whether there is excess of genetic similarity among affected relative pairs.

3.2 Measure of Genetic Similarity

For qualitative traits, such as disease status, phenotypic similarity can be easily measured by disease affection status, for example, both affected or both unaffected. For genotypic similarity, the simplest measure is based on identical by state (IBS). Two alleles are called IBS if they are the same allele. Another important measure of genetic similarity is based on identical by descent (IBD). Two alleles are called IBD if they are physical copies of the same ancestral allele. IBD implies IBS, but not vice versa. Therefore, IBD contains more information about genetic similarity than IBS.

Consider the example in Fig. 5 in which three nuclear families each with two offspring are genotyped at a marker with four alleles, denoted by 1, 2, 3, and 4. Since IBS simply compares the states of the alleles, we can easily see that the three sib pairs share 2, 1, and 1 alleles IBS, respectively, without considering parental genotypes. To determine the number of alleles shared IBD, we need to consider the parental genotypes. For the first sib pair in (A), both siblings have inherited one copy of allele 1 from the father and one copy of allele 1 from the mother. Since both parents are heterozygous, the IBD of this sib pair is 2. For the second sib pair in (B), both siblings have inherited allele 4 from the mother, but the other alleles of the two siblings are different. Therefore, their IBD is 1. For the third sib pair in (C), although both siblings have inherited allele 4 from the mother, because the mother is homozygous for allele 4, we cannot tell whether the grandmaternal allele or the

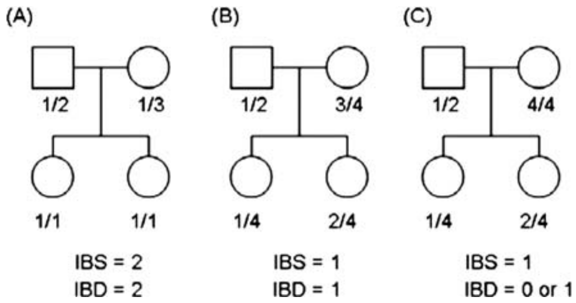


Fig. 5 Illustration of IBS and IBD. All family members are genotyped at an autosomal marker with four different alleles, represented by 1, 2, 3, and 4

grandpaternal allele is transmitted. Therefore, the number of alleles shared IBD for this sib pair is either 0 or 1.

Based on these measures of genetic similarity, we can construct statistical tests to test for linkage between the disease and marker loci by comparing the observed and expected amount of allele sharing at the marker locus. Such analysis does not require specification of the disease model and is therefore called model-free. Sometimes, they are also called nonparametric analysis methods since no disease model parameters need to be specified.

Although model-free linkage analysis can be conducted either using IBS or IBD, methods based on IBD are typically more powerful. This is because IBD can track which ancestral allele is shared in common between affected relative pairs, and therefore carries more information on linkage. IBS-based linkage tests are typically simpler to calculate but carry less information. Since IBD-based tests are more preferred in modern genetic linkage analysis, in the remaining sections, we will focus on IBD-based linkage analysis methods.

3.3 *Model-Free Linkage Analysis for Affected Sib Pairs*

A commonly used study design in gene mapping studies of complex human diseases is the affected sib pair (ASP) design. In this design, a set of affected sib pairs is collected and genotyped. This design is particularly popular for late onset diseases for which parents might not be available for study.

For fully informative data, we can count the number of ASPs that share 0, 1, or 2 alleles IBD. To conduct model-free linkage analysis using ASPs, we need to determine what kind of IBD values would be expected for a sib pair if the genetic marker is unlinked to the disease locus. Under the null hypothesis of no linkage, according to Mendel's first law, each sibling receives one copy of paternal allele and one copy of maternal allele. Since the father has two alleles, the probability that the two siblings will receive the same copy of the paternal allele is $1/2$. Similarly, the probability of receiving the same copy of the maternal allele is $1/2$. Therefore, the probability of sharing two alleles IBD (one from the father and one from the mother) is $1/2 \times 1/2 = 1/4$. Similarly, we can show that the probabilities of sharing 1 and 0 alleles IBD are $1/2$ and $1/4$, respectively.

If the genetic marker is linked to the disease locus, then the ASP are expected to have more than expected allele sharing at the marker locus, that is, the ASP are more likely to share 1 or 2 alleles IBD, and the IBD sharing probabilities will deviate from $(1/4,\ 1/2,\ 1/4)$. Based on this observation, the hypothesis to be tested in ASP linkage analysis is

$$\mathrm{H}_0 : (z_0, z_1, z_2) = (1/4,\ 1/2,\ 1/4)\ \mathrm{vs.}\, \mathrm{H}_1 : (z_0,\ z_1,\ z_2) \neq (1/4,\ 1/2,\ 1/4)\,,$$

where $z_0 = P\,(\mathrm{IBD} = 0|\mathrm{ASP})$, $z_1 = P\,(\mathrm{IBD} = 1|\mathrm{ASP})$, $z_2 = P\,(\mathrm{IBD} = 2|\mathrm{ASP})$ denote the probabilities of sharing 0, 1, and 2 alleles IBD.

A popular method for ASP linkage analysis is the likelihood-based maximum LOD score (MLS) approach [19]. In a simple case, if IBD could be observed with certainty, then each ASP can be scored as sharing 0, 1, or 2 alleles IBD, and we can evaluate the likelihood for the null and the alternative hypotheses, respectively.

Suppose there are n ASPs, and among them n_j $(j = 0, 1, 2)$ share j alleles IBD. The maximum likelihood estimates of the IBD sharing probabilities are $\hat{z}_j = n_j/n$. Under the null hypothesis of no linkage, the likelihood is

$$L = \left(\frac{1}{4}\right)^{n_0} \left(\frac{1}{2}\right)^{n_1} \left(\frac{1}{4}\right)^{n_2}. \tag{13}$$

Under the alternative hypothesis of linkage, the likelihood is

$$L = \hat{z}_0^{n_0} \hat{z}_1^{n_1} \hat{z}_2^{n_2}. \tag{14}$$

The LOD score can be calculated as

$$\mathrm{LOD} = \log_{10} \frac{\hat{z}_0^{n_0} \hat{z}_1^{n_1} \hat{z}_2^{n_2}}{\left(\frac{1}{4}\right)^{n_0} \left(\frac{1}{2}\right)^{n_1} \left(\frac{1}{4}\right)^{n_2}}. \tag{15}$$

This LOD score is also called the MLS. Since there are only two independent parameters involved (due to constraint $z_0 + z_1 + z_2 = 1$), the corresponding likelihood ratio statistic $\mathrm{LRT} = 4.605$ LOD is asymptotically distributed as a chi-squared distribution with two degrees of freedom.

Alternatively, we can conduct a chi-squared goodness-of-fit test to test for linkage. The chi-squared goodness-of-fit test is of the following form:

$$\chi^2 = \sum_{i=0}^{2} \frac{(n_i - e_i)^2}{e_i}, \tag{16}$$

where e_0, e_1, and e_2 are $n/4$, $n/2$, and $n/4$, respectively. This test statistic is approximately distributed as a chi-squared distribution with two degrees of freedom.

So far, we have discussed a simple case where IBD sharing is known. In real life, markers are not always fully informative, therefore IBD might not be known with certainty. To incorporate uncertainty in IBD estimation, we need an alternative likelihood that allows for partially informative data so that all available data are utilized in the analysis. This likelihood should have the following desirable properties: (1) depends on parameters z_0, z_1, and z_2; (2) can incorporate partial information on IBD; (3) for fully informative data, it is equivalent to the previous likelihoods in (13) and (14).

To incorporate partial information on IBD, Risch [19] introduced the following likelihood for an ASP:

$$L\left(z_0, z_1, z_2\right) = \sum_{j=0}^{2} P\left(genotypes|\mathit{IBD} = j\right) P(\mathit{IBD} = j|\mathrm{ASP}) = \sum_{j=0}^{2} w_j z_j, \tag{17}$$

Table 1 Conditional probabilities $P\left(X\,|\mathrm{IBD}\right)$ for ordered sibpair genotypes X. a, b, c, and d are distinct alleles with frequencies p_a, p_b, p_c, and p_d, respectively

	P(*X* | IBD) for		
X	IBD = 0	IBD = 1	IBD = 2
(*aa*, *aa*)	p_a^4	p_a^3	p_a^2
(*aa*, *ab*)	$2p_a^3p_b$	$p_a^2p_b$	0
(*aa*, *bb*)	$p_a^2p_b^2$	0	0
(*aa*, *bc*)	$2p_a^2p_bp_c$	0	0
(*ab*, ab)	$4p_a^2p_b^2$	$p_ap_b(p_a+p_b)$	$2p_ap_b$
(*ab*, *ac*)	$4p_a^2p_bp_c$	$p_ap_bp_c$	0
(*ab*, *cd*)	$4p_ap_bp_cp_d$	0	0

where $w_j = P\left(genotypes|IBD = j\right)$ is the joint probability of genotypes for the sib pair given that they share j alleles IBD. For sibling pairs, the joint probability of genotypes, w_j, can be obtained from Table 1 [20].

For a set of n independent ASPs, the likelihood can be written as

$$L\left(z_0, z_1, z_2\right) = \prod_{i=1}^{n}\sum_{j=0}^{2} w_{ij}z_j. \tag{18}$$

The LOD score is

$$\mathrm{LOD} = \sum_{i=1}^{n}\log_{10}\frac{w_{i0}z_0 + w_{i1}z_1 + w_{i2}z_2}{\frac{1}{4}w_{i0} + \frac{1}{2}w_{i1} + \frac{1}{4}w_{i2}}. \tag{19}$$

Maximization of the sharing probabilities, z_0, z_1, and z_2 can be performed using numerical maximization algorithms such as the EM (expectation-maximization) algorithm [21] subject to constraint $z_0 + z_1 + z_2 = 1$. The MLS is the LOD score evaluated at the maximum likelihood estimates of z_0, z_1, and z_2. MLS obtained in this way is asymptotically distributed as chi-squared distribution with two degrees of freedom. Clearly, for fully informative data, this MLS is equivalent to the MLS in (15).

The MLS method depends on the estimated IBD sharing probabilities z_0, z_1, and z_2, with constraint $z_0+z_1+z_2 = 1$. However, this constraint does not guarantee that the estimated parameter values are biologically meaningful. To solve this problem, Holmans [64] suggested that maximization should be focused on situations where IBD sharing probabilities are compatible with a genetic model, which is equivalent to restricting maximization to a possible triangle, defined by $\mathrm{z}_1 \leq 0.5$ and $z_0 \leq 0.5z_1$. Restriction to the possible triangle has been shown to increase the power of the MLS linkage analysis method (Holmans 1993).

3.4 Multipoint Analysis for Affected Sib Pairs

In Sect. 3.3, we have discussed how to analyze a single marker. Typically, a genome-wide linkage scan involves hundreds of markers. It is not efficient to analyze each marker individually, because the methods discussed before will lose information when a marker is uninformative in a particular family. Multipoint linkage analysis considers a number of markers simultaneously instead of one marker at a time. Since IBD states change infrequently along the chromosome, neighboring markers can therefore help resolve ambiguities about IBD sharing and thus extract more information on IBD.

Multipoint analysis requires the calculation of the likelihood of genotypes for a series of consecutive markers. To model marker dependency and to speed up calculation, a hidden Markov model can be utilized. In this model, it is assumed that IBD sharing at the current marker only depends on IBD sharing at the previous marker. This assumption is reasonable when there is no genetic interference, that is, the presence of a recombination event in one region does not affect the occurrence of recombination events in adjacent regions.

Ingredients for the multipoint likelihood calculation includes: (1) observed genotypes at each marker for an ASP; (2) possible IBD states at each marker; and (3) probabilities that connect IBD states along the chromosome. Let X_j denote the genotypes for an ASP at marker j $(j = 1, \ldots, M)$, and let I_j denote the IBD state for the ASP at marker j. Then the likelihood of the ASP is

$$\begin{aligned} L &= P(X_1, \ldots, X_M|\text{ASP}) \\ &= \sum_{I_1=0}^{2} \cdots \sum_{I_M=0}^{2} P(X_1, \ldots, X_M, I_1, \ldots, I_M|\text{ASP}) \qquad (20) \\ &= \sum_{I_1=0}^{2} \cdots \sum_{I_M=0}^{2} \left\{ P(I_1|\text{ASP}) P(X_1|I_1) \prod_{j=2}^{M} P(I_j|I_{j-1}) P(X_j|I_j) \right\}. \end{aligned}$$

To calculate this likelihood, we need to calculate: (1) IBD prior probability, $P\big(I_1|\text{ASP}\big)$. For an ASP, the IBD prior probabilities are simply $1/4,\ 1/2$, and $1/4$, respectively, for sharing 0, 1, and 2 alleles IBD under the null hypothesis of no linkage. Under the alternative hypothesis of linkage, these probabilities can be estimated numerically. (2) The probability of observed genotypes at each marker conditional on IBD, which can be obtained from Table 1. (3) IBD transition probabilities for adjacent markers.

Later we describe how to calculate the IBD transition probabilities between two adjacent markers. Let θ_j denotes the recombination fraction between markers $j - 1$ and j. The probability of change in IBD states at markers $j - 1$ and j for a sib pair is $\Psi_j = 2\theta_j\,(1 - \theta_j)$. Table 2 gives the IBD transition probabilities for all IBD states [19].

Table 2 IBD transition probabilities between markers $j-1$ and j. θ_j denote the recombination fraction, and $\Psi_j = 2\theta_j\,(1-\theta_j)$ is the probability of change in IBD states between the two markers

		IBD state at marker ***j***		
		0	1	2
IBD state	0	$(1-\Psi_j)^2$	$2\Psi_j\,(1-\Psi_j)$	Ψ_j^2
at marker $j-1$	1	$\Psi_j\,(1-\Psi_j)$	$(1-\Psi_j)^2+\Psi_j^2$	$\Psi_j\,(1-\Psi_j)$
	2 Ψ_j^2	2 Ψ_j	$(1-\Psi_j)$	$(1-\Psi_j)^2$

Although the calculation of the likelihood in (20) is straightforward, in general, the calculation is slow unless there are only a few markers. To speed calculations up, we could reorganize the likelihood using a forward–backward algorithm [22]. This algorithm is based on the hidden Markov model and involves four steps: (1) evaluate probability at the starting marker, (2) evaluate left chain probability for markers on the left-hand side of the starting marker, (3) evaluate right chain probability for markers on the right-hand side of the starting marker, and (4) combine all probabilities and get the final likelihood.

Let j be the starting marker, and let $X_{left} = (X_1,\ldots,X_{j-1})$ and $X_{right} = (X_{j+1},\ldots,X_M)$ denote the genotypes for markers on the left-hand side and right-hand side of the starting marker, respectively. Then, the original likelihood in (20) can be rewritten as

$$\begin{aligned} L &= P\,(X_1,\ldots,X_{j-1},X_j,X_{j+1},\ldots,X_M|\mathrm{ASP}) \\ &= \sum_{I_j=0}^{2} P\,(X_{left},X_j,X_{right}|I_j)\,P\,(I_j|\mathrm{ASP}) \\ &= \sum_{I_j=0}^{2} P\,(X_{left}|I_j)\,P\,(X_{right}|I_j)P(X_j|I_j)\,P\,(I_j|\mathrm{ASP}) \qquad (21) \\ &= \sum_{I_j=0}^{2} L_j\,(I_j)\,R_j\,(I_j)\,P\,(X_j|I_j)P(I_j|\mathrm{ASP}), \end{aligned}$$

where

$$L_j\,(I_j) = P\,(X_1,\ldots,X_{j-1}|I_j) = \sum_{I_{j-1}=0}^{2} L_{j-1}\,(I_{j-1})\,P\,(X_{j-1}|I_{j-1})\,P\,(I_{j-1}|I_j), \qquad (22)$$

and

$$R_j\,(I_j) = P\,(X_{j+1},\ldots,X_M|I_j) = \sum_{I_j=0}^{2} R_{j+1}\,(I_{j+1})\,P\,(X_{j+1}|I_j)\,P\,(I_{j+1}|I_j). \qquad (23)$$

Special cases are $L_1(I_1) = P(X_1|I_1)$ and $R_M(I_M) = P(X_M|I_M)$. Using Baum's forward–backward algorithm, the likelihood can be calculated iteratively, i.e., at each step, we can reuse the results from previous calculations. This type of recursive algorithm can greatly speed up the calculation especially when a large number of markers are involved.

Let $\hat{z}_{j0}$, $\hat{z}_{j1}$, and $\hat{z}_{j2}$denote the maximum likelihood estimates of the IBD sharing probabilities at marker j obtained from the multipoint likelihood in (21). Since the likelihood incorporates multiple linked markers, the estimates are more accurate than estimates obtained based on a single marker.

The LOD score for testing linkage at marker j is

$$\mathrm{LOD} = \log_{10} \frac{L(\hat{z}_{j0}, \hat{z}_{j1}, \hat{z}_{j2})}{L\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)}. \tag{24}$$

Since the IBD estimates are obtained based on multiple linked markers, this is called multipoint LOD score.

3.5 Model-Free Linkage Analysis for General Pedigrees

3.5.1 Inheritance Vector

When only a pair of individuals is considered, genetic similarity can be described by IBD. When more than two individuals are considered, a single number, such as IBD, is no longer sufficient to describe the complex inheritance pattern. A direct extension of IBD to multiple individuals is inheritance vector, which describes the gene flow in a family. The idea underlying the inheritance vector is that it determines for every individual in a pedigree which paternal or which maternal allele has been transmitted to his/her offspring. As a consequence, a meiosis can be described by a bit where 1 denotes the grand maternally transmitted allele and 0 denotes the grand paternally transmitted allele. All inheritance bits are collected to a vector, called the inheritance vector.

Consider a pedigree with n nonfounders. Then there are $2n$ meioses. Genetic transmission in a pedigree at a locus can be characterized by the inheritance vector $\boldsymbol{v} = (v_1, \ldots, v_{2n})$. For meiosis $i\,(i = 1, \ldots, 2n)$, the inheritance bit is defined as

$$v_i = \begin{cases} 0 \text{ if grandpaternal allele passed in meiosis } i, \\ 1 \text{ if grandmaternal allele passed in meiosis } i. \end{cases}$$

Elements in the inheritance vector completely specify, between the two alleles of a parent, which one is transmitted to the offspring. A priori, under the null hypothesis of no linkage, all inheritance vectors, $\boldsymbol{v}$ are equally likely, and therefore $P(\boldsymbol{v}) = 2^{-2n}$.

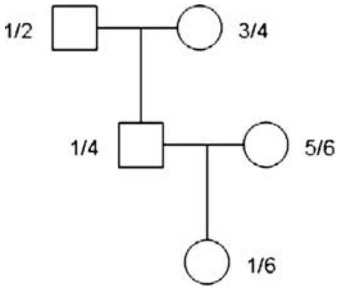


Fig. 6 Three-generation pedigree to illustrate inheritance vector determination. All individuals are genotyped at a marker with six distinct alleles, denoted by $1, 2, \ldots,$ and 6

Later we will describe how inheritance vector can be determined. Consider the three-generation pedigree in Fig. 6. In this family, there are $f = 3$ founders (grandfather, grandmother, and mother) and $n = 2$ nonfounders (father and offspring). Suppose all family members are genotyped at a genetic marker with six different alleles, represented by 1, 2, ... , and 6. For this family, the length of the inheritance vector is $2n = 2 \times 2 = 4$. If we assume each genotype is written with the paternally derived allele first, then we can determine that the inheritance vector is $\boldsymbol{v} = (0, 1, 0, 1)$. Note that, prior to typing the marker, we do not have any information about inheritance pattern, and thus $P(\boldsymbol{v}) = 1/2^4 = 1/16$ for all $\boldsymbol{v}$'s.

In reality, we do not know whether paternally derived allele is listed first or not. In this case, for the two inheritance bits corresponding to the father, we cannot determine whether allele 1 is transmitted from his grandfather or grandmother; similarly for allele 4, we cannot tell whether it is grand paternally or grand maternally derived either. Therefore, the corresponding two elements in the inheritance vector cannot be specified with certainty. For the offspring, since we know the grandfather's genotype, we can tell with certainty that allele 1 is inherited from the grandfather; however, for allele 6, we cannot determine whether it is grand paternally or grand maternally derived. Therefore, the inheritance vector can only be written in the form of $\boldsymbol{v} \in \{(v_1, v_2, 0, v_4) : v_1, v_2, v_4 = 0, 1\}$ and each possible vector $\boldsymbol{v}$ has probability 1/8.

3.5.2 NPL Score When the Inheritance Vector Is Known

In Sect. 3, we discussed how to conduct model-free linkage analysis for ASPs. Such analysis can be extended to other types of relative pairs and to families with a large number of individuals by enumerating all possible inheritance vectors. For general pedigrees, the idea is to calculate a score, S, based on IBD sharing in affected relatives. The score should have the following desirable properties: (1) has known mean and variance under the null hypothesis of no linkage, (2) the value increases under the alternative hypothesis of linkage, (3) uses all affected relatives in a pedigree, and (4) implementation is based on inheritance vectors. Based on these considerations, Whittemore and Halpern [23] proposed the following score:

$$S_{all}\left(\boldsymbol{v}\right)=2^{-a}\sum_{h}\left[\prod_{i=1}^{2f}b_{i}\left(h\right)!\right], \tag{25}$$

where h is one of the 2^a possible set of alleles obtained by choosing one allele from each affected individual, and $b_i(h)$ is the number of times that the founder allele $i\,(i=1,2,\ldots,2f)$ appears in h. $S_{all}(\boldsymbol{v})$ is an average of the terms in the bracket in (25). It gives sharply increasing weight as the number of affected individuals sharing a particular allele increases. This is an attractive feature because sharing a common allele among affected relatives is more appealing that sharing different alleles.

If the inheritance vector for a family is known, then given score $S_{all}\left(\boldsymbol{v}\right)$, we can calculate $Z\left(\boldsymbol{v}\right)=\left[S_{all}\left(\boldsymbol{v}\right)-\mu\right]/\,\sigma$, where μ and σ are the mean the standard deviation of $S_{all}\left(\boldsymbol{v}\right)$ under the null hypothesis of no linkage. If there is no linkage, the mean and variance of $Z\left(\boldsymbol{v}\right)$ are 0 and 1, respectively. To combine data for N families, let $Z=\sum_{i=1}^{N}\gamma_i Z_i$, where Z_i is the score for family i and γ_i is a weighting factor which is chosen such that $\sum_{i=1}^{N}\gamma_i^2=1$. The statistic Z is called the NPL (nonparametric linkage) statistic. For example, we can take $\gamma_i=1/\sqrt{N_i}$, where N_i is the number of individuals in family i. If the inheritance vectors can be observed for each family, then the significance of the observed NPL statistic Z can be determined by normal approximation. Under the null hypothesis of no linkage, Z is asymptotically distributed as standard normal.

3.5.3 NPL Score When the Inheritance Vector Is Uncertain

Usually, $\boldsymbol{v}$ is not observed with certainty. In this case, for each family i with observed marker data X_i, we can calculate $\bar{S}_i=\sum\limits_{v}S_i(\boldsymbol{v})P(\boldsymbol{v}|X_i)$, where $P(\boldsymbol{v}|X_i)$ can be estimated from pedigree likelihood. Let $Z_i=(\bar{S}_i-\mu)/\sigma$, then the NPL score can be calculated as $\bar{Z}=\sum_{i=1}^{N}\gamma_i Z_i$. Since under $\mathrm{H}_0:E\left(\bar{Z}\right)=0$, but $\mathrm{Var}\left(\bar{Z}\right)\leq 1$, the "complete data" ($\boldsymbol{v}$ known) approximation will be conservative for $\bar{Z}$. However, we can still evaluate significance of $\bar{Z}$ through simulation of marker genotypes.

Computer simulations can be time consuming for large pedigrees. To resolve the conservativeness of the NPL statistic when IBD information is incomplete and to avoid time-consuming simulations, Kong and Cox [24] proposed a simple analytical solution. The method is based on the observation that under the null hypothesis of no linkage, all inheritance vectors are equally likely; under the alternative model of linkage, increase of allele sharing is proportional to $S\left(\boldsymbol{v}\right)$.

The Kong and Cox procedure is the following. Let

$$P\left(\boldsymbol{v}|\delta\right)=P_{\text{uniform}}\left(\boldsymbol{v}\right)\left(1+\delta\frac{S\left(\boldsymbol{v}\right)-\mu}{\sigma}\right), \tag{26}$$

In (26), the signal parameter δ is closely related to linkage. The δ parameter represents the degree of allele sharing in a pedigree and can be thought of as measuring the size of the genetic effect, where under the null hypothesis of no linkage, $\delta = 0$, and under the alternative hypothesis, $\delta > 0$, and it corresponds to excess sharing. The likelihood is

$$L\left(\delta\right) = \prod_{i=1}^{N} \sum_{\boldsymbol{v}_i} P\left(X_i | \boldsymbol{v}_i\right) P\left(\boldsymbol{v}_i | \delta\right). \tag{27}$$

The LOD score is calculated by maximizing the single parameter δ in the numerator based on observed genotype data

$$\text{LOD} = \log_{10} \frac{L\left(\delta\right)|_{\delta=\hat{\delta}}}{L\left(\delta\right)|_{\delta=0}}. \tag{28}$$

This LOD score is called the Kong and Cox LOD score. Since under the alternative hypothesis $\delta > 0$, therefore the test is one sided. To evaluate significance of the test, define

$$Z_{lr} = \sqrt{2\left[\log L\left(\delta\right)|_{\delta=\hat{\delta}} - \log L\left(\delta\right)|_{\delta=0}\right]}. \tag{29}$$

When the number of families N is large, the p-value can be approximated by $1 - \Phi\left(Z_{lr}\right)$, where Φ is the cumulative distribution function of the standard normal distribution.

Equation (26) is called linear model [24]. In addition to this linear model, they also suggested an exponential model which is more suitable for models with large deviation from null sharing, especially when there are a small number of families, or when the information is far from complete. The exponential model is defined as the following:

$$P\left(\boldsymbol{v}|\delta\right) = P_{\text{uniform}}\left(\boldsymbol{v}\right) \exp\left(\delta \frac{S\left(\boldsymbol{v}\right) - \mu}{\sigma}\right). \tag{30}$$

Note that when δ is small, $\exp\{\delta\left[S\left(\boldsymbol{v}\right) - \mu\right] /\ \sigma\}$ is approximately $1 + \delta[S\left(\boldsymbol{v}\right) - \mu]/\sigma$, and these two models become very close to each other.

3.6 Lander–Green Algorithm

In Sect. 3.4, we discussed how to calculate the multipoint likelihood for sib pairs using Baum's forward–backward algorithm. In this section, we discuss how to generalize the calculations for general pedigrees using Lander–Green's algorithm (1987) [25]. The Lander–Green algorithm is based on the use of inheritance vectors.

Consider a pedigree with f founders and n nonfounders genotyped for M consecutive markers along a chromosome. Assuming no genetic interference and known marker order, then it can be shown that the inheritance vectors at the M markers, $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M$, constitute a Markov chain, that is, the inheritance vector at the current marker only depends on the inheritance vector at the previous marker.

Recall that for sib pair data, the multipoint likelihood is

$$L = P\left(X_1, \ldots, X_M \,|\, \mathrm{ASP}\right) \\ = \sum_{I_1=0}^{2} \cdots \sum_{I_M=0}^{2} \left\{ P\left(I_1 \,|\, \mathrm{ASP}\right) P\left(X_1 \,|\, I_1\right) \prod_{j=2}^{M} P\left(I_j \,|\, I_{j-1}\right) P\left(X_j \,|\, I_j\right) \right\}. \tag{31}$$

To calculate the likelihood for general pedigrees, we can simply replace IBD states by inheritance vectors,

$$L = P\left(X_1, \ldots, X_M \,|\, phenotypes\right) \\ = \sum_{v_1} \cdots \sum_{v_M} \left\{ P\left(\boldsymbol{v}_1 \,|\, phenotypes\right) P\left(X_1 \,|\, \boldsymbol{v}_1\right) \prod_{j=2}^{M} \right. \tag{32} \\ \left. \times\ P\left(\boldsymbol{v}_j \,|\, \boldsymbol{v}_{j-1}\right) P\left(X_j \,|\, \boldsymbol{v}_j\right) \right\}.$$

Similar to the previous calculation for ASPs, we need the following ingredients for the likelihood calculation: (1) Prior probability of inheritance vector at the starting marker. For a pedigree with n nonfounders, the prior probability of the inheritance vector is 2^{-2n} under the null hypothesis of no linkage. (2) Probability of marker genotypes given inheritance vector at a particular marker. This probability depends on genotype frequencies. (3) Transition probabilities of inheritance vectors.

Later we will describe how to calculate the transition probabilities for adjacent markers. For marker j, denote its inheritance vector by $\boldsymbol{v}_j = (v_{j,1}, v_{j,2}, \ldots, v_{j,2n})$. The transition probability depends on the recombination fraction θ_j, between markers $j-1$ and j. To calculate the transition probabilities $P\left(\boldsymbol{v}_{j-1} \,|\, \boldsymbol{v}_j\right)$, note that $P\left(v_{j-1,i} = v_{j,i}\right) = 1 - \theta_j$ and $P\left(v_{j-1,I} \neq v_{j,i}\right) = \theta_j$. Thus, with one meiosis there are two possible states, no recombination occurs between markers $j-1$ and j (denote the event by 0) and recombination occurs between markers $j-1$ and j (denote the event by 0). The transition probability matrix is

$$T = \begin{bmatrix} 1-\theta_j & \theta_j \\ \theta_j & 1-\theta_j \end{bmatrix}. \tag{33}$$

With two meioses, there are four possible states (00, 01, 10, 11), and the transition probability matrix is

$$T^{\otimes 2} = \begin{bmatrix} \left(1-\theta_j\right) T & \theta_j T \\ \theta_j T & \left(1-\theta_j\right) T \end{bmatrix}$$

$$= \begin{bmatrix} (1-\theta_j)^2 & (1-\theta_j)\,\theta_j & \theta_j\,(1-\theta_j) & \theta_j^2 \\ (1-\theta_j)\,\theta_j & (1-\theta_j)^2 & \theta_j^2 & \theta_j\,(1-\theta_j) \\ \theta_j\,(1-\theta_j) & \theta_j^2 & (1-\theta_j)^2 & (1-\theta_j)\,\theta_j \\ \theta_j^2 & \theta_j\,(1-\theta_j) & (1-\theta_j)\,\theta_j & (1-\theta_j)^2 \end{bmatrix}, \tag{34}$$

where $\otimes$ is Kronecker product. In general, we can write the transition probability matrix using the recursive formulation as follows:

$$T^{\otimes(n+1)} = \begin{bmatrix} (1-\theta_j)\,T^{\otimes n} & \theta_j T^{\otimes n} \\ \theta_j T^{\otimes n} & (1-\theta_j)\,T^{\otimes n} \end{bmatrix}. \tag{35}$$

The transition probability matrix is patterned. It depends on the number of meioses where the outcome changes and the number of meioses where the outcome does not change. The transition probabilities are powers of θ_j and $1-\theta_j$.

Computation requirement for the likelihood in (33) is of order $M\,2^{2n-f}$ [26]. The Lander–Green algorithm is suitable for very large number of markers, but is limited to relatively small pedigrees because the number of possible inheritance vectors increases exponentially with the number of individuals in the pedigree.

Similar to the Elston–Stewart algorithm, there have been many extensions and enhancements of the original Lander–Green algorithm. For example, several researchers noted that there are many redundancies within the inheritance vector space, therefore the calculation can be sped up by focusing on symmetries resulting from the transmission of alleles from single founders [27], or founder couples [28], or individuals in the pedigree [29]. The Lander–Green algorithm was implemented in several software packages. Popular programs include GENEHUNTER [25], ALLEGRO [28], and MERLIN [29, 30].

4 Practical Examples

In the previous sections, we have described the model-based and model-free linkage analysis methods. In this section, we will use a real data example to illustrate how to carry out linkage analysis in real settings.

The data we consider here is obtained from a linkage study on age-related macular degeneration (AMD; [31]). AMD is a complex multifactorial disease that affects the central region of the retina. It is the leading cause of untreatable blindness among the elderly in Western populations [32–34]. A dramatic increase in the size of the aging population makes AMD a significant public health problem and a major focus of research efforts. Various studies have demonstrated a genetic predisposition for AMD [30, 34–36]. The data we consider here is based on a 5-cM genomewide linkage study in families enriched for late-stage AMD. Families in this study were primarily ascertained and recruited from the clinical practice at the Kellogg Eye Center, University of Michigan Hospitals. Since the Retina Clinic serves as a tertiary health care center for the State of Michigan and the surrounding Great Lakes region, the patient population is biased toward late-stage AMD. The patient population used for genotyping in this study is white and primarily of Western European ancestry, reflecting the genetic constitution of the Great Lakes region. In total, the samples include 117 families, 748 individuals, including 321 founders, and 427 nonfounders. The average family size is 6.39 and the average number of generations is 2.41. In this dataset, there are 369 sibling pairs, 10 half sibling pairs, 92 cousin pairs, 843 parent–offspring pairs, 280 grandparent–grandchild pairs, and 157 avuncular pairs. A high-resolution 5-cM genomewide screen was performed at the Marshfield Clinic Research Foundation (Marshfield, WI). For illustration purpose, here we focus on the analysis of chromosome 1, where many linkage studies identified a linkage signal [31, 38–41] and where, more recently, one of the strongest genetic contributors to AMD susceptibility have been mapped [42–45].

The data of [31] include a total of 67 microsatellite markers genotyped on chromosome 1, with average marker heterozygosity of 73.2%. To illustrate the properties of the different approaches, we carried out both model-based and model-free linkage analysis. For model-based linkage analysis, a key step involves selecting parameters for the genetic model. In general, the model parameters are obtained from segregation analysis, which we will not discuss here but instead refer the reader to other more appropriate sources on segregation analysis (e.g. [46]). Even when there is linkage, misspecification of penetrances and disease allele frequencies can significantly reduce power. For complex diseases, such as AMD, where the mode of inheritance is unclear and multiple loci contribute to disease susceptibility, a common strategy is to try multiple genetic models or even to identify a genetic model that maximizes the observed LOD score (MOD score analysis; [47]).

In a recent case-control association study on AMD with 616 cases and 275 controls, Zareparsi et al. [44] estimated genetic model parameters based on genotypes for the Y402H coding variant in the CFH gene on chromosome 1 using methods described in Sect. 5. They fixed the disease prevalence at 20%, and found that a multiplicative model fits the data well. Table 3 lists the estimated genetic model parameters.

For the model-based linkage analysis, we analyzed the data using all four genetic models as specified in Table 3. Figure 7 shows the linkage curves obtained using different genetic models. The command line of running model-based linkage analysis in MERLIN is the following:

```
> merlin -d chr1.dat -p chr1.ped -m chr1.map --model model.txt --grid 0.5
```

Table 3 Estimated genetic models for AMD [43]. Disease prevalence was fixed at 20%. D denotes the disease allele, and d denotes the normal allele. p is the disease allele frequency. f_{dd}, f_{Dd}, and f_{DD} are penetrances for the three disease genotypes

Model	p	f_{dd}	f_{Dd}	f_{DD}
Multiplicative	0.39	0.08	0.20	0.49
Additive	0.41	0.07	0.23	0.38
Dominant	0.46	0.09	0.24	0.24
Recessive	0.46	0.16	0.16	0.36

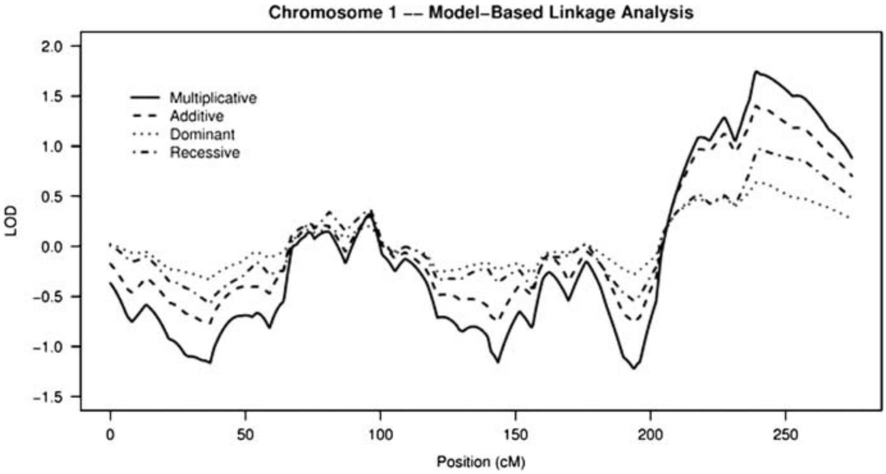


Fig. 7 Model-based linkage analysis for chromosome 1 data

where chr1.ped is the pedigree file that specifies individual relationships, genotypes and phenotypes, chr1.map is the map file that provides marker locations, and chr1.dat is the data file that helps decode the contents of the map file. File model.txt specifies the genetic model parameters. It includes four fields: affection status (matching the data file), disease allele frequency, probability of being affected for individuals with 0, 1, and 2 copies of the disease allele (penetrances), and finally a label for the analysis model. For example, to evaluate the four genetic models in a single run, model.txt would look like:

AMD	0.39	0.08,0.21,0.47	Multiplicative
AMD	0.41	0.07,0.23,0.38	Additive
AMD	0.46	0.09,0.24,0.24	Dominant
AMD	0.46	0.16,0.16,0.36	Recessive

The `--grid 0.5` option specifies that LOD scores should be calculated at 0.5-cM intervals.

Consistent with the results reported by Zareparsi et al. [44], the multiplicative model fits the data the best among the four models we considered because it yields the strongest evidence of linkage. The peak LOD score gradually decreases for the additive, dominant, and recessive models, respectively.

In this example, we chose the genetic model parameters based on estimates obtained from a maximum likelihood approach as described in Sect. 5. The likelihood optimization assumes that there is a single disease locus in the region and that the genetic model is completely specified by the disease allele frequency and the three penetrances. In general, the investigators may have information on disease prevalence or other relevant information. It is recommended that the users choose genetic model parameters that fit the prior information about the disease.

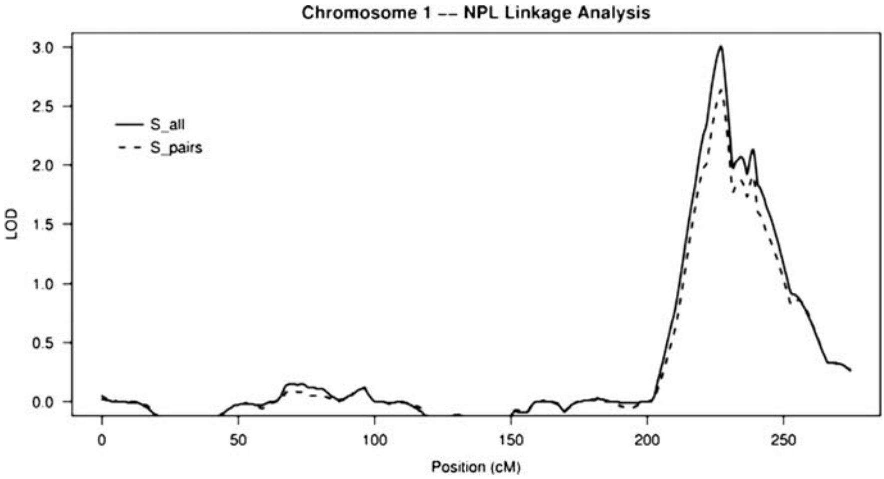


Fig. 8 Nonparametric linkage analysis of chromosome 1 using S_{pairs} or S_{all} statistics

For complex diseases, we also recommend the users to conduct model-based linkage analysis with different genetic models.

A simple and popular alternative to linkage analysis using a parametric model, are nonparametric analyses. In general, these will be less powerful than a parametric analysis when the disease model can be specified correctly; on the other hand, they can be extremely helpful in settings where the disease model is unknown. Next, let us consider model-free linkage analysis for the chromosome 1 data. As discussed in Sect. 3.5, we can either use the S_{pairs} statistic or the S_{all} statistic in the NPL analysis. Figure 8 displays the linkage curves obtained using the NPL linkage analysis method. The MERLIN command line for running the S_{pairs} analysis is the following:

```
> merlin -d chr1.dat -p chr1.ped -m chr1.map --pairs --grid 0.5
```

And the MERLIN command line for running the S_{all} analysis is:

```
> merlin -d chr1.dat -p chr1.ped -m chr1.map --npl --grid 0.5
```

By default, MERLIN uses the linear model [24] for the NPL analysis. As pointed out in MERLIN's documentation, this model is designed to identify small increases in allele sharing spread across a large number of families, and this is what one usually expects for a complex disease. If you are searching for a large increase in allele sharing in a small number of families, then the exponential model [24] specified by the `--exp` option is usually more appropriate. This alternative model is more computationally intensive and requires more memory, but provides a better linkage test if a large increase in allele sharing among affected individuals is expected.

Since many of the families in the dataset include more than one affected sib pair, the NPL analysis using the S_{all} statistic yields slightly stronger evidence of linkage than the analysis using the S_{pairs} statistic. This is consistent with the expected relative power of the two statistics [48].

We note that both the model-based and the model-free linkage analysis yielded a linkage peak around 239 cM on chromosome 1. However, the evidence of linkage using the model-free linkage analysis method is much stronger: the LOD score at the linkage peak using the S_{all} statistic is 2.92, whereas the LOD score at the linkage peak using the model-based analysis assuming a multiplicative model is only 1.74. Although the genetic model parameters are obtained from maximum likelihood in the model-based linkage analysis, the parameter estimation procedure assumes that there is only a single disease locus in the region. It is possible that the true genetic model is different from the four models that we considered in Table 3. In fact, there is now good evidence that there are multiple disease susceptibility alleles in the CFH region on chromosome 1 [45,49]. As shown in previous studies [19], when the genetic model parameters are misspecified, the power of detecting linkage might be reduced.

5 Identifying SNPs Responsible for a Linkage Signal

Linkage analysis is an important first step in position cloning of complex human diseases. However, linkage analysis often results in candidate region of 10–20 Mb. To localize the susceptibility allele more precisely, disease-marker association analyses using dense genetic markers, typically SNPs, specific to the linked region can be carried out. If an SNP shows evidence for association, it is useful to know whether the linkage result can be explained in part or in full by the candidate SNP.

Population-based association analysis often compares marker allele frequencies between unrelated case and control subjects (see the chapter “Population-Based Association Studies”). As alternatives, family-based association methods [50–54] have been developed and they offer a compromise between traditional linkage studies and case-control association studies (see the chapter “Family-Based Association Studies”). A shortcoming of the family-based association methods is that they cannot distinguish between potentially causal SNPs and other variants showing weaker evidence of association. In this section, we will describe a statistical method that identifies candidate SNPs which can explain the observed linkage signal in part or in full through joint modeling of linkage and association using ASPs [55], sibship data and nuclear families [45].

5.1 Assumptions and Definitions

Assume a set of ASPs is genotyped for a candidate SNP and $M \geq 0$ flanking markers that help to evaluate evidence for linkage. It is assumed that the candidate SNP and the flanking markers are in linkage equilibrium. Consider a diallelic disease locus with disease predisposing allele D (frequency p_D) and wild type allele d (frequency $p_d = 1 - p_D$), and a nearby SNP with alleles A (frequency p_A) and a (frequency $p_a = 1 - p_A$). Denote the four disease-SNP haplotypes by *DA*, *Da*, *dA*, and da (frequencies p_{DA}, p_{Da}, p_{dA}, and p_{da}). The superlocus formed by combining the disease and SNP loci is assumed to be in Hardy–Weinberg equilibrium in the general population. Let $f_g = P\,(\text{affected} \mid g)$ be the penetrance for a given genotype $g \in \{dd, Dd, DD\}$ at the disease locus. By definition, the population prevalence of the disease $K = f_{dd}{p_d}^2 + 2f_{Dd}p_dp_D + f_{DD}{p_D}^2$.

Let $X = (X_1, \ldots, X_k, X_{\text{SNP}}, X_{k+1}, \ldots, X_M)$ be the observed marker genotypes for the ASP. Let I_m, I_{SNP}, and I_D be the possibly unknown number of alleles shared IBD by an ASP at marker m, at the candidate SNP, and at the putative disease locus, respectively. It is assumed that there is no recombination between the candidate SNP and the disease locus so that $I_{\text{SNP}} = I_D$. Denote disease locus IBD sharing probabilities for an ASP by $z_i = P\,(I_D = i \mid \text{ASP})$, $i = 0, 1, 2$. For ease of computation, it is assumed that there is no genetic interference so that $\{I_m\}$ forms a hidden Markov chain.

5.2 *Conditional Probability of Marker Data Given ASP*

Since the ASPs are sampled according to their disease status, it is natural to consider the retrospective likelihood $P\,(X \mid \text{ASP})$. To calculate this conditional probability, we can apply Baum's forward and backward algorithm (1972) so that the probability can be calculated as

$$\begin{aligned} P\,(X \mid \text{ASP}) &= \sum_{I_D} P\,(X \mid I_D; \text{ASP}) P\,(I_D \mid \text{ASP}) \\ &= \sum_{I_D} P\,(X_1, \ldots, X_k \mid I_D)\, P\,(X_{k+1}, \ldots, X_M \mid I_D)\, P\,(X_{\text{SNP}}, I_D \mid \text{ASP}) \qquad (36) \\ &= \sum_{I_D} \left\{ \left(\sum_{I_k} P\,(I_k \mid I_D)\, L_k\,(I_k) \right) \left(\sum_{I_{k+1}} P\,(I_{k+1} \mid I_D)\, R_{k+1}\,(I_{k+1}) \right) \right. \\ &\quad \left. \times P\,(X_{\text{SNP}}, I_D \mid \text{ASP}) \right\}. \end{aligned}$$

where k and $k+1$ are flanking markers on the left- and right-hand side of the candidate SNP.

At an arbitrary marker $m\,(1 \le m \le M)$,

$$L_m\,(I_m) = P(X_1, \ldots, X_m \mid I_m) = \sum_{I_{m-1}} L_{m-1}\,(I_{m-1})\, P(X_m \mid I_m)\, P(I_{m-1} \mid I_m)\,, \tag{37}$$

and

$$R_m\,(I_m) = P(X_m, \ldots, X_M \mid I_m) = \sum_{I_{m+1}} R_{m+1}\,(I_{m+1})\, P(X_m \mid I_m)\, P(I_{m+1} \mid I_m)\,. \tag{38}$$

Special cases are $L_1(I_1) = P(X_1 | I_1)$ and $R_M(I_M) = P(X_M | I_M)$. The conditional probabilities of the genotype data given the number of alleles shared IBD for the sib pair at marker m, $P(X_m | I_m)$, are presented in Table 1 in Sect. 3.3. IBD transition probabilities, $P(I_{m+1} | I_m)$, are given in Table 2 in Sect. 3.4. Recursive calculation of $L_m(I_m)$ and $R_m(I_m)$ allows the rapid evaluation of $P(X | \mathrm{ASP})$ in a manner linear in the number of markers M.

To calculate $P(X_{\mathrm{SNP}}, I_D | \mathrm{ASP})$, let G_j denote the disease-SNP haplo-genotype for sib $j = 1, 2$. Summing over all ordered haplo-genotypes that are consistent with the observed SNP genotypes,

$$\begin{aligned} P(X_{SNP}, I_D | \mathrm{ASP}) &= \sum_{(G_1,G_2)\sim X_{\mathrm{SNP}}} P(G_1, G_2, I_D | \mathrm{ASP}) \\ &= \sum_{(G_1,G_2)\sim X_{\mathrm{SNP}}} \frac{P(\mathrm{ASP} | G_1, G_2)\, P(G_1, G_2 | I_D)\, P(I_D)}{P(\mathrm{ASP})} \\ &= \sum_{(G_1,G_2)\sim X_{\mathrm{SNP}}} \frac{f_{G_1} f_{G_2} P(G_1, G_2 | I_D)\, P(I_D)}{P(\mathrm{ASP})}, \end{aligned} \tag{39}$$

where $P(G_1, G_2 | I_D)$ can be calculated from Table 1 in Sect. 3.3 by regarding each haplo-genotype as a genotype of the superlocus that has up to four alleles. For a sib pair, $P(I_D)$ takes values $(1/4, 1/2, 1/4)$ in the general population. Similarly, we can obtain the probability of an ASP, where

$$P(\mathrm{ASP}) = \sum_{I_D} \sum_{(G_1,G_2)} f_{G_1} f_{G_2} P(G_1, G_2 | I_D)\, P(I_D). \tag{40}$$

It is worth noting that the above calculation allows analysis with missing genotypes. For example, to accommodate ASPs where only one sib is genotyped at the candidate SNP, we can sum over all possible SNP genotypes for the sib with missing genotype.

5.3 Relationship Between Disease Locus and Candidate SNP

A useful measure of LD between two loci is the squared statistical correlation, defined as $r^2 = (p_{DA} - p_D p_A)^2 / [p_D(1 - p_D)\, p_A(1 - p_A)]$ in a sample of phased haplotypes. r^2 measures the degree of linkage disequilibrium (LD) between the candidate SNP and the putative disease locus as represented by the observed linkage signal, and can quantify the degree to which the linkage signal is explained by the candidate SNP. The candidate SNP and the putative disease locus can be in linkage equilibrium $\left(r^2 = 0\right)$, in complete LD $\left(r^2 = 1\right)$, or in partial LD $\left(0 < r^2 < 1\right)$. Under linkage equilibrium, the candidate SNP is not associated with the putative disease locus and plays no causal role in the linkage signal. Under complete LD, the candidate SNP or a marker in complete LD with it can fully account for the linkage signal; we call this model plausible causality. Given partial LD, the candidate SNP partially accounts for the linkage signal.

The models can be reparameterized using three penetrances, f_{dd}, f_{Dd}, f_{DD}, and (1) allele frequencies p_D and p_A for the linkage equilibrium model, (2) single allele frequency $p = p_D = p_A$ for the complete LD model, and (3) haplotype frequencies p_{DA}, p_{Da}, p_{dA} for the general model. Given only ASPs, each of these models is identifiable, except the linkage equilibrium model, where parameters $(f_{dd}, f_{Dd}, f_{DD}, p_D, p_A)$ are not all identifiable since the data contain information only for p_A and (z_0, z_1, z_2), corresponding to a total of three degrees of freedom since $z_0 + z_1 + z_2 = 1$. To achieve an identifiable model, note that under linkage equilibrium, $P\left(X_{SNP}, I_D \mid \text{ASP}\right) = P\left(X_{SNP} | I_D\right) P\left(I_D \mid \text{ASP}\right)$ and that $P\left(X_{SNP} \mid I_D\right)$ depends only on p_A. Thus, the linkage equilibrium model can be reparameterized in terms of (z_0, z_1, p_A), resulting in likelihood similar to the traditional MLS linkage test [19] but with an additional parameter p_A. IBD sharing probabilities (z_0, z_1, z_2) should satisfy the triangle constraint: $0 \leq z_1 \leq 0.5$, and $0 \leq z_0 \leq 0.5z_1$ (Holmans 1993). The previous models assume that the candidate SNP is completely linked to the putative disease locus. If the candidate SNP is unlinked, then IBD sharing probabilities at the SNP should be $(1/4, 1/2, 1/4)$, and the only estimable parameter is p_A.

For a sample of independent ASPs, the retrospective likelihood of the data is

$$L = \prod P(X \mid ASP), \tag{41}$$

where the product is taken over all independent ASPs. Here, a retrospective likelihood is chosen because the data are ascertained through their disease affection statuses. Using a retrospective likelihood can avoid the problem of ascertainment bias so that the parameter estimates are valid for the general population.

To maximize the likelihood in (41), we can use a simplex algorithm [56], an optimization method that does not require derivatives. In what follows, we represent the maximum of a particular likelihood subject to its parameter constraints by $\hat{L}$. In addition, r^2 can be estimated from frequency estimates for disease-SNP haplotype frequency. The estimate of r^2 is of particular interest given partial disease-SNP LD; it reflects the degree to which a linkage result is explained by the candidate SNP.

5.4 Hypothesis Testing

Given different relationships between the candidate SNP and the disease locus, we can test for linkage, association, and plausible causality. Let $\hat{L}_{\text{LE}}$, $\hat{L}_{\text{LD}}$, $\hat{L}_{\text{GM}}$, and $\hat{L}_{\text{UL}}$ denote the likelihoods maximized under the models that assume linkage equilibrium, complete LD, a general model that allows LD to vary freely, and no linkage, respectively. Then, we can evaluate evidence for linkage with the maximum LOD score $\mathrm{MLS} = \log_{10}\left(\hat{L}_{\mathrm{LE}}\right) - \log_{10}\left(\hat{L}_{\mathrm{UL}}\right)$. We can evaluate evidence for association by testing whether the candidate SNP is in linkage equilibrium with the disease locus with the likelihood ratio statistic $\mathrm{T}_{\mathrm{LE}} = 2\left[\ln\left(\hat{L}_{\mathrm{GM}}\right) - \ln\left(\hat{L}_{\mathrm{LE}}\right)\right]$. Rejection of linkage equilibrium between the disease and SNP loci suggests the candidate SNP is associated with the disease locus and can account (in part) for the observed linkage signal. We examine plausible causality by testing whether the candidate SNP is in complete LD with the disease locus with the likelihood ratio statistic $\mathrm{T}_{\mathrm{LD}} = 2\left[\ln\left(\hat{L}_{\mathrm{GM}}\right) - \ln\left(\hat{L}_{\mathrm{LD}}\right)\right]$. Rejection of complete LD for an associated SNP suggests that the SNP cannot fully account for the observed linkage signal. If there is a single disease causal variant in the region, then it must be another SNP; or there might be other disease causal variants in the region.

The asymptotic distributions of T_{LE} and T_{LD} under the null hypotheses might in principle be approximated by mixture of chi-squared distributions [57], but the degrees of freedom and mixing parameters are difficult to derive owing to the complexity of parameter constraints and boundaries. Alternatively, significance of the tests can be assessed empirically by simulating marker genotypes under the null hypothesis and comparing the observed statistic with the simulated null distribution. Below, we describe the simulation procedures to obtain the null distributions. When linkage equilibrium is assumed under the null, we can sample SNP genotypes conditional on flanking marker genotypes, which are fixed, and estimated parameters. In contrast, when assuming complete LD between the SNP and the disease loci, we can sample flanking marker genotypes conditional on the observed SNP genotypes and estimated parameters.

For the linkage equilibrium model, we use the observed data to obtain the SNP allele frequency estimate $\hat{p}_A$ and the IBD sharing probability estimates $(\hat{z}_0, \hat{z}_1, \hat{z}_2)$ at the candidate SNP. To obtain a simulated sample under linkage equilibrium, for each ASP, we retain flanking marker data and simulate the IBD configuration at the candidate SNP according to

$$P\left(I_D \mid X_1, \ldots, X_M, \mathrm{ASP}\right) \propto P\left(X_1, \ldots, X_k \mid I_D\right) P\left(X_{k+1}, \ldots, X_M \mid I_D\right) \hat{z}_{I_D}, \tag{42}$$

for $I_D = 0, 1, 2$, where $P\left(X_1, \ldots, X_k \mid I_D\right)$ and $P\left(X_{k+1}, \ldots, X_M \mid I_D\right)$ are the left- and right-chain probabilities calculated in (37) and (38). Given the IBD configuration at the candidate SNP, the ASPs candidate SNP genotypes can then be sampled based on the estimated candidate SNP allele frequency $\hat{p}_A$. We obtain the null distribution of T_{LE} by calculating the statistic for each simulated data set.

The null distribution simulation procedure for the test of complete LD is different. For each ASP, we simulate the IBD configuration at the candidate SNP conditional on the observed SNP genotypes for the ASP and the estimated parameters $\left(\tilde{f}_{dd}, \tilde{f}_{Dd}, \tilde{f}_{DD}\right)$ and $\tilde{p} = \tilde{p}_D = \tilde{p}_A$ obtained from the complete LD model according to

$$P(I_D \mid X_{\mathrm{SNP}}, \mathrm{ASP}) \propto P(X_{\mathrm{SNP}}, I_D \mid \mathrm{ASP}), \tag{43}$$

which can be obtained from (39). We leave the SNP genotypes for the ASP unchanged from their observed values. To avoid excess IBD sharing explained by the flanking markers, we resample their genotypes conditional on the IBD configuration at the candidate SNP. Specifically, we sample genotypes at marker k according to transition probabilities $P\,(I_k \mid I_D)$ and marker k's allele frequencies. Marker $k+1$'s genotypes are sampled similarly but with transition probabilities $P\,(I_{k+1} \mid I_D)$. Moving left and right along the chromosome, we simulate flanking marker genotypes based on $P\,(I_{m-1} \mid I_m)$ and $P\,(I_{m+1} \mid I_m)$, respectively. The rest of the simulation procedure is the same as that for T_{LE}.

5.5 Extension to Sibship Data and Nuclear Families

The methods described earlier for ASPs can be readily extended to general sibship data using inheritance vectors. Let $Y = (Y_1, \ldots, Y_s)$ denote the phenotypes of all s siblings in a sibship. The conditional probability of marker genotypes X given disease phenotypes Y is

$$P\,(X \mid Y) = \sum_{G \sim X_{\mathrm{SNP}}} P\,(X_1, \ldots, X_M, G)\, P\,(Y \mid G)\, /\, P\,(Y), \tag{44}$$

where the summation is taken over all disease-SNP haplo-genotypes that are consistent with the observed SNP genotypes. Summing over all possible inheritance vectors at the disease locus and applying Baum's forward and backward algorithm,

$$\begin{aligned} P(X_1, \ldots, X_M, G) &= \sum_{\boldsymbol{v}_D} P(X_1, \ldots, X_k \mid \boldsymbol{v}_D)\, P(X_{k+1}, \ldots, X_M \mid \boldsymbol{v}_D) P(G, \boldsymbol{v}_D) \\ &= \sum_{\boldsymbol{v}_D} \left[\sum_{\boldsymbol{v}_k} L_k\,(\boldsymbol{v}_k)\, P\,(\boldsymbol{v}_k \mid \boldsymbol{v}_D) \right] \\ &\quad \times \left[\sum_{\boldsymbol{v}_{k+1}} R_{k+1}\,(\boldsymbol{v}_{k+1})\, P\,(\boldsymbol{v}_{k+1} \mid \boldsymbol{v}_D) \right] P\,(G \mid \boldsymbol{v}_D)\, P\,(\boldsymbol{v}_D)\,, \end{aligned} \tag{45}$$

where k and $k+1$ are flanking markers on the left- and right-hand side of the candidate SNP. The summation over all possible inheritance vectors allows the handling of incomplete inheritance information and phase ambiguity by incorporating prior probabilities of the inheritance vectors. At any marker $m\,(1 \le m \le M)$,

$$\begin{aligned} L_m\,(\boldsymbol{v}_m) &= P\,(X_1, \ldots, X_m \mid \boldsymbol{v}_m) \\ &= \sum_{\boldsymbol{v}_{m-1}} L_{m-1}\,(\boldsymbol{v}_{m-1})\, P\,(X_m \mid \boldsymbol{v}_m)\, P\,(\boldsymbol{v}_{m-1} \mid \boldsymbol{v}_m), \end{aligned} \tag{46}$$

and

$$\begin{aligned} R_m(\boldsymbol{v}_m) &= P(X_m, \ldots, X_M \mid \boldsymbol{v}_m) \\ &= \sum_{\boldsymbol{v}_{m+1}} R_{m+1}(\boldsymbol{v}_{m+1}) P(X_m \mid \boldsymbol{v}_m) P(\boldsymbol{v}_{m+1} \mid \boldsymbol{v}_m). \end{aligned} \tag{47}$$

The calculation of (45) requires three probabilities: (1) the prior probability of inheritance vector $\boldsymbol{v}_D$. (2) the inheritance vector transition probability between two consecutive markers, and (3) the conditional probability of marker genotypes given the inheritance vector at that marker. Clearly, the prior probability $P(\boldsymbol{v}_D) = 2^{-2s}$. The transition probability between inheritance vectors at markers m and $m+1$ can be obtained based on derivations in Sect. 3.6.

$$P(X_m \mid \boldsymbol{v}_m) = \sum_{O_m^{\mathrm{dad}}} \sum_{O_m^{\mathrm{mom}}} P\left(X_m \mid O_m^{\mathrm{dad}}, O_m^{\mathrm{mom}}, \boldsymbol{v}_m\right) P\left(O_m^{\mathrm{dad}}\right) P\left(O_m^{\mathrm{mom}}\right), \tag{48}$$

where $P\left(X_m \mid O_m^{dad}, O_m^{mom}, v_m\right)$ takes value of 1 if the sibship's genotype data X_m are consistent with the ordered parental genotypes O_m^{dad} and O_m^{mom} and the inheritance vector $\boldsymbol{v}_m$, and 0 otherwise. The summation is taken over all ordered parental genotypes. $P(G \mid \boldsymbol{v}_G)$ can be calculated in a similar fashion by regarding each haplo-genotype as a genotype of the superlocus formed by combining the disease and SNP loci.

Recursive calculation of $L_m(\boldsymbol{v}_m)$ and $R_m(\boldsymbol{v}_m)$ using these three probabilities allows (45) to be evaluated in a manner linear in the number of marker loci M. This calculation is similar to the Lander–Green algorithm discussed in Sect. 3.6. Equation (44) is an extension of the retrospective likelihood calculation for ASPs described in Sects. 4.1–4.3. Here, the sibship size can be >2 and sibs can be either affected or unaffected.

The above calculation can be readily extended to accommodate parental genotypes. Following the derivation of (44), the critical part in the calculation is the conditional probability of marker genotypes for the sibs and their parents given the inheritance vector at a particular marker. Let X_m^{dad} and X_m^{mom} represent the observed unordered parental genotypes at marker m. Then the conditional probability of the observed genotypes given the inheritance vector at marker m is

$$\begin{aligned} &P\left(X_m, X_m^{\mathrm{dad}}, X_m^{\mathrm{mom}} \mid \boldsymbol{v}_m\right) \\ &= \sum_{O_m^{\mathrm{dad}} \sim X_m^{\mathrm{dad}}} \sum_{O_m^{\mathrm{mom}} \sim X_m^{\mathrm{mom}}} P\left(X_m \middle| O_m^{\mathrm{dad}}, O_m^{\mathrm{mom}}, \boldsymbol{v}_m\right) P\left(O_m^{\mathrm{dad}}\right) P\left(O_m^{\mathrm{mom}}\right), \end{aligned} \tag{49}$$

where the summation is taken over all ordered parental genotypes that are consistent with the observed unordered parental genotypes. This extension enables us to analyze nuclear families with genotyped parents, including parent-affected offspring trios, which are the basic sampling units used by the TDT [49].

Under the assumption that the disease phenotypes are independent given the genotypes at the disease locus, $P\,(Y\,|\,G)$ is the product of simple functions of penetrances. An affected sib $j\,(1 \leq j \leq s)$ with disease-SNP haplo-genotype G_j contributes a term f_{G_j}, and an unaffected sib j contributes a term $1 - f_{G_j}$. By the law of the total probability, the probability of the disease phenotypes for the sibship

$$P\,(Y) = \sum_{G} \left\{ P\,(Y\,|\,G) \sum_{v_G} \left[P\,(G|v_G)\, P\,(v_G) \right] \right\}. \tag{50}$$

Substituting (45), $P\,(Y\,|\,G)$, and $P\,(Y)$ into (44), we can obtain the conditional probability for the sibship $P\,(X\,|\,Y)$ as a function of model parameters $\{f_{dd}, f_{Dd}, f_{DD}, p_{DA}, p_{Da}, p_{dA}\}$.

For sibship data and nuclear families, the tests of linkage equilibrium and complete LD can be carried out in a similar fashion as those for ASPs described in Sect. 4.3. A key advantage of this likelihood calculation is that it allows the joint analysis of different sampling units in a unified statistical framework, leading to more efficient use of the available data.

5.6 Summary

In this section, we have described a unified likelihood framework to estimate useful genetic parameters and to test for both linkage equilibrium and complete LD between a candidate SNP and the putative disease locus through joint modeling of linkage and association. Results from these two tests complement each other in answering whether the candidate SNP can account in part or in full for the observed linkage signal. Estimate of the disease-SNP LD provides a measure to quantify the degree of contribution of the candidate SNP to linkage evidence. Taken together with the disease locus and the SNP allele frequency estimates, this approach will be valuable in helping researchers to evaluate the role of a candidate SNP in disease susceptibility and fine map disease genes. Methods described in this section [45,55] are implemented in the computer program LAMP.

6 Comparison of Model-Based and Model-Free Linkage Analysis Methods

Linkage analysis of pedigree data is a powerful tool for mapping genomic regions that are likely to contain genes influencing human diseases. In this chapter, we described two types of linkage analysis strategy in detail: model-based linkage analysis and model-free linkage analysis. There are numerous model-based and model-free statistics available for the linkage analysis of pedigree data.

Model-based linkage analysis was originally developed for mapping Mendelian diseases using large pedigrees with multiple affected family members. For Mendelian diseases, it is often assumed that the phenotype arises from a single major gene effect with full penetrance, that is, it is expressed regardless of other genetic or environmental factors. Model-based linkage analysis has been successfully applied in mapping of hundreds of Mendelian diseases including Huntington's disease and cystic fibrosis. Model-based linkage analysis is the most powerful linkage analysis strategy when the parameters of the genetic model are known.

However, Mendelian diseases are generally rare. For many common diseases with high prevalence, their penetrances are often incomplete and phenocopies may exist. In this situation, it is difficult to specify the genetic model parameters. For model-based linkage analysis, misspecification of the model parameters due to incomplete penetrances and/or phenocopies may lead to reduction in power to detect linkage [19]. This is also shown in the real data example that we considered in Sect. 4. The reduction in linkage power is because incomplete penetrances can reduce evidence for linkage and phenocopies can give false information about recombination patterns.

Many of the common complex diseases are late onset. For example, the average age at diagnosis for type 2 diabetes is about 45, and the average age at diagnosis for AMD is about 70. For such diseases, it is hard to collect large pedigrees with extended family members. For mapping of complex common diseases, the recent trend is to focus on small families such as affected sib pairs and their parents when available. With the rapid development of genotyping technology, a typical linkage genome scan involves a few hundred microsatellite markers and thousands of SNPs. To simultaneously incorporate all available data, the model-free linkage anlaysis methods are more preferred because they are based on the Lander–Green algorithm which is suitable for small pedigrees with a large number of markers.

Compared to model-based linkage analysis methods, model-free approaches have several advantages, including conceptual simplicity, no need for specification of a genetic model, and a tendency to use small pedigrees which are more easily to obtain than large pedigrees. The traditional mode-free linkage analysis methods assume that the samples to be analyzed are homogeneous. More recently, extensions of the model-free approaches have been developed which can incorporate phenotypic heterogeneity by weighting families or individuals based on covariate information [58] or by incorporating covariates directly into the linkage analysis [59, 60]. In addition, methods that can account for locus heterogeneity have also been developed [61].

Although the model-free linkage analysis is attractive, there are some disadvantages as compared to the model-based linkage analysis. For example, the model-free linkage analysis does not produce estimates of important genetic model parameters. In addition, when the genetic model parameters are correctly specified, the model-based analysis is more powerful. Therefore, we recommend that researchers consider both the advantages and disadvantages of different approaches when determining which linkage analysis methods to choose for a particular disease.

6.1 Software Packages for Linkage Analysis

Various software packages have been developed for linkage analysis. The first widely used computer program was LIPED [10], which implemented singlepoint parametric linkage analysis method for general pedigrees based on the Elston–Stewart algorithm and its extensions. Later computer programs for parametric linkage analysis include LINKAGE (Lathrop et al., [68]), MENDEL [62], and VITESSE [17]. These programs can do multipoint linkage analysis but only for a few genetic markers.

Exact multipoint linkage calculations involving many markers for general pedigrees was first made practical in 1996 with the development of GENEHUNTER [26]. GENEHUNTER implemented the Lander–Green algorithm, and the computing time increases linearly with the number markers but exponentially with the number of individuals per family. Unlike those earlier programs, GENEHUNTER implemented both model-based and model-free linkage analysis methods. However, the NPL method implemented in GENEHUNTER is conservative when IBD sharing is incomplete. To solve this issue, a one-degree-of-freedom method introduced by Kong and Cox [24] was introduced and implemented in a modified version of GENEHUNTER, GENEHUNTER-PLUS.

Although GENEHUNGER has greatly improved the computing speed as compared to earlier programs, it is still slow when family size is large. In 2000, computer program ALLEGRO was introduced [28]. ALLEGRO has much of the functionality of GENEHUNTER but improved GENEHUNTER in several ways, including speed improvement with new computational algorithms, additional scoring functions, improved input and output, and use of disk-swapping to reduce memory requirements.

Another popular modern linkage analysis software package is MERLIN. In the real data example that we considered in Sect. 4, we illustrated how to carry out model-based and model-free linkage analysis using MERLIN. Similar to GENEHUNTER and ALLEGRO, MERLIN also uses a hidden Markov model. However, MERLIN further improves the computational speed of ALLEGRO by representing patterns of gene flow in pedigrees with sparse binary trees. In addition to model-based and model-free linkage analysis, MERLIN can also detect genotyping errors and conduct variance-components based linkage analysis for quantitative traits (see the chapter "Qualitative Trait Linkage Analysis").

Web Resources

LIPED:	http://linkage.rockefeller.edu/soft/liped.html
LINKAGE:	ftp://linkage.rockefeller.edu/software/linkage
VITESSE:	http://linkage.rockefeller.edu/soft/vitesse
MENDEL:	http://www.genetics.ucla.edu/software/mendel
ALLEGRO:	http://www.decode.com/software
GENEHUNTER:	http://www.broad.mit.edu/ftp/distribution/software/genehunter

GENEHUNTER-PLUS: http://www.stat.uchicago.edu/genehunterplus
LAMP: http:// www.sph.umich.edu/csg/abecasis/lamp
MERLIN: http://www.sph.umich.edu/csg/abecasis/merlin

Acknowledgements We would like to thank Dr. Anand Swaroop for allowing us to present results from the analysis of the AMD data.

References

1. Gusella JSF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, Young AB, Shoulson I, Bonilla E, Martin JB (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234–238
2. Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM (1987) Complete cloning of the duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell 50:509–517
3. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080
4. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak, Zielenski J, Lok S, Plavsic N, Chou JL et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066–1073
5. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N et al. (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 245:1059–1065
6. Wald A (1947) Sequential analysis. Wiley, New York
7. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:227–318
8. Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542
9. Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399
10. Ott J (1974) Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597
11. Lange K, Elston RC (1975) Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. Hum Hered 25:95–105
12. Lange K, Boehnke M (1983) Extensions to pedigree analysis. V. Optimal calculation of Mendelian likelihoods. Hum Hered 33:291–301
13. Lange K, Goradia TM (1987) An algorithm for automatic genotype elimination. Am J Hum Genet 40:250–256
14. Ott J (1976) A computer program for general linkage analysis of human pedigrees. Am J Hum Genet 26:588–597
15. Lathrop GM, Lalouel J, Julier C, Ott J (1984) Strategies for multilocus linkage in humans. Proc Nl Acad Sci USA 81:3443–3446
16. Cottingham RW Jr, Idury RM, Schaffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252–263
17. O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402–408
18. Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugenics 6:133–138
19. Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46: 229–241
20. Thompson EA (1975) The estimation of pairwise relationships. Ann Hum Genet 39:173–188

21. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39:1–38
22. Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8
23. Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127
24. Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188
25. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Nl Acad Sci USA 84:2363–2367
26. Kruglyak L, Daly M, Reeve-Daly, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
27. Kruglyak K, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454
28. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12–13
29. Abecasis GR Cherny SS, Cookson WO, Cardon LR (2002) Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101
30. Abecasis GR, Wigginton JE (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77:754–767
31. Abecasis GR, Yashar BM, Zhao Y, Ghiasvand NM, Zareparsi S, Branham KEH, Reddick AC, Trager EH, Yoshida S, Bahling J, Filippova E, Elner S, Johnson MW, Vine AK, Sieving PA, Jacobson SG, Richards JE, Swaroop A (2004) Age-related macular degeneration: a high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease. Am J Hum Genet 74:482–494
32. Klein R, Klein BE, Linton KL (1992) Prevalence of age-related maculopathy. The Beaver Dam Eye Study. Ophthalmology 99:933–943
33. Vingerling JR, Dielemans I, Hofman A, Grobbee DE, Hijmering M, Kramer CF, De Jong PT (1995) The prevalence of age-related maculopathy in the Rottermam study. Ophthalmology 102:205–210
34. Bird AC (2003) Towards an understanding of age-related macular disease. Eye 17:457–466
35. Seddon JM, Ajani UA, Mitchell BD (1997) Familial aggregation of age-related maculopathy. Am J Ophthalmol 123:199–206
36. Klaver CCW, Wolfs RCW, Assink JJM, van Duijn CM, Hofman A, de Jong TVM (1998) Genetic risk of age-related maculopahty. Arch Ophthalmol 116:1646–1651
37. Gorin MB, Breitner JC, De Jong PT, Hageman GS, Klaver CC, Kuehn MH, Seddon JM (1999) The genetics of age-related macular degeneration. Mol Vis 5:29
38. Klein ML, Schultz DW, Edwards A, Matise TC, Rust K, Berselli CB, Trzupek K, Weleber RG, Ott J, Wirtz MK, Acott TS (1998) Age-related macular degeneration: clinical features in a large family and linkage to chromosome 1q. Arch Ophthalmol 116:1082–1088
39. Weeks DE, Conley YP, Mah TS, Paul TO, Morse L, Chang NJ, Dailey JP, Ferrell RE, Gorin MB (2000) A full genome scan for age-related maculopathy. Hum Mol Genet 9:1329–1349
40. Majewski J, Schultz DW, Weleber RG, Schain MB, Edwards AO, Matise TC, Acott TS, Ott J, Klein ML (2003) Age-related macular degeneration – a genome scan in extended families. Am J Hum Genet 73:540–550
41. Seddon JM, Santangelo SL, Book K, Chong S, Cote J (2003) A genomewide scan for age-related macular degeneration provides evidence for linkage to several chromosomal regions. Am J Hum Genet 73:780–790
42. Edwards AO, Ritter R 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA (2005) Complement factor H polymorphism and age-related macular degeneration. Science 308:421–424
43. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308:419–421

44. Zareparsi S, Branham KEH, Li M, Shah S, Klein RJ, Ott J, Hoh J, Abecasis GR, Swaroop A (2005) Strong association of the Y402H variant in Complement Factor H at 1q32 with susceptibility to age-related macular degeneration. Am J Hum Genet 77:149–153
45. Li M, Boehnke M, Abecasis GR (2006) Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am J Hum Genet 78:778–792
46. Elandt-Johnson RC (1971) Probability models and statistical methods in genetics.Wiley, New York
47. Hodge SE, Elston RC (1994) Lods, words, and mods: the interpretation of lod scores calculated under different models. Genet Epidemiol 11:329–342
48. Sengul H, Weeks DE, Feingold E (2001) A survey of affected-sibship statistics for nonparametric linkage analysis. Am J Hum Genet 69:179–190
49. Maller J, George S, Purcell S, Fagerness J, Altshuler D, Daly MJ, Seddon JM (2006) Common variation in three genes, including a noncoding variant in *CFH*, strongly influences risk of age-related macular degeneration. Nat Genet 38:1055–1059
50. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516
51. Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333
52. Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961
53. Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458
54. Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 67:146–154
55. Li M, Boehnke M, Abecasis GR (2005) Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am J Hum Genet 77:149–153
56. Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313
57. Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610
58. Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M (2004) Ordered subset analysis in genetic linkage mapping of complex traits. Genetic Epid 27:53–63
59. Olson JM (1999) A general conditional-logistic model for affected-relative-pair linkage. Am J Hum Genet 65:1760–1769
60. Greenwood CMT, Bull SB (1999) Analysis of affected sib pairs, with covariates – with and without constraints. Am J Hum Genet 64:871–885
61. Schaid DJ, McDonnell SK, Thibodeau SN (2001) Regression models for linkage heterogeneity applied to familial prostate cancer. Am J Hum Genet 68:1189–1196
62. Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471–472
63. O'Connell JR, Weeks DE (1998) PedCheck: A program for identifying genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266
64. Holmans P (1993) Asymptotic properties of affected sib-pair linkage analysis. Am J Hum Genet 52:362–374
65. Weeks DE, Conley YP, Tsai HJ, Mah TS, Schmidt S, Postel EA, Agarwal A, Haines JL, Pericak-Vance MA, Rosenfeld PJ, Paul TO, Eller AW, Morse LS, Dailey JP, Ferrell RE, Gorin MB (2004) Age-related maculopathy: a genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions. Am J Hum Genet 75:174–189
66. Schmidt S, Scott WK, Postel EA, Agarwal A, Hauser ER, De La Paz MA, Gilbert JR, Weeks DE, Gorin MB, Haines JL, Pericak-Vance MA (2004) Ordered subset linkage analysis supports a susceptibility locus for age-related macular degeneration on chromosome 16p12. BMC Genet 5:18
67. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
68. Lathrop GM, Lalouel JM, White RL. (1986). Construction of human linkage maps: likelihood calculations for multilocus linkage analysis. Genet Epidemiol. 3 (1), 39–52

Linkage Analysis of Quantitative Traits

Christopher I. Amos, Bo Peng, Yaji Xu, and Jianzhong Ma

Abstract Nearly three quarters of a century of statistical innovations have resulted from the development of methods to perform genetic linkage analysis in humans and other outbred organisms. Lionel Penrose was among the first investigators to develop methods that could be used to identify genetic linkages for quantitative traits. His methods predated the development of modern likelihood methods or wide acceptance of analysis of variance techniques. He initially sought to partition variance among sibs according to marker similarity [73], assuming particular modes of inheritance. His later publications provided approaches that could be applied for a range of potential inheritance patterns [74]. Oscar Kempthorne [52, 53] developed analysis of variance methods that form a basis for some linkage analytical approaches, building on the earlier work of Sir Ronald Fisher [37]. Fisher developed u-scores which form a basis for efficient score statistics for linkage analysis [38]. Many of the methods developed by these pioneers remain in use, with some modifications to allow their application in a modern era in which thousands of markers are available for analysis in extended families.

This chapter reviews the statistical approaches that are now in use for linkage analysis of quantitative data. We first describe the data that we used to demonstrate methods of analysis. Then, we provide a statement of the genetic model and typical likelihood formulation that are applicable for pedigrees. Next, we discuss a variety of linkage methods that have been developed for model-free linkage analysis. Finally, we describe models for multivariate analysis.

1 Introduction and Description of Data

Genetic loci that influence a trait may either be linked or unlinked to marker loci, which have known chromosomal locations and for which the genotype can be readily deduced from the observable phenotype. When two loci are tightly linked,

C.I. Amos (✉)
Department of Epidemiology, The University of Texas, M. D. Anderson Cancer Center,
1155 Pressler Blvd, Unit 1340, Houston, TX, 77030,
e-mail: camos@mdanderson.org

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_4, © Springer-Verlag Berlin Heidelberg 2010

typically residing within a few megabase pairs of each other, alleles that arose from a common ancestor on a chromosome will be coinherited in passage through a family. The further apart the two loci are located on a chromosome the more likely they are to show independent assortment during meiosis. If a trait locus is linked to a marker locus, then within families individuals with similar marker alleles will show trait values that are more similar than other similarly related members of the same family. When performing linkage studies, it is important to note that familial membership induces correlation in relatives, and the correlation structure depends upon the effects of the trait loci on the quantitative trait. Often in linkage analysis we assume that the marker and trait alleles in the general population do not show any correlation, a condition called dilocus equilibrium but which is often denoted as linkage equilibrium. The term "linkage equilibrium" should more properly be reserved for alleles at two loci that are closely located upon the same chromosome and that are not correlated in the population.

Human traits (such as height, weight, blood pressure, and cholesterol level) and diseases (such as cystic fibrosis and Alzheimer's disease) are most frequently studied because they have direct influence on public health. Other types of quantitative traits such as the level of RNA expression may also vary among individuals and have a genetic component. In this chapter, we analyze these expression values, treating them as quantitative traits. We use data from recent studies of the genetic basis of variation in human gene expression [23, 24, 65]. Because the impact of DNA sequences is manifested through transcription, variations in transcript level can be considered an intermediate stage between DNA sequence differences and complex human traits and diseases. For many RNA transcripts, interindividual variations in expression level of genes are smaller in monozygotic twins than among individuals of other relationships, thus suggesting a genetic component to this variation [23].

The dataset we use to demonstrate methods of analysis consists of RNA expression levels of lymphoblastoid cell lines using Human Focus Affymetrix arrays containing probes from 8,500 transcripts. RNA expression levels of specific genes are obtained by averaging results obtained by multiple probes, which are called probe sets, within the gene. The expression studies were performed on 14 three-generation Centre d'Etude du Polymorphisme Humain (CEPH) Utah families (approximately 14 individuals per family including grandparents, parents, and about eight children per family). For 3,554 of the 8,500 genes tested, [65] greater variation was observed among individuals than between replicate determinations on the same individuals. These 3,554 expression phenotypes (expressed genes) have been made available as a part of the Genetic Analysis Workshop (GAW) XV and are available by request to the organizers of the workshops (see Web Resources). These data were extensively analyzed as a part of the workshops [72, 92].

Like many studies of quantitative traits, such as the NHLBI Family Heart Study, which collect not a single observation on each subject but rather dozens of quantitative traits (e.g., height, weight, cholesterol fractions, blood pressure, fasting glucose, and insulin), the dataset used as an example here includes information from a large number of quantitative traits. Unlike other studies in which quantitative traits are often considered independent, however, expression values of many different probe

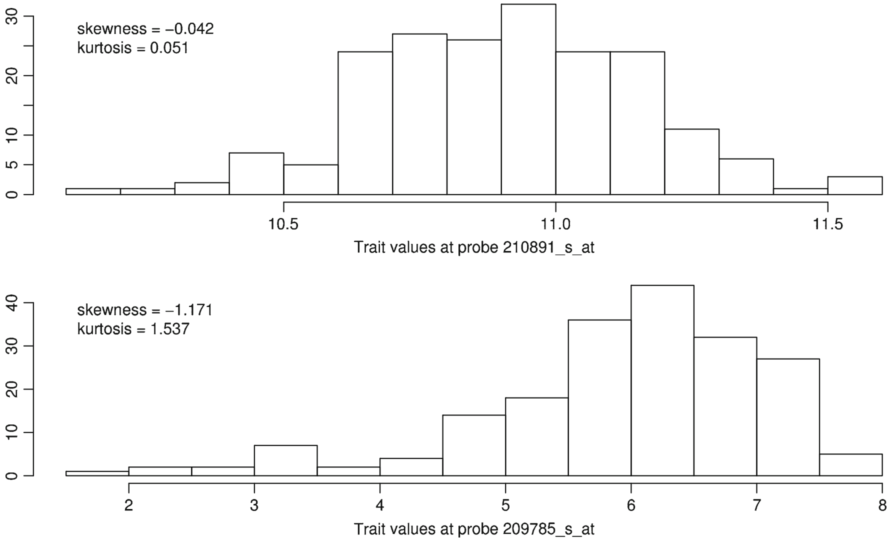


Fig. 1 Histogram of RNA expression levels at probes 210891_s_at and 209785_s_at

sets, which measure RNA transcript levels, are expected to be correlated with each other, and specific pathways through which genes interact have been described, for example by the *Kyoto Encyclopedia of Genes and Genomes* or *Gene Ontology* (see Web Resources). Statistical methods that can efficiently incorporate data from multiple phenotypes, and correlations between them can increase the power to detect linkages [8, 62].

Expression values at different probe sets vary greatly in distributional form. Although many of these traits display approximate normality, the majority show variable departures from normality. Figure 1 plots the histogram of expression values at two probe sets, 210891_s_at and 209785_s_at. These two probe sets are not closely linked and are not on chromosome 5, so are not expected to show linkage. Probe set 210891_s_at has skewness and kurtosis close to zero, while probe set 209785_s_at has significant skewness (-1.171) and kurtosis (1.537). We will use these two probe sets to demonstrate all quantitative trait linkage analysis methods, using 160 markers on chromosome 5.

2 Methods

Many genetic linkage analytical methods have been developed for the study of qualitative rather than quantitative traits. Many of the initially studied severe genetic syndromes such as phenylketonuria or hypercholesterolemia due to LDL receptor defects represent one or a collection of extreme phenotypes for which the development of a model to capture variation along this continuum would not be particularly useful, given a lack of individuals with intermediate phenotypes. When studying more common conditions such as high blood pressure or body mass index, however, incorporating information from the quantitative trait distribution can improve the power to detect linkages and increase the precision of estimates. Quantitative variation in a trait often occurs because there are a number of factors influencing its expression. For a few quantitative phenotypes such as apolipoprotein(a) levels (a major risk factor for heart disease), quantitative variation results from the effects of a large number of alleles at a single locus that determine the trait levels [18, 46]. For many traits, a few genetic loci, along with environmental exposures and measurement error, can explain the trait's continuous interindividual variation. For traits that are influenced by environmental or demographic factors, adjustment for these factors as a part of the linkage process is analytically efficient, but some existing procedures cannot jointly allow for covariates while performing linkage analysis. When using these less sophisticated methods, analyses to adjust for covariate effects must precede analysis of the residuals.

Many models used to map genes responsible for quantitative traits assume normality of the studied quantitative traits. They perform optimally when the trait values of family members follow a multivariate normal distribution. Violation of this assumption can have detrimental effects on the type I error and power, particularly for variance components (VC) methods [2, 11]. Various methods have been proposed to transform trait values, including simple transformations such as square root and logarithmic transformations; more advanced transformations such as Box–Cox; and rank-based transformations. The choice of transformation is often arbitrary, however, and different choices can lead to conflicting results. [31] proposed a method that treats the transformation as part of the parameter space and estimates the transformation along with other parameters. The resulting transformation is rank based and is asymptotically efficient among all order-preserving transformations. However, existing implementations are computationally intensive.

A computationally rapid and simple approach to normalizing data is to apply an inverse probit transformation to the ranked trait data (ENQT) [72]. This method ranks the trait values and scales the ranks to (0, 1). It then transforms the scaled ranks to a normal distribution using an inverse normal transformation. This method is computationally efficient and can be applied to most quantitative traits, resulting in perfectly normal trait values. It is especially suitable, therefore, for studies with a large number of quantitative traits, when manual, customized transformations are not feasible (such as this dataset, although we only described analyses of two traits). Although ENQT is usually an efficient approach for transforming data, it will not result in adequately normal data when many values are tied, for example when values above or below a threshold are truncated [31]. For most of the statistical methods described in this chapter, we will plot the p-values or logarithm of odds (LOD) scores of untransformed data using solid lines, and the p-values or LOD scores of data transformed by ENQT using dashed lines. The impact of normalization is reviewed in the discussion section.

2.1 Classical Model-Based Linkage Analysis

Let X_{ip} denote the trait observation on the ith individual for the pth trait. Let M_i denote the marker phenotype for the i_{th} individual. For this individual, we let g_{is} denote the s^{th} major genotype at a trait-affecting locus, and m_{ir} denote the rth marker genotype at the marker locus. Generally, the marker locus has simple Mendelian codominant expression, so that $P(M_i|m_{ir}) = 1$ for one genotype and 0 for all other genotypes. At the trait locus, the genotype–phenotype relationship $f(X_{ip}|g_{is})$ is the distribution of phenotypes for a given genotype. This genotype–phenotype relation depends on parameters such as the genotype-specific mean μ_s, nongenetic variance σ_e^2, and covariate effects β. Decomposition of the major genetic interindividual variability, σ_g^2, into additive (σ_a^2) and dominance (σ_d^2) components of variance is possible for both diallelic and multiallelic systems [27, 37]. For a diallelic locus, the major gene variance components are $\sigma_a^2 = 2pq[a - (p - q)d]^2$ and $\sigma_d^2 = 4p^2q^2d^2$ where $a = \frac{1}{2}$ the displacement of the two homozygous means and $d =$ the difference between the upper homozygous mean and the heterozygous mean (see the chapter "Population Genetics" for more details).

For univariate phenotypes, we omit the trait-specific indices in the interest of clarity. Let

$$X_i = \mu + \sum_k \beta_k z_{ik} + g_{is} + G_i + e_i, \tag{1}$$

where μ is the average of the genotype-specific means; β_k the kth regression coefficient; z_{ik} the kth covariate observation; e_i the residual variation from the model, $\mathrm{Cov}(e) = I\sigma_e^2$; G_i is a polygenic source of variation, $\mathrm{Cov}(G) = R\sigma_G^2$, where R is the coefficient of relationship between pairs of individuals, $R_{ij} = 1$ if $i = j$, and $(\frac{1}{2})^k$ otherwise, and where k is the degree of relationship between the relative pair (with $E(G_i) = E(e_i) = 0$). Ordinarily, we assume that g_{is} is an unobservable effect from the sth genotype. Because g_{is} is unobservable, we have, without loss of generality, $E(X_i) = \mu + \sum \beta z$. The case in which specific allelic effects are directly observable has been called the measured genotype approach [16, 17] and is subsumed by (1). We denote the total variance by σ_T^2. Total heritability (assuming no gene–environment interactions), h^2, is $\sigma_g^2 + \sigma_G^2 = \sigma_T^2$.

The fitting of VC models that do not include linked genetic markers has been extensively reviewed previously [49]. Here, we assumed that $\mathrm{Cov}(e_i, e_j) = 0$, although this assumption can easily be relaxed to estimate parent–parent, parent–offspring, and shared sibling environments. Finally, we usually assume no covariances between unobservable variables such as $\mathrm{Cov}(g_{is}, e_i) = \mathrm{Cov}(G_i, e_i) = \mathrm{Cov}(G_i, g_{is}) = 0$. These restrictions might be relaxed, but large sample sizes or specialized sampling schemes would be needed to assess the relevant parameters. To identify the $Cov(g_{is}, e_i)$, for example, one could study identical twins reared either together or reared separately so that some of the twin pairs would experience different familial environments and others would experience similar environments [66].

Mendelian segregation of a single major locus imposes a dependence structure that can be exploited to create efficient algorithms for evaluating the likelihood of the data [34, 70]. The general approach consists of a sequential conditioning of the data, which is efficiently accomplished by conditioning children's phenotypes on their parents' phenotypes. The Elston–Stewart algorithm [34] further assumes that the probability of an individual's phenotype conditional on his genotype (the genotype–phenotype relationship) is independent of other pedigree members' phenotypes and that the probability of an individual's genotype depends only on the genotypes of that individual's parents. This algorithm is efficient for analysis of extended pedigrees but is computationally intensive when more than a few markers are being studied jointly. The Lander–Green algorithm [56] is an alternative procedure that uses a hidden Markov model (see Chap. 7) to more efficiently incorporate data from multiple markers, but is limited to computation for fewer than about 20 individuals. A third approach, implemented in Merlin [1], uses a graph theoretical approach to eliminate impossible marker genotypes, and is therefore able to study larger families than the Lander–Green algorithm when multiple relatives have been genotyped.

When performing linkage analysis using a quantitative trait, modeling the genotype–phenotype relationship requires specifying a fixed number of alleles and treating the major gene component, g_i in 1, as a fixed but unobservable effect. We call this modeling approach as fixed effects maximum likelihood (FEML) analysis. The usual FEML approach consists of segregation analysis in which the trait-related parameters are estimated by using only the trait data. The parameter estimates are then typically fixed, and evidence for linkage is then assessed in a separate analysis. For this analysis, the evidence for linkage is evaluated by fitting a dilocus or multilocus model in which the phenotype data of the pedigree members and genotype–phenotype relation are used to infer probabilities of the trait genotypes, and the marker data are used to infer probabilities of marker phenotypes in the pedigree members. Evidence for linkage is assessed by evaluating whether the alleles of the trait and marker loci are co-inherited; evidence for co-inheritance is provided by the estimated recombination fraction θ.

This approach is called "classical," or "traditional linkage analysis," but a more precise name for this type of analysis is maximum pseudolikelihood analysis (MPLA), since all parameters of the entire likelihood are not jointly modeled. Quantitative traits are often influenced by several loci, but most segregation analytic approaches model only a single locus so that the correct genetic model cannot be obtained. If the trait model is not accurate, then biases will occur in the estimates from MPLA. For qualitative traits, which have been well studied, the most serious of these biases occurs in the recombination fraction (θ), which tends to be overestimated [25,51]. In the study of a quantitative trait, if the interindividual variance is underestimated, then the recombination fraction is overestimated and vice versa [7]. If preliminary segregation analysis does provide an accurate description of the genetic sources of variability of a trait, then MPLA is statistically more efficient than the model-free methods described later in this chapter. Hence in those situations in which a simple genetic mechanism explains interindividual variability, MPLA could be a preferred method of analysis. In the context of studying variability of RNA expression levels, MPLA would be a preferred approach for analysis if, for example, all variabilities were due to effects of DNA variation within the coding sequence of the gene.

To perform MPLA, we first used the SEGREG module of the software package SAGE (see Web Resources). For this analysis, we had to assume that an unknown single locus with two genotypes is influencing the trait levels of probe sets in pedigrees. To reduce the parameters to estimate, we further assumed that all genotypes have the same variance. We then used maximum likelihood methods to estimate genotypic means that best fitted the data. The models on how genotype at an unknown gene modulates RNA expression values at probe sets 209785_s_at and 210901_s_at are listed in Table 1. For example, an individual with genotype AA at this marker is expected to have expression value 5.23 for probe set 209785_s_at.

A number of software applications can be used to estimate the location of this unknown marker, given a model. As shown in Fig. 2, we scanned the 160 markers on chromosome 5 and plotted the $-\log_{10}(p - \text{value})$ at each marker. We use the LODLINK module of SAGE to perform the analyses, although other programs

Table 1 Segregation analyses of two traits

Trait	Mean of AA	Mean of Aa	Mean of aa	Residual Variance
209785_s_at	5.23	3.74	6.32	0.6
210891_s_at	10.94	10.93	10.49	0.05

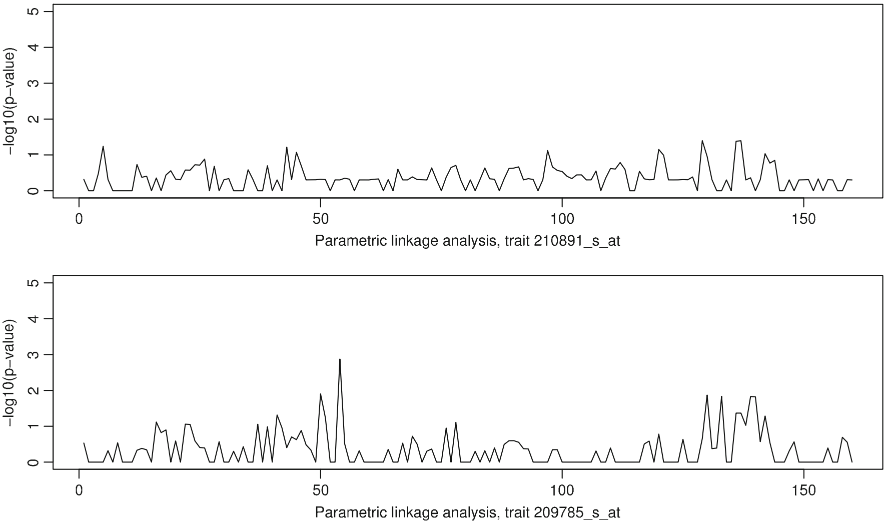


Fig. 2 Result of parametric linkage analyses
$-\log_{10}(p\text{-value})$ of the parametric linkage analyses of two traits, 210891_s_at (*top*) and 209785_s_at (*bottom*), at 160 markers on chromosome 5. Note that, although the markers are equally spaced in the figure, they are not actually unevenly distributed

such as Fastlink [26] also can be used. LODLINK requires that the trait locus has only two alleles but can jointly adjust for covariate effects, while Linkage (see Web Resources) can perform analysis assuming multiple alleles at the trait locus but cannot adjust for covariate effects. Compared with results obtained from nonparametric linkage analyses (see below), the signals obtained in the parametric linkage analysis show less evidence for linkage than model-free results.

Many existing software programs are not equipped for analysis of quantitative data using multiple marker data (called multipoint analysis in the genetic literature). The Elston–Stewart algorithm implementations in the Linkage [58] and Fastlink [26] programs are efficient for analysis of extended pedigrees but can manage only five or six markers jointly. To analyze the markers on chromosome 5 requires interleaving results from multiple separate analyses. As an alternative, [80] showed how existing software applications (such as Merlin and GeneHunter) that have been developed for analysis of qualitative data using liability classes can be applied for the analysis of quantitative data using a FEML approach. For each subject, the genotype-specific value of the probability density function is used to model the probability of observing a phenotype given each of the possible genotypes for an individual. This approach to modeling was also shown to adjust adequately for effects of nongenetic cofactors.

Rather than performing segregation analysis followed by linkage analysis, segregation and linkage analyses can be performed jointly by using a variety of analytical methods. Using Linkage [58] or Fastlink [26], one can optimize a likelihood expression that models the recombination fraction between a single trait locus and a marker locus. Application of these procedures for joint segregation and linkage analysis is not well described in the accompanying software manuals, but is permitted by using the optimization program Gemini, specifying a string of 0's and 1's in the last line of the parameter file, in which the 1's represent those parameters to be estimated. The order of estimated parameters consists of the recombination fraction, followed by parameters for each locus. If a trait parameter is indicated, then the parameters that will be estimated include the allele frequencies, trait means, and residual variance, and if specified, a parameter allowing the heterozygote residual variance to deviate from the homozygote residual variance. While Gemini is computationally rapid, it is sensitive to initial conditions and may not converge well if too many parameters are estimated jointly. Practically, therefore, analysis using this software should be performed by first estimating marker allele frequencies in a first step. Subsequent steps would be estimating the trait means and variances along with the recombination fraction holding marker allele frequencies fixed. Provided dilocus equilibrium holds, estimating the marker allele frequencies first in a separate step would lead to only a minimal decrease in efficiency in estimating the recombination fraction or trait-related parameters.

As an alternative, model-based joint segregation and linkage analysis can be accomplished by implementations of Monte-Carlo Markov Chain (MCMC) procedures (see the chapter "Markov Chain Monte Carlo Linkage Analysis Methods"). The most widely used method is implemented in the program Loki [48]. This program uses the Metropolis-Hastings algorithm to perform joint segregation and linkage analysis. The program requires that effects from unmeasured genetic factors affect the trait in an additive fashion and requires that each locus be diallelic. It uses a reversible-jump algorithm to modify the likelihood, allowing for a variable number of loci influencing the trait. [82] recently studied the behavior of an MCMC approach to joint segregation and linkage analysis. [60] provided a Bayesian framework for analysis of selected samples and extended the Loki program to allow for selection.

2.2 Model-Free Haseman–Elston Regression Approach

To circumvent these issues in modeling, a variety of model-free tests for genetic linkage have been developed. Model-free methods evaluate the similarity among pairs of individuals for both the marker and trait phenotypes. In contrast to FEML, which usually fixes the genotype–phenotype relationship and then evaluates genetic relationships, model-free methods estimate the genetic relationships among individuals in a first step and then evaluate the evidence for a trait-influencing locus at the specified location.

The model-free methods described here are based on identity-by-descent (IBD) sharing, conditional on pedigree marker information. IBD measures genetic similarity among pairs of relatives from their common inheritance of particular marker alleles. We say that a pair of relatives shares an allele IBD if that allele can be traced to a common ancestor. Often not all family members are available for study (typically one or both parents are missing). In this case, the probabilities that pairs of individuals share zero, one, or two alleles IBD are calculated by using the available family members and population genotype frequencies. These proportions can then be used to create an estimated proportion of alleles IBD. Algorithms for evaluating IBD sharing for pedigree data conditional on marker data have been extensively developed [5, 8, 9, 28, 29, 54, 55, 88]. Similarity in IBD sharing is then used to evaluate trait similarity by using either regression or VC analyses, as discussed later. The *model-free* methods require specifying forms describing measured covariates affecting the first moments, and they estimate genetic and environmental factors affecting interindividual variance and the covariances among pairs of relatives. Thus, the modeling strategy depends only upon observable quantities, unlike *model-dependent* strategies, which must infer unmeasured genotypes, and this inference is conditional upon correct knowledge of additional unobservable and often confounded elements of the model, such as the number of alleles, the allele effects, and the within-genotype variances.

One of the most popular model-free tests for linkage is the Haseman–Elston regression approach. The approach is conceptually simple and provides a rapid test for linkage. For a particular family, we let $Y_{ij} = (X_i - X_j)^2$. Then a linear regression relationship has been established under linkage [8,47] between the squared pair differences between pairs of relatives, Y_{ij}, and the estimated proportion of marker alleles IBD. The regression coefficient, β, for all pairs of relatives is a function of

$-2(1-2\theta)^2\sigma_a^2$ with additional terms involving $(1-\theta)$ depending upon the relative pair type [8]. Thus when $\theta = 0$ (complete linkage), $-\frac{1}{2}\beta$ estimates σ_a^2 for all relative pairs. [35] and [32] developed a revised version of the Haseman–Elston test, which we call covariance Haseman–Elston (CH-E). This procedure consists of regressing the sib-pair covariance on to IBD sharing. [36] further developed the CH-E to allow for correlation among the sib-pairs. Allowing for the intrasibship correlation considerably improves the power and efficiency of the CH-E when there is positive correlation among the sib-pairs. [81] proposed optimal weighting procedures that weight sums and differences of trait values according to the correlation in trait values among relatives. The Haseman–Elston approach has been extended to permit analysis of arbitrarily related individuals [5, 69], which is implemented in the LODPAL module of SAGE.

Figure 3 plots the $-\log_{10} p-$ values of probe sets 210901_s_at and 209785_s_at at the 160 markers on chromosome 5, using basic and weighted Haseman–Elston regression analyses. For this particular dataset, the two methods yield almost identical results. In all figures, dotted red lines represent results for ENQT-transformed trait values. It is clear that ENQT has no impact on normal trait 210901_s_at, but changes the signal of nonnormal trait 209785_s_at significantly. We used the SIBPAL module of SAGE to perform the analyses.

2.3 Variance-Components Approaches

Although tests for genetic linkage can be developed by considering the squared pair differences for the trait values, this transformation can lead to a reduction of power to detect linkage [91]. An alternative to the Haseman–Elston approach consists of directly modeling the covariance structure of the data, conditional on the IBD sharing of the relative pairs. This approach can jointly model covariate effects along with VC. The VC are decomposed into the major genetic component, which is linked to a genetic marker; polygenic or major genes unlinked to the genetic marker; and nongenetic sources of variability. If we let π_{ij} be the proportion of marker alleles IBD and v_{ij} be the probability of sharing two alleles IBD, then

$$\mathrm{Cov}(X_i,\, X_j \mid \pi_{ij},\, v_{ij}) = \begin{cases} \sigma_a^2 + \sigma_d^2 + \sigma_G^2 + \sigma_e^2 \text{ if } i = j \\ b_{ij}(\theta, \pi_{ij})\sigma_a^2 + c_{ij}(\theta, \pi_{ij}, v_{ij})\sigma_d^2 + R_{ij}\sigma_G^2 \text{ otherwise} \end{cases}$$

For sib-pairs, $b_{ij}(\theta, \pi_{ij}) = \frac{1}{2} + (1-2\theta)^2(\pi_{ij} - \frac{1}{2})$ and $c_{ij}(\theta, \pi_{ij}, v_{ij}) = 4\theta^2(1-\theta)^2 + (1-2\theta)^2\pi_{ij} + (1-2\theta)^4 v_{ij}$. When π is estimated from marker data, we use the estimate in this expression [42]. Table 1 of [6] provides these values for other relative pairs and [3] provide algorithms for studying extended pedigrees. The effects of all unlinked genetic factors are captured in the term σ_G^2, which denotes the polygenic component of variance. If multiple marker loci are available for study, however, the variance component expression can be further partitioned into components for each

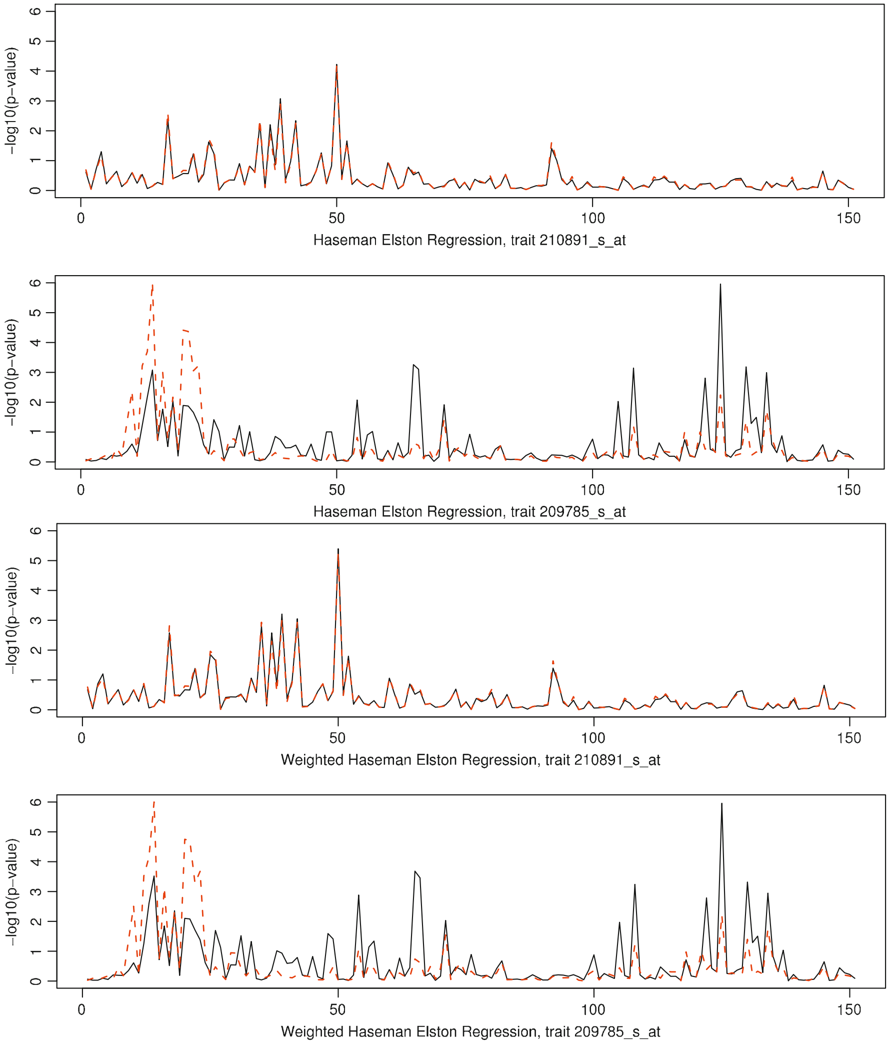


Fig. 3 Basic and weighted Haseman–Elston regression
$-\log_{10}$ p-values of the basic (*top two*) and weighted (*bottom two*) Haseman–Elston regression analyses. *Dotted lines* are the results for ENQT-transformed data. Results obtained from basic and weighted Haseman–Elston analyses are numerically, but not visually, different

of the marker loci. When multiple relative pair types are considered, the recombination fraction can be estimated. An interval-mapping approach [39] usually is used, however, in which evidence for a genetic effect is studied at multiple markers, and the strongest evidence for a major locus is taken to be the point where the major gene heritability, σ_a^2, is largest (see the multipoint mapping section below). Generalized estimating equation (GEE) methods [68, 75] can be adapted for the purposes

Table 2 Bivariate and univariate analyses

Component	Estimate	Standard error	Wald p-value
Bivariate analysis: LRT = 8.48 (rs879253)			
Mean 1	10.9182	0.0306	0
Mean 2	6.1513	0.1093	0
Major gene variance 1	0.0150	0.0101	0.0686
Major gene variance 2	0.2420	0.1543	0.0584
Major gene covariance	−0.0220	0.0272	0.7906
Polygene variance 1	0.0101	0.0147	0.2463
Polygene variance 2	0.0399	0.2078	0.4238
Polygene covariance	0.0200	0.0285	0.2406
Residual variance 1	0.0371	0.0084	4.8643E-06
Residual variance 2	0.6114	0.1221	2.7825E-07
Univariate analysis for trait 1 only: LRT = 3.54695 (rs879253)			
Mean 1	10.9204	0.0304	0
Major gene variance 1	0.0164	0.0104	0.0586
Polygene variance 1	0.0080	0.0148	0.2951
Residual variance 1	0.0378	0.0084	3.5294E-06
Univariate analysis for trait 2 only: LRT = 4.381593 (rs879253)			
Mean 1	6.1673	0.1080	0
Major gene variance 1	0.2523	0.1576	0.1107
Polygene variance 1	0.0168	0.2084	0.3977
Residual variance 1	0.6216	0.1221	9.3191E-07

of fitting expressions given by expressions (1) and (2) [10], and robust variance estimation can be used to provide robust estimates of the variance of each variance component [10,21,59]. Gessler and Xu [42] found virtually identical results for VC methods that use IBD sharing expressed either as probabilities of sharing zero, one, or two alleles or as the proportion of alleles shared.

We have previously used simulation studies to evaluate the power and efficiency of the Haseman–Elston and VC procedures for varying sample sizes, sibship sizes, and deviations from the assumptions that the residual nongenetic source of variance is normally distributed. Table 2 [10] compared the power of Haseman–Elston, a GEE method, and maximum likelihood estimation assuming normality of the trait. The GEE method was generally less biased than maximum likelihood, but also had a higher median squared error. [21] further developed GEE methods for variance components analysis and derived an approach that is nearly as efficient as maximum likelihood, but does not require normality.

Power for univariate VC analyses has been studied by simulation studies [6, 10, 40,43,71,76]. Numerous researchers also applied Haseman–Elston and VC methods in the simulated data as a part of the GAW X [89]. Together, results show similar power of the VC and FEML methods [44], and both methods are more powerful in general than the Haseman–Elston method.

Sample size requirements for linkage analysis depend most heavily on the major gene heritability of the trait, and effects of less than 10% require a generally prohibitively large sample size. Several investigations have provided analytical power and sample size by describing noncentrality parameters for univariate VC analysis [83, 90, 91]. For evaluating the effect from a linked genetic factor, [90] find the noncentrality parameter from a chi-square distribution to be

$$\Lambda = \frac{q^4}{2(4-h^4)^2}(4+h^2),$$

where q^2 is the major genetic heritability, σ_a^2/σ_T^2. Additional forms, including the contributions from additional sibs and parents, are given by [15]. These forms show rather dramatic increases in power for increasing sibship size and particularly for extended families. [20] provided analytical power approximations that can be applied for large pedigrees.

Although VC methods are generally more powerful than regression methods, such as the Haseman–Elston or covariance Haseman–Elston methods, they can be sensitive to normality assumptions. [2] showed that for sib-pairs, when multivariate normality is assumed but the data are skewed or kurtotic, VC methods have excessively large type I error rates whenever there is correlation among the sib-pairs. As a particularly bad example, using a Laplace distribution that had a skewness of 0 but a standardized kurtosis of 3, the observed empirical 5% power corresponding to a theoretical significance of 5% (i.e., the size of the test) was 17% when the residual sibling correlation was 50%. These results indicate the need to evaluate distributional assumptions when applying VC methods, the need to develop diagnostic tools, and the need for robust VC tools such as least squares estimation [12], GEE [10, 21, 87], permutation testing [46] and M-estimation [86]. Inferences using the likelihood ratio test and inappropriately assuming multivariate normality can be inaccurate when the underlying trait data display either skewness or kurtosis.

As a part of GAW XI, data were distributed for a quantitative trait, MAO-B, which is correlated with alcoholic behavior and was studied by several groups for possible linkages. The distribution of the data did not significantly deviate from a normal distribution, but one family included three individuals with extremely high values (over ten standard errors beyond the mean) along with two individuals having normal levels. Analyses were extremely sensitive to this family [14], and LOD scores for some genomic regions were much larger when this family was included than when it was excluded. This finding underscores the need for methods to identify families that contribute greatly to LOD scores during analysis and also the potential need to routinely use a robust method in VC analyses. [85] proposed a data-trimming approach in which extreme observations are replaced by a specified quantile. For instance, if the data are trimmed to the 99th percentile, then any observations more extreme than the 99th percentile are replaced by the 99th quantile of the data. This approach was shown to be more powerful than usual maximum likelihood VC when the data were influenced by unusual families having extreme values [22].

As an alternative approach, [31] proposed replacing the trait observations with a semiparametric transformation in which the same VC model is assumed, but the trait that is analyzed is that inverse normal transformation of the data that maximize the likelihood. This procedure, applied to the GAW XI data, showed only minimal evidence for linkage and so appears to have provided an accurate assessment of linkage for this unusually distributed trait. While the approach that they developed was effective for this situation, it is highly computationally intensive. Therefore, [72] proposed an alternative approach in which the trait is transformed via a probit transformation to normality as a first step and then the transformed data are analyzed. In principle, this approach should provide less power than the method of [31], since the method does not seek to jointly maximize the likelihood with the transformation; in practice, however, [72] observed little loss in power using their approach compared to that of [31].

Multilocus methods have been extensively developed for VC methods. Analytical formulas for epistatic effects of trait loci were provided by [83]. Analytical procedures for multilocus analysis are available in ACT (see Web Resources), but have been more extensively developed in the software package SOLAR [3]. When studying a trait influenced by multiple trait-affecting loci, joint modeling of the effects will lead to an improvement of power [15]. Although the SOLAR package is somewhat cumbersome to install and maintain because it uses TCL to integrate analysis from multiple programs it has a great degree of flexibility in analytical and modeling schemes. In particular, the user can specify various types of joint effects of the genotypes of the different loci that are to be modeled. The implementations of VC procedures vary somewhat among packages. ACT and Merlin do not constrain VC (other than to assume they are nonnegative). SOLAR requires that the overall trait heritability be specified and then partitions the linked and unlinked components of variance to equal the total heritability. A usual approach would be to assume that the traits have independent effects upon the trait, but various forms of epistasis can also be modeled. Modeling each of these additional epistatic terms will introduce another variance component. [21] introduced similar multilocus models using GEE methods that are relatively less sensitive to nonnormality than maximum likelihood-based methods.

We used Merlin [1] to run VC analyses on the two probe sets, using 160 markers on chromosome 5. Figure 4 plots the $-\log_{10}p$-values of these two traits. The dotted lines represent results for the ENQT-transformed dataset. It is clear that normalization has a strong impact on VC analyses when considering the trait 209785_s_at.

In genetic analyses, we often analyze selected samples. For traits that are influenced by uncommon or rare alleles that confer an extreme phenotype, sampling through individuals with extreme phenotypes (a process named "ascertainment" by Sir Ronald Fisher) increases the proportion of subjects with the uncommon alleles in the sample, thus improving the ability to observe segregation within families and to identify linkages. When samples have been selected for study on the basis of the phenotypes of one or more individuals, an ascertainment correction is needed to adequately describe the data. [30] reported results from VC analyses of selected

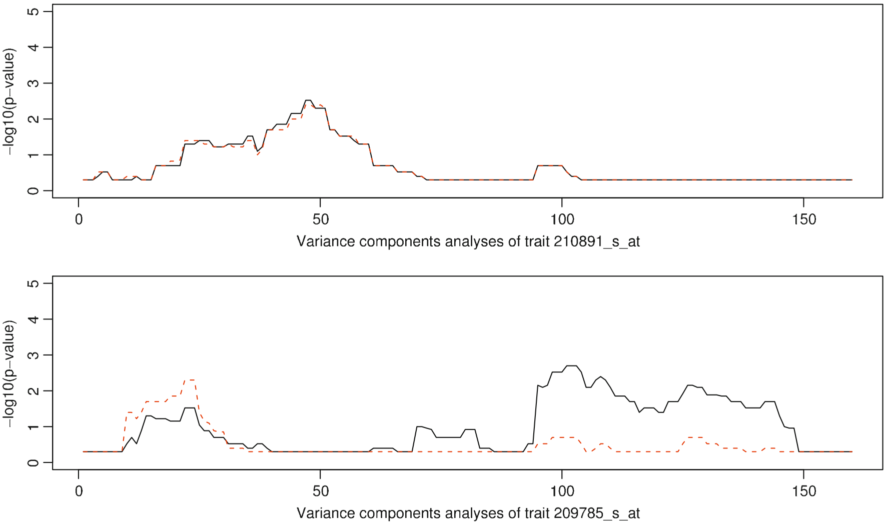


Fig. 4 Results of VC method
$-\log_{10}$ p-values of two probe sets of 160 markers on chromosome 5, using VC analyses. The *dotted red lines* represent results for the ENQT-transformed dataset

samples when various ascertainment approaches were applied. Failing to perform any ascertainment correction caused little effect on the linked component of variance and minimal bias in its estimate, but caused profound biases in the estimates of the unlinked genetic and residual VC estimates. This finding suggests that requiring the user to specify an overall heritability of the trait (as required by SOLAR) is not optimal when studying nonrandomly selected pedigrees. Alternative choices for ascertainment correction include conditioning on the trait values exceeding a threshold, which is statistically efficient if the threshold has been defined and followed carefully, or conditioning on the trait values of the probands. Results obtained by [30] showed little difference in results between these two corrections if sampling probands from the tail of the distribution (e.g., upper 5% of trait values), and either method greatly reduced biases in the nongenetic and unlinked genetic VC estimates. Ascertainment correction procedures are available in ACT. Ascertainment corrections that are available include correcting for selection of one or two individuals per family.

An alternative procedure for ascertained samples groups the individuals according to whether they exceed user-specified thresholds, and then performs linkage analysis on the dichotomized data to evaluate whether concordantly extreme relates show similar marker IBD and discordantly extreme individuals show dissimilar IBD [93]. Since much of the genetic information useful for a linkage study is indicated by those subjects with extreme values, this approach to analysis was a reasonable approach for a preliminary assessment of linkage when marker genotyping was expensive. An alternative procedure [79] described in the Sect. 2.4 regresses

IBD sharing onto trait variation and therefore conditions on the ascertainment process. This method can be applied effectively when studying samples that have been collected following complex ascertainment procedures, but requires that the user specify population parameters such as the overall population mean, variance, and heritability.

2.4 Model-Free Variance Regression

A different procedure was proposed by [79] in which the proportion of alleles shared identical by descent among relatives is regressed onto a trait similarity and dissimilarity matrix. Because the trait is conditioned upon, this approach should be less affected by departures from normality than procedures that condition the trait on marker similarity. Furthermore, the procedure can in principle be applied to samples that have been selected because of extreme phenotypes, provided that a valid population mean, variance, and heritability of the trait are all available.

The regression equation proposed by [79] is

$$\hat{\mathbf{\Pi}}_{\mathbf{c}} = \mathbf{\Sigma}'_{\mathbf{Y}\hat{\mathbf{\Pi}}}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_{\mathbf{c}} + \mathbf{e},$$

where $\hat{\mathbf{\Pi}}_{\mathbf{c}}$ is the mean-centered vector of pairwise IBD sharing proportions calculated as $\hat{\mathbf{\Pi}}_{\mathbf{c}} = \hat{\mathbf{\Pi}} - \mathbf{E}(\mathbf{\Pi})$, and $\mathbf{Y}_{\mathbf{c}}$ is the mean-centered vector of stacked pairwise squared sums and squared differences of standardized traits $(S_{ij} = (X_i + X_j)^2$, $(D_{ij} = (X_i - X_j)^2$ for $i \neq j$. The vector of squared sums, **S**, and the vector of squared differences, **D**, are collinear for families containing more than two sibs, since each element of **S** and **D** is a linear combination of two squares and a cross-product, and there are n squares and n(n−1)/2 cross-products (overall n(n+1)/2 elements), whereas there are $m = n\,(n-1)\,/2$ elements in each of the vectors **S** and **D** (corresponding to the number of pairs among n individuals) [79]. To remove this collinearity between the vectors **S** and **D**, the latter are trimmed by removing the last n(n−3)/2 elements from it, retaining exactly n elements. This ensures that collinearity is removed, while each individual is represented at least once. The trimmed vector **D** is denoted **d**. Thus, vector **Y**, defined as $\mathbf{Y} = [\mathbf{S}, \mathbf{d}]'$, has m+n elements, instead of 2m, because of the trimming of **D**; $\mathbf{Y}_{\mathbf{c}} = \mathbf{Y} - E(\mathbf{Y})$. $\mathit{\Sigma}_{\mathbf{Y}}$ is the variance–covariance matrix of the vector **Y**, and $\mathbf{\Sigma}_{\mathbf{Y}\hat{\mathbf{\Pi}}}$is the covariance matrix of stacked $\mathbf{\Sigma}_{\mathbf{S}\hat{\mathbf{\Pi}}}$and $\mathbf{\Sigma}_{\mathbf{d}\hat{\mathbf{\Pi}}}$. The statistic used in the final linkage testing is denoted as T,

$$T = \hat{Q}\sum_{i=1}^{k}[\mathbf{B}'\hat{\mathbf{\Pi}}_{\mathbf{c}}]_i = \hat{Q}^2\sum_{i=1}^{k}[\mathbf{B}'\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}\mathbf{B}]_i,$$

where k is the number of pedigrees and $\hat{Q}$ is the phenotypic variance explained by the additive effects of the QTL, a scalar weighted across all pedigrees and calculated as

$$\hat{Q} = \frac{\sum_{i=1}^{k} \left[\mathbf{B}'\hat{\mathbf{\Pi}}_{\mathbf{c}}\right]_i}{\sum_{i=1}^{k} \left[\mathbf{B}'\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}\mathbf{B}\right]_i},$$

where $\mathbf{B} = \mathbf{H}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_{\mathbf{c}}$, and $\mathbf{H}$ constitutes a matrix composed of two blocks stacked horizontally, the first block being an m×m square matrix with diagonal elements 2 and off-diagonal elements 0, the second block being an m×n matrix subtracted from a similar square matrix with diagonal elements –2. $\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}$is the variance–covariance matrix of the IBD sharing proportion vector $\hat{\mathbf{\Pi}}$.

Simulation studies conducted by [79] showed that, when the trait mean and variance are correctly specified, this procedure has power equivalent to VC methods and performs better than VC procedures for nonnormal data. [45] studied an extension of this method that allows for genetic imprinting (alleles transmitted from one parent are not expressed in the child) and showed that in general the method is very sensitive to accurate mean and variance specification, but relatively insensitive to the specified heritability. Another issue in the application of this method as it has to date been derived is that it does not allow the analyst to jointly adjust for covariates, so that these effects would have to be removed statistically before analysis can proceed. A procedure for model-free regression analysis is available in Merlin [1], and analysis can be performed efficiently using multipoint data. Application of this approach is shown in Fig. 5.

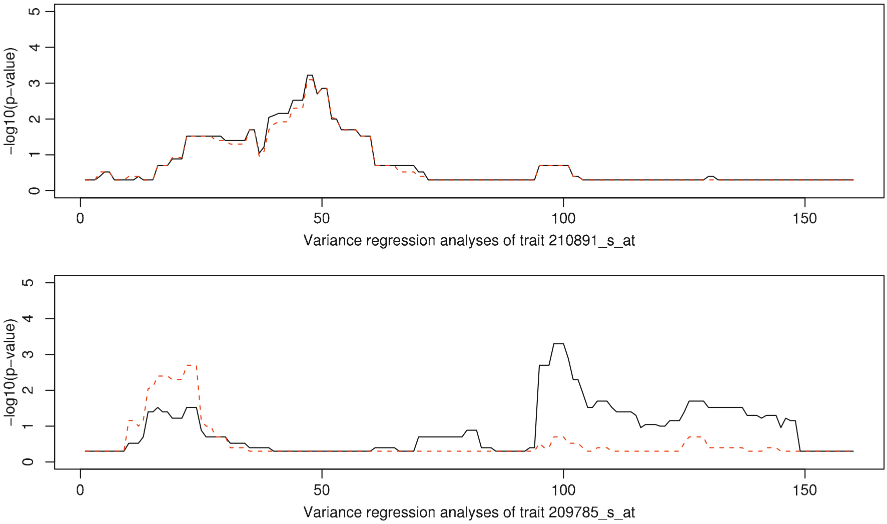


Fig. 5 Result of variance regression
$-\log_{10} p$-values of two probe sets of 160 markers on chromosome 5, using variance regression analyses. The *dotted red lines* represent results for the ENQT-transformed dataset

2.5 Multivariate Models

FEML methods have been difficult to apply for multivariate phenotypes because of the large number of parameters that need to be modeled. However, model-free methods are readily modified for multivariate data. Let $X_f = (X_{11}, \ldots, X_{1n}, \ldots, X_{mn})'$ be a vector of m multivariate trait values for n members of the fth family. Let N be the total number of families, β a vector of dimension mk of the regression coefficients for the k covariates; $Z_f = I_m \otimes Z_{n \times m}$ an $mn \times mk$ known matrix of covariate values for the fth family; V_f a variance–covariance matrix of dimension $mn \times mn$, with $V_f = A \otimes G_f + B \otimes \pi_f + C \otimes I_f$; G_f the $n \times n$ matrix of the coefficients of relationship for the family; π_f an $n \times n$ matrix of the estimated proportion of alleles IBD for pairs of related individuals for the fth family; I_f the $n \times n$ identity matrix; and A, B, and C, respectively, polygenic, major gene, and residual variance–covariance matrices each of dimension $m \times m$. The log likelihood for the data, then, assuming that it arises from a multivariate normal distribution, is

$$L_f = \sum_{f=1}^{N} - \left\{ \frac{1}{2}\ln(V_f) - \frac{k}{2}\ln(2\pi) - \frac{1}{2}(X_f - Z_f\beta)'V_f^{-1}(X_f - Z_f) \right\}.$$

Simulations to evaluate the characteristics of multivariate VC procedures have been completed [13, 33, 63]. These simulations document improvement in power for most configurations of genetic covariances by using multivariate phenotypes, and particularly when covariances are oppositely signed because of linked and unlinked genetic factors. Programs that can perform multivariate analysis using VC approaches include Mendel [57], ACT [6, 30], and SOLAR [3]. [84] provided an elegant development of multivariate models for application to structural equations modeling, with an emphasis on twins or sib-pairs. Structural equation models have been extensively developed for analysis of twin pairs [19, 67], and these approaches can assess evidence for linkage while providing very intricate partitioning of variance.

A regression expression can be formulated to develop a multivariate Haseman–Elston test, as follows:

$$E\left(\sum_{k=1}^{p} (c_k(X_{ik} - X_{jk}))^2 | \pi_{ij} \right) = \alpha + \beta\pi_{ij}.$$

Estimation of the coefficients c must be subject to constraints to ensure that total variance of the expression is constant. A conservative test statistic can be formed in the usual manner for multivariate regression by comparing the ratio of the hypothesis and error sums of squares to an F-distribution having $m - 1$ and $n - m - 1$ degrees of freedom. Comparisons that were performed by simulation of multivariate tests are discussed later. [35] suggested performing a preliminary principal components analysis when applying a similar procedure to the covariances among sib-pairs instead of the sib-pair differences. The purpose of principal components analysis is to create variables that are pairwise independent and so simplify the constraints on the parameters c. However, principal components analysis could also be effective in eliminating collinearity in the data. To assess collinearity of the data, the value of each of the eigenvalues from the variance–covariance matrix is assessed. Collinear variables are extracted in the first few principal components (eigenvectors) and can be retained for analysis, while the last principal components might be discarded. [61] proposed an alternative approach for modeling multivariate data, in which they first perform principal components analysis of the traits. The derived principal components are independent, and these can be tested separately.

One issue in modeling multivariate data is how to perform hypothesis testing. As suggested by the statistical literature [77], the VC parameters are routinely constrained to be nonnegative. However, this introduces complexities for hypothesis testing [78]. In the univariate case, the distribution of the likelihood ratio under the null hypothesis is a 1:1 mixture of χ_1^2 and χ_1^2 distributions, so that the p-value can be obtained by simply multiplying the p-value from comparison with a χ_1^2 distribution by $\frac{1}{2}$. For bivariate data, however, the mixing distribution becomes complex mixtures of chi-square distributions. For principal components analysis applied to multivariate data from model organisms [61] or from application of the Haseman–Elston test [36], there are no random components, and the distribution of test statistics follows a binomial expansion of the number of parameters times chi-square distributions with the appropriate degrees of freedom. In the bivariate case, for example, the distribution of the sum of test statistics for principal components is given by a 1:2:1 mixture of χ_2^2, χ_1^2, and χ_0^2 distributions. When studying bivariate data from humans that include random effects, the mixture distribution has not been derived but has an upper bound described by a 1:2:1 mixture of χ_3^2, χ_1^2, and χ_0^2 distributions. [4] suggested constraining the major gene covariance to ± 1 to reflect the pleiotropic effects of a major gene. Theoretically, this constraint results in a 1:1 mixture of chi-squared distributions having, respectively, one and two degrees of freedom (B. Mangin, 2008, personal communication).

Simulation studies comparing multivariate Haseman–Elston and VC methods are available in [7]. These results show that the unconstrained and constrained (pleiotropic) VC models have similar power that is superior to that of the multivariate Haseman–Elston and univariate tests for most models. Although the power of the multivariate Haseman–Elston method was less than that of the VC method, it was still greater for most models than the univariate VC test and also is computationally rapid. Power was greatest when the polygenic and major genic correlations in traits were opposite in sign, as was previously noted for multivariate mapping studies of inbred mouse lines [50]. Interestingly, when the correlations are equal and opposite, the marginal correlation between traits is zero. Thus, the naive investigator cannot rely on the observed correlation among traits to choose traits for study using multivariate tests. Because the multivariate Haseman–Elston test is computationally rapid and more powerful than univariate VC tests when the major gene and polygenic effects have opposite signs, it can still be used as a rapid screen of traits to identify those warranting further study with multivariate VC tests. Multivariate VC analysis was applied to the two traits under study in Fig. 6. As the two traits being

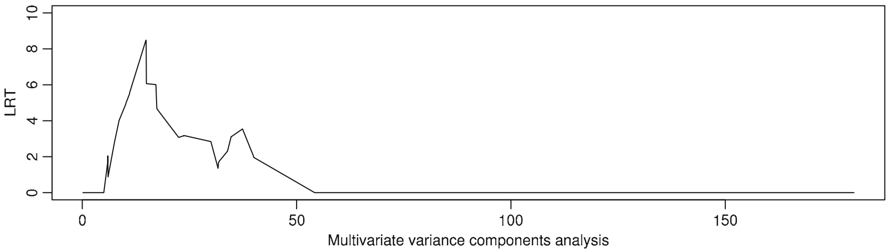


Fig. 6 Result of multivariate VC
Likelihood ratio test of two probe sets of 160 markers on chromosome 5, using multivariate VC analysis

studied in this example do not have known biological relationships, the results from this analysis do not increase evidence for linkage, as can be seen in Table 2 which shows no increase in Wald test for linkage and major gene variance components.

2.6 *Joint Linkage and Association Analysis*

The statistical methods for linkage analysis discussed so far assume linkage equilibrium between the trait and marker loci. If linkage disequilibrium is conjectured with one or several of the marker loci and the trait locus, then specialized procedures are required to partition variance among the main effects due to the marker locus and residual effects of the genetic locus in the region on interindividual variability, which is reflected by effects on the VC. [6] provided a framework by which variability could be partitioned according to effects from the marker on the trait and effects not associated with a marker locus due to top linkage disequilibrium on interindividual variability. [41] developed an approach that performed association analysis by partitioning variability to within sibship and among sibship variability. This approach allows identification of associations conditional on the parental genotypes and performance of association analysis in families allowing for potential confounding due to population stratification. The approach has been further developed and implemented in Merlin [1]. Versions of the transmission disequilibrium test for quantitative traits, in which association due to a marker locus is performed conditionally on the marker genotypes of the parents, were developed further by [2] and are reviewed further in the chapter “Family-Based Association Studies”.

3 Discussion

A wide variety of analytical approaches have been described in this chapter for linkage analysis in humans and other outbred animals. As shown by their application to data from probe sets, results from applying different approaches can yield somewhat contradictory results, according to the requirements and assumptions of each method. Therefore, we suggest a hierarchy to the analysis of data. In our experience, the first step in the analysis of quantitative data should consist of a preliminary descriptive evaluation of the distribution of the data. Are there extreme outliers? If so, a first step is to check that these outliers do not reflect coding errors or other causes for unreliable data. If extreme outliers do exist in the data, do they cluster within a few families? If so, these families may become highly influential in the analysis unless transformations are effected, such as the ENQT transformation described here. While ENQT is simple to apply, it should be remembered that in genetic studies it is occasionally the most extreme subjects who are informative, and further studies of such extreme subjects may be warranted.

[30] have developed procedures and tools for identifying and characterizing the origin of linkage signal for quantitative traits, so that influential families, if they exist, can be identified. In family studies, incorrect assignment of familial relationships can lead to invalid linkage inferences. For example, if a large sibship is studied and one unrelated individual having different trait values from the rest of the family is specified as a sibling, then strong false evidence for linkage will occur at all locations in which the indentity is inferred by descent information suggests no alleles IBD with the other siblings (which could be much of the genome for a multiallelic marker). Thus, an additional safeguard against false inference is to preliminarily check that the reported relationships among subjects are accurately specified, using available software programs like PREST [64] or RelCheck (see Web Resources).

Another preliminary assessment that should be considered is characterization of the heritability of the trait being studied. Standard VC programs like ACT, SOLAR, or MERLIN will provide estimates of heritability. Traits showing low heritability (e.g., less than 10%) are difficult to map using linkage analysis procedures unless very large samples or very extensive pedigrees are studied. If nongenetic covariates have been studied, do these affect interindividual covariation? If so, either they must be adjusted for prior to analysis or a program that jointly adjusts for these effects should be used. It is also useful to check whether sex affects the trait of interest (and it can be treated as a covariate in analyses of autosomes). In analysis of CEPH cell lines, as another example, it has been noted that samples from Yorubans show generally higher expression levels of probe sets than samples from Utah, perhaps reflecting the older derivation of the Utah samples.

Once quality control procedures are completed, further studies may proceed. If sampling of the families was not based upon any particular phenotype and the data are approximately normally distributed, VC analysis might be done next, without transformation. Analyses could be conducted either by using a maximum-likelihood based approach or by applying score-based approaches. Results of such approaches will be similar for normally distributed data, but the score-based approaches are generally not well developed for general pedigrees. If the data are nonnormally distributed, then the analyst should consider performing more robust analyses first, such as the Haseman–Elston approach or a score-based procedure or variance regression, provided data are restricted to nuclear families. If the data comprise extended pedigrees and the distribution is nonnormal, then a transformation of the data such as ENQT can be considered. While this transformation ensures normality and protects against an excess of false-positive inferences, it may lead to a loss of information for linkage if those extreme individuals are segregating uncommon alleles that influence the trait. For approximately normal data, VC procedures can be applied to identify one or more loci influencing the trait. If multiple loci appear to be influencing the trait of interest, then analyses to characterize the relationships among the traits can be conducted. MCMC procedures such as those implemented in LOKI can be applied to characterize effects from multiple loci.

If the data arise from a selected sample, then VC procedures that allow for ascertainment can be applied, provided the selection process uses at most two individuals in the families. If more than two subjects per family were used in the selection process, then the variance regression approach must be used. Implementing variance regression requires that population means and variances for the trait be specified, and so the analyst must request this information from the data provider.

If multivariate data are available, the analyst may consider performing multivariate analysis. The easiest way to perform preliminary multivariate analysis is to first perform principal component analysis and then to conduct individual univariate analyses on the principal components. Once a genetic region of interest is identified, subsequent multivariate VC analysis can be conducted to characterize the linked and unlinked components of variance. The multivariate version of the Haseman–Elston procedure permits a more robust analysis to be conducted to check that results are not unduly influenced by extreme observations and to narrow analysis, if multiple traits are being analyzed to a few traits that show higher loadings during the analysis. Once one or more genetic regions have been identified, further studies can be considered using multivariate linkage analysis to estimate the components of variance due to linkage of each trait and the covariance.

4 Web Resources

ACT:

- http://www.epigenetic.org/Linkage/act.html
- http://www.epigenetic.org/Linkage/act.tar.gz

LOKI:

- http://www.stat.washington.edu/thompson/Genepi/Loki.shtml
- http://loki.homeunix.net

LINKAGE:

- ftp://linkage.rockefeller.edu/software/linkage

MERLIN:

- http://www.sph.umich.edu/csg/abecasis/merlin

SAGE:

- http://darwin.cwru.edu

SOLAR:

- http://solar.sfbrgenetics.org

PREST:

- http://www.stat.uchicago.edu/˜mcpeek/software/prest

RelCheck:

- http://www.biostat.wisc.edu/˜kbroman/software/#relcheck

Genetic Power Calculator:

- http://pngu.mgh.harvard.edu/˜purcell/gpc

Genetic Analysis Workshop:

- http://www.gaworkshop.org, (supported by R01 GM031575)

Kyoto Encyclopedia of Genes and Genomes:

- http://www.genome.jp/kegg

Gene Ontology:

- http://www.geneontology.org

References

1. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics 30(1):97–101
2. Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J (1999) Testing the robustness of the likelihood ratio test in a variance-component quantitative trait loci (QTL) mapping procedure. American Journal of Human Genetics 65:531–544
3. Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. American Journal of Human Genetics 62:1198–1211
4. Almasy L, Dyer TD, Blangero J (1997) Bivariate quantitative trait linkage analysis: Pleiotropy versus co-incident linkages. Genetic Epidemiology 14:953–958
5. Amos CI (1988) Robust methods for detection of genetic linkage for data from extended families and pedigrees. PhD dissertation, Louisiana State University Medical Center-New Orleans
6. Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. American Journal of Human Genetics 54:535–543
7. Amos CI, de Andrade M (2001) Genetic linkage methods for quantitative traits. Statistical Methods for Medical Research 10:3–25
8. Amos CI, Elston RC (1989) Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genetic Epidemiology 6:349–361

9. Amos CI, Dawson DV, Elston RC (1990) The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. American Journal of Human Genetics 47:842–853
10. Amos CI, Zhu D, Boerwinkle E (1996) Assessing genetic linkage and association with robust components of variance approaches. Annals of Human Genetics 60:143–160
11. Amos CI, Krushkal J, Thiel TJ, Young A, Zhu DK, Boerwinkle E, de Andrade M (1997) Comparison of model-free linkage mapping strategies for the study of a complex trait. Genetic Epidemiology 14:743–748
12. Amos CI, Gu X, Chen J, Davis BR (2000) Least squares estimation of variance components for linkage. Genetic Epidemiology 19(Supplement 1):S1–S7
13. Amos CI, de Andrade M, Zhu D (2001) Comparison of multivariate tests for genetic linkage. Human Heredity 51:133–144
14. Barnholz JS, de Andrade M, Page GP, King TM, Peterson LE, Amos CI (1999) Assessing linkage of monoamine oxidase b (MAOB) in a genome-wide scan of 285 markers on 22 chromosomes using univariate variance components approach. Genetic Epidemiology 17(Supplement):S49–54
15. Blangero J, Williams JT, Almasy L (2001) Variance component methods for detecting complex trait loci. Advances in Genetics 42:151–181
16. Boerwinkle E, Sing CF (1987) The use of measured genotype information in the analysis of quantitative phenotypes in man: III. Simultaneous estimation of the frequencies and effects of the apolipoprotein e polymorphism and residual polygenetic effects on cholesterol, betalipoprotein, and triglyceride levels. Annals of Human Genetics 51:211–226
17. Boerwinkle E, Chakraborty R, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man: I. Models and analytical methods. Annals of Human Genetics 50:181–194
18. Boerwinkle E, Leffert CC, Lin J, Lackner C, Chiesa G, Hobbs HH (1992) Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. Journal of Clinical Investigation 90:52–60
19. Cardon LR, Fulker DW (1994) The power of interval-mapping of quantitative trait loci using selected sib pairs. American Journal of Human Genetics 55:825–833
20. Chen WM, Abecasis GR (2006) Estimating the power of variance component linkage analysis in large pedigrees. Genetic Epidemiology 30:471–484
21. Chen WM, Broman KW, Liang KY (2004) Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genetic Epidemiology 26:265–272
22. Chen WM, Broman KW, Liang KY (2005) Power and robustness of linkage tests for quantitative traits in general pedigrees. Genetic Epidemiology 28(1):11–23
23. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen K, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nature Genetics 33: 422–425
24. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369
25. Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J (1986) Effects of misspecifying genetic parameters in LOD score analysis. Biometrics 42:393–399
26. Cottingham R, Idury RM, Schäffer AA (1993) Faster sequential genetic linkage computations. American Journal of Human Genetics 53(1):252–263
27. Crow JF (1986) Basic concepts in population, quantitative, and evolutionary genetics. W.H. Freeman and Co., New York
28. Curtis D, Sham PC (1994) Using risk calculation to implement an extended relative pair analysis. Annals of Human Genetics 58:151–162
29. Davis S, Schroeder M, Goldin LR, Weeks DE (1996) Nonparametric simulation-based statistics for detecting linkage in general pedigrees. American Journal of Human Genetics 58:867–880

30. de Andrade M, Amos CI, Thiel T (1999) Methods to estimate the genetic component of variance for quantitative traits in families. Genetic Epidemiology 17:64–76
31. Diao G, Lin DY (2005) A powerful and robust method for mapping quantitative trait loci in general pedigrees. American Journal of Human Genetics 77:97–111
32. Drigalenko E (1998) How sib-pairs reveal linkage. American Journal of Human Genetics 63:1243–1245
33. Eaves LJ, Neale MC, Maes H (1996) Multivariate multipoint linkage analysis of quantitative trait loci. Behavior Genetics 26:519–525
34. Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Human Heredity 21:523–542
35. Elston RC, Buxbaum S, Jacobs KB, Olson JM (1998) Haseman and elston revisited. Presented at the International Genetic Epidemiology Society
36. Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genetic Epidemiology 19(1):1–17
37. Fisher RA (1918) The correlation in relatives on the supposition of Mendelian inheritance. Transactions of the Royal Statistical Society, Edinburgh 52:399–433
38. Fisher RA (1935) The detection of linkage with autosomal dominant abnormalities. Annals of Eugenics 6:187–201
39. Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. American Journal of Human Genetics 54:1092–1103
40. Fulker DW, Cherny SS (1996) An improved multipoint sib-pair analysis of quantitative traits. Behavior Genetics 26:527–532
41. Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. American Journal of Human Genetics 64:259–267
42. Gessler DD, Xu S (1996) Using the expectation or the distribution of the identity by descent for mapping quantitative trait loci under the random model. American Journal of Human Genetics 59:1382–1390
43. Goldgar DE, Oniki RS (1992) Comparison of multipoint identity-by-descent method with parametric multipoint linkage analysis for mapping quantitative traits. American Journal of Human Genetics 50:598–606
44. Göring HH, Williams JT, Blangero J (2001) Linkage analysis of quantitative traits in randomly ascertained pedigrees: comparison of penetrance-based and variance component analysis. Genetic Epidemiology 21 Supplement:S783–S788
45. Gorlova OY, Weng S, Zhang Y, Amos CI, Spitz MR (2007) Aggregation of cancer among relatives of never-smoking lung cancer patients. International Journal of Cancer 121(1): 111–118
46. Guerra R, Wan Y, Jia A, Amos CI, Cohen JC (1999) Testing for linkage under robust genetic models. Human Heredity 49(3):146–153
47. Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3–19
48. Heath SC (1997) Markov Chain Monte Carlo segregation and linkage analysis for oligogenic models. American Journal of Human Genetics 61:748–760
49. Hopper JL (1993) Variance components for statistical genetics: applications in medical research to characteristic related to diseases and health. Statistical Methods in Medical Research 2:199–223
50. Jiang C, Zeng ZB (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140:1111–1127
51. Kammerer CM, MacCluer JW (1985) Comparison of two preliminary methods of quantitative linkage analysis. Human Heredity 35:319–325
52. Kempthorne O (1957) An Introduction to Genetic Statistics. Wiley, New York
53. Kempthorne O, Horner TW (1955) The theoretical correlations of relatives in random mating populations. Genetics 40:153–167
54. Kong A, Cox NJ (1997) Allele sharing models – LOD scores and accurate linkage tests. American Journal of Human Genetics 61:1179–1188

55. Kruglyak L, Lander ES (1995) Complete multipoint sib pair analysis of qualitative and quantitative traits. American Journal of Human Genetics 57:439–454
56. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences USA 84:2363–2367
57. Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E (2001) Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. American Journal of Human Genetics 69(Supplement):504
58. Lathrop GM, Lalouel JM, Julier C, Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. American Journal of Human Genetics 37(3):482–498
59. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
60. Ma J, Amos CI, Daw EW (2007) Ascertainment correction for Markov chain Monte Carlo segregation and linkage analysis of a quantitative trait. Genetic Epidemiology 31(6):594–604
61. Mangin B, Thoquet P, Grimsley N (1998) Pleiotropic QTL analysis. Biometrics 54:88–99
62. Marlow AJ, Fisher SE, Francks C, MacPhie IL, Cherny SS, Richardson AJ, Talcott JB, Stein JF, Monaco AP, Cardon LR (2003) Use of multivariate linkage analysis for dissection of a complex cognitive trait. American Journal of Human Genetics 72(3):561–570
63. Martin N, Boomsma D, Machin G (1997) A twin-pronged attack on complex traits. Nature Genetics 17:387–392
64. McPeek MS, Sun L (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. American Journal of Human Genetics 66:1076–1094
65. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:733–734
66. Neale BM, Ferreira MAR, Medland SE, Posthuma D (eds) (2007) Statistical Genetics: Gene Mapping Through Linkage and Association. Taylor and Francis, London
67. Neale MC, Cardon LR (1992) Methodology for the Study of Twins and Families (NATO ASI Series D: Behavioral and Social Sciences, Vol. 67). Kluwer Academic, Dordrecht, The Netherlands
68. Nelder JA, Pregibon D (1987) An extended quasi-likelihood function. Biometrika 74:221–232
69. Olson JM, Wijsman E (1993) Linkage between quantitative trait and marker locus: Methods using all relative pairs. Genetic Epidemiology 10:87–102
70. Ott J (1999) Analysis of Human Genetic Linkage, 3rd edn. Johns Hopkins University Press, Baltimore, MD
71. Page GP, Amos CI, Boerwinkle E (1998) Exclusion and linkage using the QLOD approach. American Journal of Human Genetics 62:962–968
72. Peng B, Yu RK, DeHoff KL, Amos CI (2007) Normalizing a large number of quantitative traits using empirical normal quantile transformation. BMC Proceedings 1(Supplement 1):S156
73. Penrose LS (1938) Genetic linkage in graded human characters. Annals of Eugenics 8:233–238
74. Penrose LS (1946) A further note on the sib-pair linkage method. Annals of Eugenics 13:25–29
75. Prentice RL, Zhao LP (1991) Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47:825–839
76. Schork NJ (1993) Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency power, and modeling considerations. American Journal of Human Genetics 53:1306–1319
77. Searle SR, Casella G, McCulloch CE (1992) Variance Components. Wiley, New York
78. Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. Journal of American Statistical Association 82:605–610
79. Sham PC, Purcell S, Cherny SS, Abecasis GR (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. American Journal of Human Genetics 71:238–253
80. Shete S, Amos CI, Hwang SJ, Strong LC (2002) Individual-specific liability groups in genetic linkage, with applications to kindreds with Li-Fraumeni syndrome. American Journal of Human Genetics 70(3):813–817

81. Shete S, Jacobs KB, Elston RC (2003) Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. Human Heredity 55:79–85
82. Sung YJ, Thompson EA, Wijsman EM (2007) MCMC-based linkage analysis for complex traits on general pedigrees: multipoint analysis with a two-locus model and a polygenic component. Genetic Epidemiology 31(2):103–114
83. Tiwari HK, Elston RC (1997) Deriving components of genetic variance for multilocus models. Genetic Epidemiology 14(6):1131–1136
84. Todorov AA, Vogler GP, Gu C, Province MA, Li Z, Heath AC, Rao DC (1998) Testing causal hypotheses in multivariate linkage analysis of quantitative traits: general formulation and application to sibpair data. Genetic Epidemiology 15:263–278
85. Wang J, Guerra R, Cohen J (1998) A statistically robust variance component approach for quantitative trait linkage analysis. Annals of Human Genetics 62:349–359
86. Wang J, Guerra R, Cohen J (1999) Least squares estimation of variance components for linkage. Annals of Human Genetics 63:249–262
87. Wang K, Huang J (2002) A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. American Journal of Human Genetics 70:412–424
88. Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127
89. Wijsman EM, Amos CI (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: Summary of GAW10 contributions. Genetic Epidemiology 14:719–735
90. Williams JT, Blangero J (1999) Power of variance component linkage analysis to detect quantitative trait loci. Annals of Human Genetics 63:545–563
91. Wright FA, Kong A (1997) Linkage mapping in experimental crosses: the robustness of single-gene models. Genetics 146:417–425
92. Yu R, DeHoff K, Amos CI, Shete S (2007) Seeking gene relationships in gene expression data using support vector machine regression. BMC Proceedings 1(Supplement 1):S51
93. Zhang H, Risch N (1996) Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: selected sampling by parental phenotypes. American Journal of Human Genetics 59:951–957

Markov Chain Monte Carlo Linkage Analysis Methods

Robert P. Igo, Jr., Yuqun Luo, and Shili Lin

Abstract As alluded to in the chapter "Linkage Analysis of Qualitative Traits", neither the Elston–Stewart algorithm nor the Lander–Green approach is amenable to genetic data from large complex pedigrees and a large number of markers. In such cases, Monte Carlo estimation methods provide a viable alternative to the exact solutions. Two types of Monte Carlo methods have been developed for linkage analysis, haplotype inference, and other kinds of genetic analysis. They are Markov chain Monte Carlo (MCMC) methods and Monte Carlo methods that are based on independent samples. Approaches based on Markov chain Monte Carlo methods are more widely applicable; there is practically no limit on the size or complexity of the pedigrees, nor on the number of markers to be considered simultaneously. In this chapter, we will review the basic principles of MCMC methods for multipoint linkage analysis with extended pedigrees. Both simulations and application to data from the Framingham study will be used to compare three MCMC software packages: LOKI, MORGAN, and SIMWALK.

1 Introduction

The search for genetic risk factors of human diseases has focused squarely on complex traits, which run in families but which are not inherited in simple Mendelian fashion. It has been a decade-long notion, since being foreseen by Risch and Merikangas [38], that genetic association analysis offers the best hope of mapping common genetic variants with small effects. Although genomewide association scans have indeed surged recently to popularity as technological advances have made large-scale genotyping more feasible, linkage analysis remains an important tool. It can be used to identify broad genomic regions that harbor genetic variants contributing to the genetic influence on the trait. Such information may be used in follow-up candidate region association mapping or in providing a corroborating line

S. Lin (✉)
Department of Statistics, The Ohio State University.
e-mail: shili@stat.osu.edu.

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_5, © Springer-Verlag Berlin Heidelberg 2010

of evidence to confirm the existence of genetic influences from genomewide association scans. Moreover, linkage analysis offers better power than association analysis to detect rare disease-causing variants with relatively large effect, which may also play a role in complex diseases.

To detect loci contributing to complex traits by linkage analysis, it is crucial to collect and efficiently analyze data that contain maximum information on inheritance. In practice, this entails the use of extended pedigrees, when available, and analysis of multiple markers simultaneously [50]. There are two main categories of approaches to linkage analysis: model-based or Identity-by-Descent (IBD) based, with the former requiring specification of the mode of inheritance and the latter free of this requirement. When the penetrance model can be specified correctly, model-based methods are more powerful and offer greater resolution in localizing disease genes [2]. IBD-based methods, on the other hand, are more robust to model misspecification, but usually at the expense of reduced power [27]. Computational demands, however, limit either the size of pedigrees or the number of markers that can be incorporated into a multipoint linkage analysis with exact computation of model-based LOD scores, IBD sharing statistics, or IBD-based likelihoods. In this chapter, we focus on model-based methods and related issues.

The two widely used approaches for exact calculation of multipoint likelihoods on pedigrees (leading to LOD scores) are the Elston–Stewart algorithm [11] and its extensions, first implemented in the program LINKAGE [30] for multipoint analysis and refined in FASTLINK [7] and VITESSE [36], and the Lander–Green algorithm [28], implemented in GENEHUNTER [27]. In recent years, MERLIN [1] has emerged as the package of choice for many past GENEHUNTER users, partly due to its versatility and its computational efficiency, as discussed later. The computational complexity of Elston–Stewart algorithm increases in a linear fashion with the number of pedigree members, but exponentially with the number of markers. Multiallelic markers, such as microsatellites, require a vast amount of memory to allow for all possible haplotypes. Hence, on the one hand, the Elston–Stewart algorithm is usually restricted to analyzing only a few microsatellite markers jointly. The Lander–Green algorithm, on the other hand, is efficient in the number of markers but not in pedigree size: the computational demand is linear in the number of markers but exponential in the number of inheritance bits (defined as twice the number of non-founders minus the number of founders). GENEHUNTER was originally not recommended for use on pedigrees larger than 16 bits [27]. MERLIN uses a different algorithm, incorporating sparse gene flow trees, but faces similar limitations for pedigrees: memory requirement is still exponential in the number of inheritance bits, limiting the maximum pedigree size to 24 inheritance bits as its default in its latest release (MERLIN-1.1.1).

The program SUPERLINK, another exact method developed more recently, uses a Bayesian Networks representation and various tactics to optimize elimination order [12]. Consequently, it can often accommodate data of greater complexity and/or size than the above programs. But many real data sets are still too complex to be analyzed intact by any of the exact algorithms, especially when there are many markers and a high proportion of untyped individuals [9, 12].

When the size and/or the complexity of the data become prohibitive for exact calculations, Monte Carlo approaches provide viable alternatives. In a nutshell, Monte Carlo methods provide consistent estimates of quantities of interest through averaging over samples from some underlying distributions. There are two main classes of Monte Carlo methods employed in linkage analysis, depending on whether the samples are independently drawn or whether the samples form a Markov chain (dependent samples). The latter is usually referred to more specifically as Markov chain Monte Carlo (MCMC) rather than simply Monte Carlo.

Early Monte Carlo approaches such as gene dropping [35] take independent samples from a distribution that is not conditioned on observed phenotypic data, and as such, are rather inefficient. A brief account of the history of Monte Carlo methods (based on independent samples) in linkage analysis up to the early 1990s can be found elsewhere [31]. Sequential imputation is another Monte Carlo approach that does take observed data into account in the sampling distribution [24, 26]. In addition to LOD score computation, it has been extended to IBD-based linkage analysis [43] and haplotype analysis on pedigrees [32]. Many MCMC methods for both model-based and IBD-based linkage analysis have also been developed over the last decade; see Thompson [48, 49] for reviews and many references therein. In this chapter, we focus on comparing several MCMC methods on their performance in estimating LOD scores and drawing inferences based on them.

2 Test Data

2.1 Data from the Framingham study

Permission to use the Genetic Analysis Workshop (GAW) 13 Problem 1 data was obtained from the Framingham Heart Study (http://www.nhlbi.nih.gov/about/framingham/index.html). The Problem 1 sample comprises 330 multiplex pedigrees, including 2,464 individuals from Cohorts I and II for whom phenotype data were available [8]. Genome-wide short tandem repeat (STR) marker data on roughly 400 markers (Marshfield screening set 9) were available for 1,702 individuals. This sample includes numerous large families: the number of inheritance vector bits is greater than 32 for twelve pedigrees in the sample, and greater than 64 for one pedigree. In addition, two pedigrees contain marriage loops. Hence, linkage analysis on this data set poses a considerable computational challenge.

A continuous measure of lipid level, TH, was derived from the data on serum triglyceride (TG) and HDL-cholesterol (HDL) levels. Specifically, TH is the residual of linear regression of log(TG/HDL) on age, age^2, age^3, BMI, smoking, and alcohol use. High values of log(TG/HDL) are associated with the metabolic syndrome [18]. Shearman et al. [41] performed linkage analysis on log(TG/HDL) adjusted for a slightly different set of covariates. Although the adjustment for covariates was done differently, we refer to the trait used in Shearman et al.[41] also as "TH" in discussions hereafter for ease of reference. TH roughly follows a normal distribution. The two largest, outlying values were Winsorized downward to be equal to the third largest.

A binary trait, bTH, was also derived from TH by classifying the top quartile of TH values as "affected" and the remainder as "unaffected," following George et al. [13].

In a genomewide scan of 332 Framingham pedigrees (including the 330 pedigrees in our analysis) using variance-components methods, Shearman et al. [41] reported some evidence for linkage to TH near the q terminus of chromosome 7 (max. LOD score $=$ 2.5 near marker D7S2195, 155 Kosambi cM). Two groups participating in GAW13, who studied a similar phenotype on the 330 Framingham pedigrees, also identified regions on chromosome 7q with suggestive evidence for linkage [15, 20]. Thus, we based the comparison of the MCMC methods on multipoint linkage analysis performed on chromosome 7 (220 Haldane cM), where data on 22 microsatellite markers are available.

2.2 Simulated data

To explore in greater depth the similarities and differences, performances of the MCMC approaches were also compared through analysis of simulated data. A quantitative trait was simulated with one or two underlying QTLs of known locations and properties. These artificial data sets were patterned after the Framingham pedigrees, the 22 chromosome 7 microsatellite markers (locations and allele frequencies), and the TH phenotype. Trait values, together with marker genotypes, were generated for the 330 Framingham pedigrees, using the genedrop program in the MORGAN package. The pattern of missing data followed that in the real data. The trait models were those identified using the Loki package on the Framingham data, as detailed in Sect. 4.1. Specifically, for one-QTL simulated data, a recessive QTL situated at 190 cM accounted for 20% of the total variance, and 10% of the variance was assigned to an additive polygenic component. For two-QTL simulated data, two QTLs situated at 51.7 cM (QTL1, additive) and 190 cM (QTL2, recessive) accounted for 20% and 15% of the total variance, respectively, with an additive polygenic component accounting for 10% of the variance. The remaining 70% and 55% of the trait variability, for the one-QTL and the two-QTL models, respectively, was assumed to be due to non-genetic factors. Details of the QTL models are given in Table 1, including the allele frequency for the low-risk allele A (p_A), the mean values of the three genotypes ($\mu_{AA}, \mu_{AB}, \mu_{BB}$), the QTL genetic variance (σ^2_{g}), the additive polygenic variance (σ^2_{poly}), and residual variance (σ^2_{e}).

Some of the MCMC approaches were also applied to simulated binary traits. The binary indicators of affection status were derived from the simulated quantitative trait values by designating persons whose measurements fall on the top 25% as affected, following the same protocol as in the real data analysis.

Table 1 Quantitative trait simulation models

		Genotype Means			Var. Comp.		
Model	p_A	μ_{AA}	μ_{AB}	μ_{BB}	σ^2_{g}	σ^2_{poly}	σ^2_{e}
One-QTL (190 cM)	0.6	−0.195	−0.195	1.025	0.2	0.1	0.7
Two-QTL							
QTL1 (51.7 cM)	0.25	−1.095	−0.365	0.365	0.2	0.1	0.55
QTL2 (190 cM)	0.6	−0.169	−0.169	0.888	0.15		

3 MCMC Methods and Packages

The MCMC methodology is frequently employed in statistical applications in which exact calculations are infeasible. In MCMC, dependent samples forming a Markov chain are generated with the property that the sampling distribution will converge to the desired target distribution, which may be known only up to a constant. This limiting property is guaranteed if the Markov chain constructed is aperiodic and irreducible [39]. After discarding realizations in the initial, “burn-in”, period to allow the Markov chain to converge to the target distribution, the remaining samples can then be used to make inference about the quantities of interest. This general MCMC methodology has been adapted and tailored to linkage analysis. In this section, we provide a brief description of three publicly available MCMC linkage analysis packages: MORGAN (version 2.8.2), SimWalk2, and Loki (version 2.5.4).

The MORGAN suite of programs for pedigree analysis includes two MCMC linkage analysis programs: lm_markers and lm_bayes [16,47–49]. Both approaches compute LOD scores from a sample of inheritance vectors conditional on the marker data or all available data. Sampling of inheritance vectors is through an “LM sampler,” a block Gibbs sampler that consists of both an L-sampler (updating all meioses at one locus) and an M-sampler (updating all loci for one meiosis) [19, 51]. In lm_markers, the sampling is conditioned only on the marker data, and likelihoods are computed by the technique of Lange and Sobel [14, 29]. The LOD score is obtained in the usual fashion by comparing the likelihood at each proposed trait location to that of an unlinked trait locus. Either binary or quantitative trait data may be analyzed using lm_markers. The inheritance model for continuous traits accommodates an additive polygenic component in addition to a single diallelic quantitative trait locus (QTL). This approach has been extended to simultaneous mapping of two trait loci [46], but the implementation of the two-locus scheme was not officially released in the most current version of MORGAN. The pseudo-Bayesian approach of lm_bayes estimates LOD scores for a binary trait with a fully specified model through two rounds of MCMC sampling [13]. A pseudo-prior distribution is constructed in an initial phase of MCMC sampling, and in a second round of sampling from the posterior distribution, the LOD score is then estimated using the inverse of the pseudo-prior. George et al. [16] provide a concise summary of the two-phase MCMC procedure. As both the lm_markers and lm_bayes programs are based on the LM sampler, which requires each pedigree to be single-locus peelable, the applicability of the programs are limited to not-too-complex pedigrees.

SimWalk2, another model-based LOD-score method, calculates the likelihood in a fashion similar to lm_markers. It differs in using a random-walk algorithm to sample the inheritance vectors conditional on the marker data [29, 44, 45], which employs different types of transition matrices between nodes of the Markov chain. Unlike the LM sampler, this sampler is not restricted to single-locus-peelable pedigrees, hence it can deal with pedigrees of greater complexities. However, the trade-off is its slow convergence to the limit distribution, especially in "near reducibility" situations. SimWalk2 also provides heterogeneity LOD scores and the fraction α of linked pedigrees in addition to standard LOD scores. SimWalk2, like lm_bayes, applies only to binary trait data and inheritance models with only a single diallelic trait locus.

The Bayesian oligogenic MCMC program Loki carries out combined oligogenic segregation and linkage analysis on quantitative traits and employs an approach distinct from those of the LOD-score methods [19]. No prior specification of the mode of inheritance is required, as segregation analysis is conducted concurrently to, or in the absence of, linkage analysis. Reversible-jump MCMC [17] enables the number of QTLs in the overall model to change during the Loki MCMC run. Hence, estimation of the number of QTLs is possible as well as their locations – a distinct advantage in studying complex traits. Loki also uses the LM-sampler for drawing realizations from the Markov chain, and as such requires the pedigrees to be single-locus peelable. Although the Bayesian oligogenic approach is highly versatile, it has unique limitations. Prior distributions must be supplied for some parameters, such as the number of QTLs, and in general the quantity of data will not be sufficient to overwhelm their influence on the posterior distribution [52]. Because the dimensionality of the sample space is very high and continually changing, it is difficult to assess whether a representative sample from the parameter space has been obtained. Finally, inference under the Bayesian paradigm complicates interpretation and comparisons of results with other software, as no LOD scores or frequentist p values are available.

4 Comparison of Methods

4.1 Analysis Strategies

Among the programs considered, MORGAN and SimWalk2 require the prespecification of single-locus trait models to compute multipoint LOD scores. SimWalk2 and lm_bayes analyze only binary traits, while lm_markers can be applied to both quantitative and binary traits. In what follows, the same binary or quantitative trait models were applied to all the programs that require them. The penetrances of each of the three trait genotypes for the binary trait from the Framingham data, bTH, were derived from applying the same 75% trait value cutoff for affection status to TH as in Sect. 2.1 and the genotype means of the corresponding continuous trait models.

For simulated data, penetrances for binary trait data were similarly obtained from the true models used to generate the simulated continuous trait. These trait models were supplied to the linkage analysis software for simulated data. In what follows, we first provide some details on the estimation of the continuous trait models for TH, and then describe linkage anlayses performed by the three software packages.

4.1.1 Estimation of Segregation Models for TH

Two different approaches, in the absence of any marker data, were employed to estimate a segregation model for TH. One of them was based on Bonney's Class D regressive model [5], as implemented in the program SEGREG in the S.A.G.E. package, version 5.4 [40]. The two pedigrees containing marriage loops had to be omitted because they could not be handled by SEGREG.

The other approach was Bayesian oligogenic segregation analysis, which was performed using Loki [19]. Inference was based on the posterior distribution of the parameters. The effects of priors on the outcome have been investigated empirically [52] and the understanding is still evolving [22]. Furthermore, given the complexity of the Loki model, meaningful summarization of the output requires some thought on the part of the user. In our analysis, we specified the priors and summarized the outputs according to the recommendations contained in the above two references. The number of QTLs, k, was treated as a random variable, and the prior was set to be a Poisson distribution with mean 2 in our analysis. Each QTL was assumed to be diallelic, with the genotype effects ε_{AB} and ε_{BB} parametrized as the difference between mean trait values for the AB and the AA genotypes ($\mu_{AB} - \mu_{AA}$), and between those for the BB and the AA genotypes ($\mu_{BB} - \mu_{AA}$), respectively. Note that in this parametrization, AA is treated as the reference genotype. The allele A frequency, p_A, was assigned a prior Uniform [0,1] distribution. In summarizing the output from Loki, the high-risk allele was assigned label B. We specified independent normal prior distributions for ε_{AB} and ε_{BB} with mean 0 and variance τ_β. Wijsman and Yu [52] found that a poor selection of τ_β would result in slower convergence of the Markov chain. They suggested performing several short MCMC runs for a grid of values of τ_β, and to select for the prior the τ_β value that provided the greatest genetic variance for the QTLs. We followed this suggestion and used $\tau_\beta = 1$ for the Loki segregation analysis.

The estimates of $(p_A, \varepsilon_{AB}, \varepsilon_{BB})$ were obtained from the modes of the joint trivariate posterior density surface. Genotype effects were then converted to genotype means, with the constraint that the overall trait mean be 0. Specifically, in every iteration of the MCMC run, Loki reports a realization of $(p_A, \varepsilon_{AB}, \varepsilon_{BB})$ for each QTL in the current model. The aggregate of these realizations, pooled across all QTL models in all iterations, forms the posterior density for the model parameters. Because the segregation model from a given iteration may contain more than one QTL, or none at all, the posterior density is not a true probability distribution. Hence, standard Bayesian estimation from the posterior distribution is not possible. Instead, a graphical method is used. A surface or contour plot of $(\varepsilon_{AB}, \varepsilon_{BB})$ displays peaks of posterior density where major QTLs with similar parameter values were accepted in the model with high frequency. Estimates of $(\varepsilon_{AB}, \varepsilon_{BB})$ are obtained from each peak as the pair of genotype effects corresponding to the peak location, or posterior mode, and p_A for each of these models was obtained from the distribution of allele frequency realizations corresponding to these models in the MCMC runs. Loki provides, via the loki_ext.pl script included with the program, estimates of the total QTL variance σ_{g}^2 and residual variance σ_{e}^2. Specifically, σ_g^2 is estimated using the posterior means of the total genetic variance, summed over all QTLs, while σ_{e}^2 is estimated as the total trait variance minus the total genetic variance.

4.1.2 Linkage Analysis Based on Loki

Combined segregation and linkage analysis for quantitative traits was conducted using Loki for each of the data sets (real or simulated). Priors were chosen as described earlier. To gauge the evidence against the null model of no linkage, Bayes Factors (BF) were calculated over 2-cM segments of chromosome 7 from the posterior distribution of indicator of QTL presence. Bayes Factor is the ratio of the posterior odds of two competing models vs. the corresponding prior odds. According to Kass and Raftery [25], the range 1 → 3.2, 3.2 → 10, 10 → 100, >100 provides evidence against the null model that is scant, substantial, strong, and decisive, respectively.

4.1.3 Linkage Analysis Based on MORGAN

LOD scores were estimated at the markers and at approximately 2-cM intervals between markers, using both lm_bayes and lm_markers in MORGAN. Modes of inheritance were obtained from SEGREG and Loki segregation analysis as described earlier for the real data. For the simulated data, the generating (true) model (for one-QTL models) or the marginal model of each QTL (for two-QTL models) were utilized. Both programs require initial configuration of inheritance vectors to start the Markov chain sampling. This is obtained via sequential imputation (SI) method [14] because it has been found to yield “more accurate results” with lm_markers than the alternative independent-locus setup [53]. In addition, lm_bayes conducts an initial MCMC run to estimate the pseudo-prior distribution. The ratio of run length for finding initial configurations via SI for estimating pseudo-prior and for burn-in to that for main analysis were specified following the recommended (default) settings (Table 2).

4.1.4 Linkage Analysis Based on SimWalk2

Similarly, LOD scores were obtained using SimWalk2, using default settings of the software (see footnote of Table 2). The authors of SimWalk [44] do not recommend

Table 2 Run length and computational time of the MCMC pedigree analyses

Method	No. of Iterations for inference (burn-in)	Trait	Time[a] (hh:mm)
Loki, segregation analysis only	50,000 (1,000)	Quant.	2:50[b]
Loki, segregation and linkage analysis	200,000 (1,000)	Quant.	9:44[b]
lm_markers	100,000 (10,000) plus 30,000 SI setup	Quant. Binary	18:09[c] 17:20[d]
lm_bayes	50,000 (5,000) plus 15,000 SI setup plus 25,000 pseudo-prior	Binary	30:31[d]
SimWalk2	Default[e]	Binary	22:25[d]

[a] Total run time on a 2.9-GHz AMD64 processor, running Linux with job management under a Sun Grid Engine

[b] Total CPU time including nine similar Loki segregation scans for optimizing τ_β

[c] Mean of four chromosome 7 scans, run with inheritance models S1, L1, L2, and L3, respectively

[d] Mean of four chromosome 7 scans, run with inheritance models S1-bin, L1-bin, L2-bin, and L3-bin, respectively

[e] The default setting in SimWalk2 is cumbersome to describe in the table format. It involves various specifications, including the run length, the number of simulated annealing iterations, and the subsampling frequency. Details can be found in the cited references and the software

deviating from these settings, and no comparison of the effects of the choices of different settings has been published [53].

4.2 Comparison of the Three Linkage Analysis Software

4.2.1 Framingham Data

Segregation Analysis

SEGREG identified a recessive model with a single QTL for TH, with high-risk allele frequency being 0.27 and the genotype mean for BB being 1.07 (Table 3, model S1). For the binary trait, bTH, the likelihood surface was very flat near the maximum, and hence inheritance parameters could not be estimated using SEGREG. Instead, we used the strategy as described in Sect. 4.1 to derive the corresponding binary trait model, S1-bin, whose detailed specifications are given in the second segment of Table 3.

From Loki analysis on TH, three major QTLs were identified through visual inspection of the genotype effects distribution (Fig. 1). The marginal genetic models for each of these QTLs are given as L1, L2, and L3 in Table 3, together with the penetrances of their corresponding binary trait models, L1-bin, L2-bin, and L3-bin. Each of these models was supplied as the sole segregation model to MORGAN and

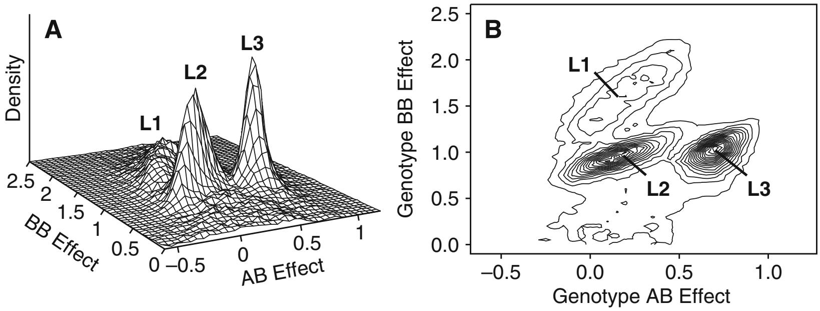


Fig. 1 Posterior density of genotype effects from Loki segregation analysis on TH. The relative posterior density of QTLs fitted into the oligogenic inheritance model (vertical axis, arbitrary scale) throughout the MCMC run is plotted as function of the genotype effects of genotypes AB and BB relative to AA, with B denoting the high risk allele. The peaks marked L1, L2, and L3 correspond to the QTL models in Table 3. (**a**) surface plot; (**b**) corresponding contour plot

Table 3 Estimated inheritance models

	QTL Model[a]			
Parameter	S1	L1	L2	L3
p_A	0.728	0.870	0.580	0.240
μ_{AA}	−0.083	−0.097	−0.232	−0.825
μ_{AB}	−0.122	0.203	−0.092	−0.144
μ_{BB}	1.068	1.643	0.698	0.175
σ_g^2	0.094	0.062	0.108	0.067
σ_e^2	0.341	0.380	0.334	0.375
	Binary Trait Model[b]			
Penetrance[c]	S1-bin	L1-bin	L2-bin	L3-bin
AA	0.205	0.196	0.138	0.021
AB	0.186	0.367	0.198	0.188
BB	0.874	0.984	0.698	0.357

[a] S1, model from SEGREG; L1, L2, L3: models 1, 2 and 3 from Loki segregation analysis
[b] Binary counterparts of the QTL models
[c] Penetrance is the probability of having the disease, given the genotype at the trait locus

SimWalk2 in subsequent linkage analysis, given that these two programs currently only accept single-locus trait models. However, it is worth noting that the three modes of inheritance are interdependent, as they are in fact the effects of the three QTLs in a single segregation model. These three distinct QTLs are visible as peaks in the posterior density plot of genotype effects (Fig. 1).

Linkage Analysis

We conducted linkage analysis on the continuous trait TH, using Loki, on the derived binary trait bTH using lm_bayes and SimWalk2, and on both using lm_markers. Loki, lm_bayes, and SimWalk2 all detected evidence for linkage to

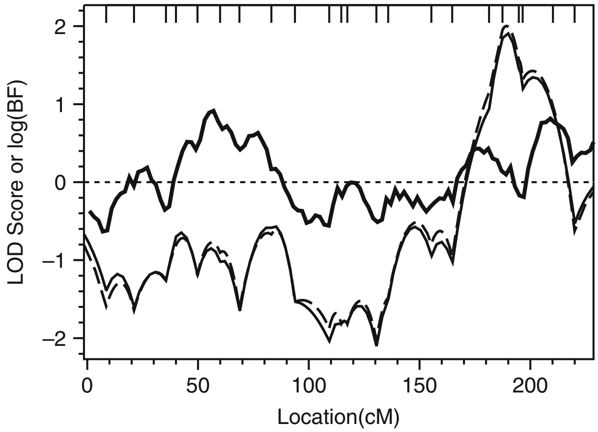


Fig. 2 Chromosome 7 linkage analysis on TH. Multipoint LOD scores from lm_bayes (*thin solid line*) and SimWalk2 (*dashed line*), and $\log_{10}$ BF from Loki (*heavy solid line*) are plotted against map position in Haldane cM. *Tick marks* at the top show the locations of the 22 chromosome 7 markers. A *thin dotted line* indicates the scores expected in the absence of linkage

a region of chromosome 7q (Fig. 2), the former applied to TH (no model specification needed) and the latter two applied to bTH (with each of S1-bin, L1-bin, L2-bin, and L3-bin specified). SimWalk2 gave results nearly identical to lm_markers for binary traits (data not shown), and very similar to lm_bayes, both here (Fig. 2) and in later analyses. For this reason, we present only results from Loki and lm_markers applied to continuous traits, and from SimWalk2 applied to continuous traits. Combined segregation and linkage analysis using Loki yielded substantial evidence for linkage in two regions of chromosome 7: a broad peak on 7p (max. BF = 8.2 at 57 cM) and a slightly weaker signal on 7q (max. BF = 6.5 at 209 cM). The signal on 7q resembles a linkage peak previously reported for the TH trait, in both location and strength [15], and the 7p peak overlaps a signal at 71 cM reported from a variance component s(VC) linkage scan [41], albeit not extensively. Both lm_bayes and SimWalk2, in analysis of the dichotomized bTH trait, found suggestive evidence for linkage on chromosome 7q between markers D7S2195 and D7S1805 (Fig. 2; max. LOD scores = 1.91 at 190 cM and 2.00 at 189 cM, respectively) when L2-bin was specified as the segregation model. This linkage peak overlaps with the strongest VC LOD score for TH [41]. Surprisingly, LOD scores from analyzing the continuous TH trait using lm_markers remained below zero across all but the q-terminal 10 cM of chromosome 7 (data not shown). Not only did this diverge widely from the results obtained from the binary trait, it also indicated that the continuous trait provided less linkage evidence.

Linkage results from SimWalk2 and lm_bayes were, in general, robust to the choice of inheritance model, although strength of linkage signals did vary somewhat (Fig. 3). While model L2-bin yielded the strongest LOD scores (1.9–2.0) in the 7q linkage region, LOD scores between 1.5 and 1.7 were obtained at the same location under models S1-bin and L1-bin. However, under model L3-bin, which

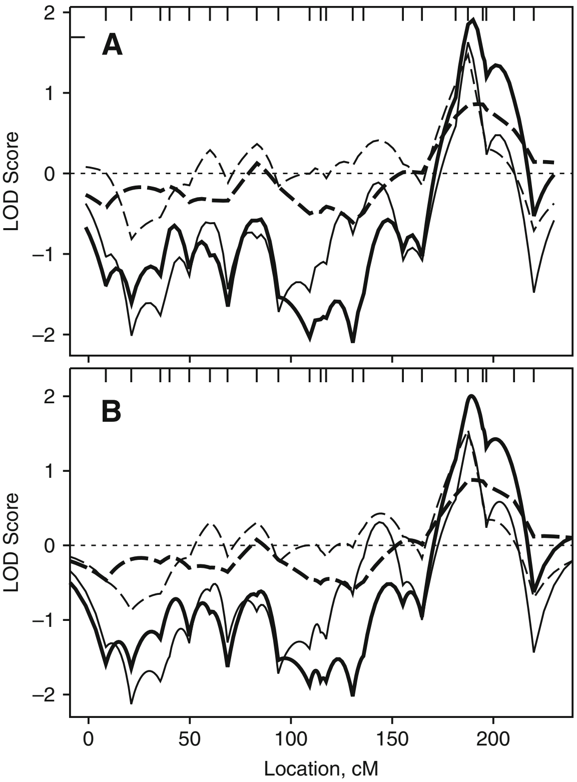


Fig. 3 LOD scores obtained from lm_bayes (**a**) and SimWalk2 (**b**) under four inheritance models. Each plot displays the LOD scores under each of these four models (Table 3): S1 (*thin solid line*), L1 (*thin dashed line*), L2 (*heavy solid line*), and L3 (*heavy dashed line*)

was markedly different from the others in having a highly negative AA genotype mean, the peak LOD score in the 190-cM region was only around 0.8. In analysis of the continuous trait, LOD scores from lm_markers remained near or below zero in the vicinity of the 7q linkage region, regardless of the trait model. However, the lm_markers scan under model L3 found moderate evidence for linkage on chromosome 7p (LOD $=$ 1.33 at 58 cM) near the 7p peak detected by Loki (data not shown).

Of the four MCMC programs, Loki required the least CPU time to complete a full analysis of chromosome 7 (Table 2). lm_markers needed nearly twice as much time, whether given binary or quantitative trait data, to conduct the same analysis; and SimWalk 2 needed more than twice as much as Loki. Slowest of all was lm_bayes, whose analysis under the recommended run length of 100,000 MCMC iterations lasted roughly 60 h. A run of 50,000 iterations is listed in Table 2 because we found that halving the run length did not substantially alter the results.

4.2.2 Simulated Data

As described earlier, we observed intriguing commonalities, as well as stark contrasts, in the linkage analysis results from the four MCMC programs on the Framingham data. To obtain generalizable conclusions, we further carried out similar analysis on simulated data sets. Simulation was based on the same 330 pedigree structures as in the real data. Genotype and continuous trait data were simulated for trait loci (one or two linked QTLs) of known positions and effects, details of which can be found in Sect. 2.2 and Table 1. The continuous trait was dichotomized in the same way as that done with the real data to obtain a binary trait. Genotypes for 22 microsatellite markers, with characteristics similar to those on chromosome 7 in the real data, were then simulated conditional on the genotypes at the trait loci. Because the samples were generated de novo (i.e., not conditioning on existing trait values), the information for linkage varied among replicates. This had the advantage of providing us with linkage signals with a range of strengths, but complicated our efforts to assess power.

Loki and lm_markers were run on the simulated continuous trait, while SimWalk2 and lm_bayes run on the derived binary trait. For data simulated from the one-QTL model, the true QTL model including the correct additive polygenic variance was given to lm_markers, and the corresponding genotype penetrances were given to SimWalk2, for LOD score computation. For data generated from two-QTL models, each of the two marginal QTL models was supplied in turn to two separate analyses carried out using each of the two programs that require mode of inheritance. Consequently, the results from analysis on the simulated replicates represent a best-case scenario. In what follows, comparison of the performance of the three programs will be roughly structured into three categories: (1) Consistency among programs on the overall pattern of LOD score (BF) across the chromosome; (2) Agreement in magnitude of LOD score (BF); and (3) Ability in detecting the true QTL location.

Linkage Analysis on Data Generated from the One-QTL Model

Figure 4 presents results from five independently simulated data sets with one linked QTL at 190 cM. The patterns of rises and falls of the LOD score (Bayes Factor) across the chromosome were similar for all programs, though lm_markers exhibited wilder oscillation, with dramatically negative LOD scores in most unlinked regions. SimWalk2 was least able to locate the true trait locus, failing to detect any linkage in replicates C and D, whereas both Loki and lm_markers provide strong linkage signals in replicate D and some evidence of linkage in replicate C. Similarly, although all three programs show some evidence of linkage around the trait locus in replicate A, SimWalk2 shows a higher linkage peak around 100 cM away from the true locus. This might reflect that the binary trait contains less information than the continuous trait, but other unknown factors may also play a role.

It is difficult to compare linkage evidence based on Bayes Factor and that based on LOD score directly. But overall Loki and lm_markers seem to be able to detect

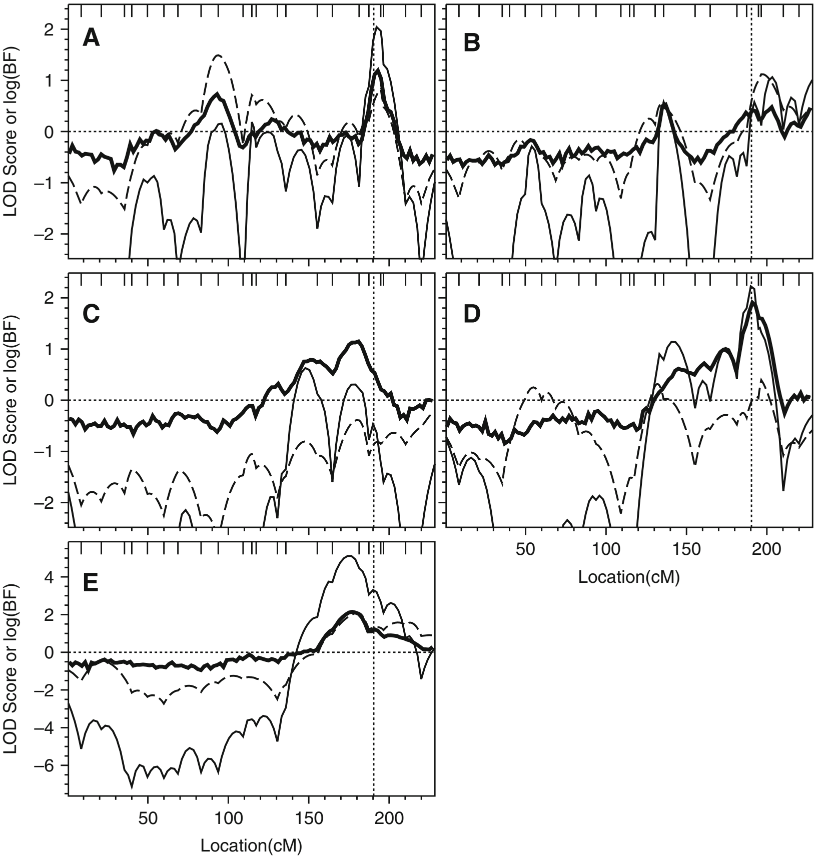


Fig. 4 Linkage analysis on data simulated from the one-QTL model. Each of the five panels (**a**–**e**) represents results from one simulated data set. Plotted are log(BF) values (*heavy solid lines*) from Loki, LOD scores from lm_markers for the quantitative trait (*thin solid lines*), and from SimWalk2 (*dashed lines*). For clarity, the entire range of negative LOD scores from lm_markers is not shown except for in replicate E. The *vertical dotted line* marks the position of the QTL

true linkage with good fidelity with the quantitative trait. Loki was able to localize the QTL within 15 cM of the actual position, with the location at maximum chromosome-wide BF ranging from 177 to 201 cM, regardless of the signal strength. On the other hand, the peak BF varied from 3.4 (Fig. 4, replicate B) to 135.8 (replicate E). lm_markers, despite its much wider range of LOD scores across the chromosome, seems best in detecting true linkage. It reported LOD scores of at least 1.0 near the QTL in all replicates except in C, and found highly significant evidence for linkage in replicate E (LOD = 5.11 at 175 cM). Remarkably, all programs agreed very well in the location of the major linkage signal when substantial evidence for linkage was obtained (replicate E).

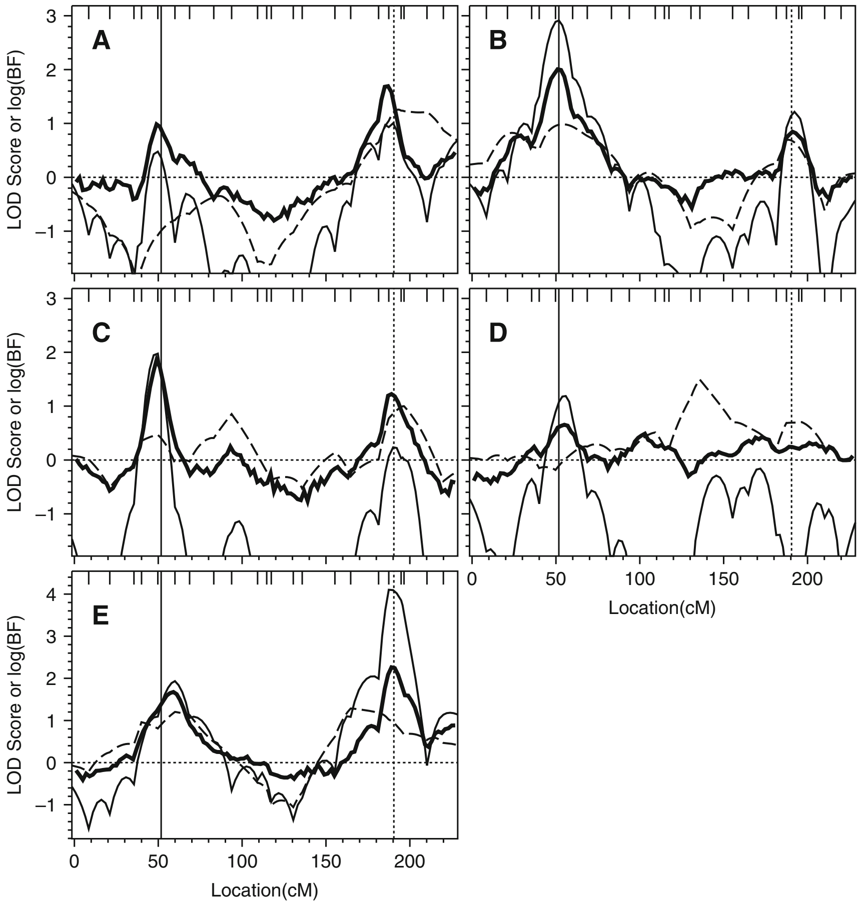


Fig. 5 Linkage analysis on data simulated from the two-QTL model. The marginal model for QTL1 (51.7 cM) was supplied. Each of the five panels (**a**–**e**) represents results from one simulated data set. Plotted are log(BF) values (*heavy solid lines*) from Loki, LOD scores from lm_markers for the quantitative trait (*thin solid lines*), and from SimWalk2 (*dashed lines*). Locations of the QTLs at 51.7 cM and 190.4 cM, are marked with *a vertical solid line and a vertical dotted line, respectively*

Linkage Analysis on Data Generated from the Two-QTL Model

Results from linkage analysis on data simulated from the two-QTL model are presented in Figs. 5 and 6, the former from using the marginal trait model at 51.7 cM and the latter from that at 190.4 cM. The Loki results in Fig. 6 are identical to those in Fig. 5, and are shown for clarity. Although the comparison is more complex, now that there are two regions that house true trait loci, there is remarkable similarity between the conclusions here and those drawn from the one-QTL simulation. Again, SimWalk2 is least able to detect true linkage, with the LOD scores greater

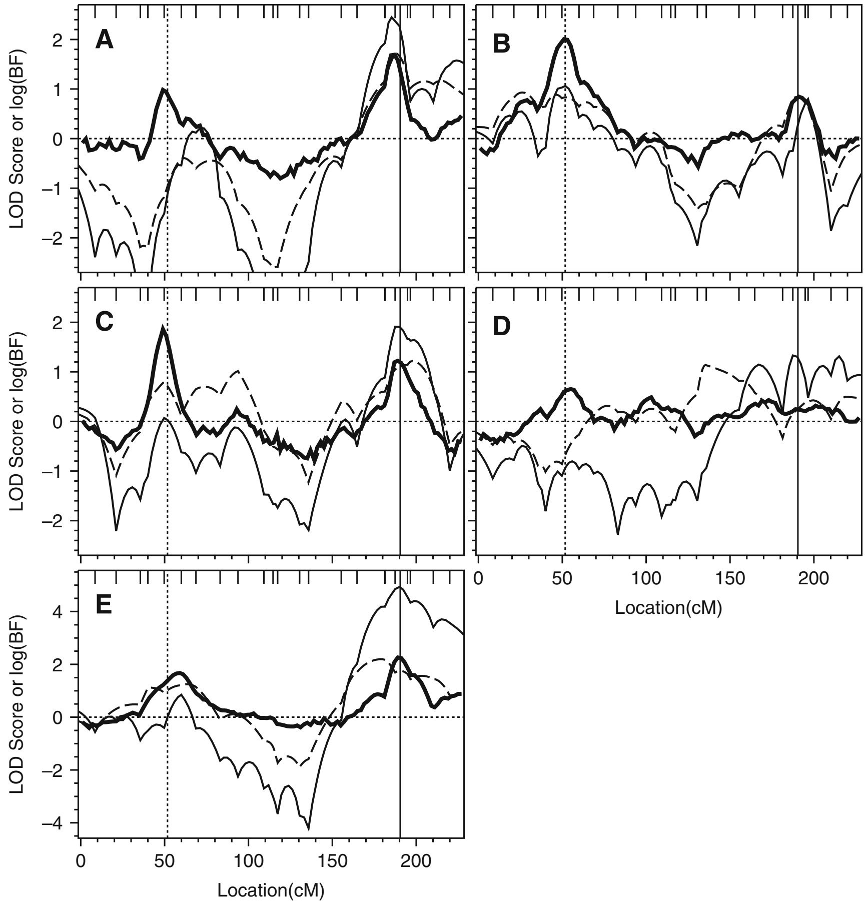


Fig. 6 Linkage analysis on data simulated from the two-QTL model. The marginal model for QTL2 (190.4 cM) was supplied. Each of the five panels (**a**–**e**) represents results from one simulated data set. Plotted are log(BF) values (*heavy solid lines*) from Loki, LOD scores from lm_markers for the quantitative trait (*thin solid lines*), and from SimWalk2 (*dashed lines*). Locations of the QTLs at 51.7 cM and 190.4 cM, are marked with *a vertical dotted line and a vertical solid line, respectively*

near QTL1 than near the QTL2 (replicates A, C, and D) when the QTL1 trait model was supplied, whereas lm_markers shows some evidence of linkage at QTL1 for replicates B and E. Except for the peak near QTL2 in replicates A and E (Fig. 6), LOD scores from SimWalk2 never exceeded 1.5. However, note that because these analyses are marginal ones in the sense that disease loci are searched one at a time, one may identify a "ghost" QTL in between two real QTLs, especially if the effects of the two loci are similar. This phenomenon is seen in the results from replicates C and D (Figs. 5 and 6).

Loki performs well in detecting true linkage. Except for replicate D (where lm_markers also has difficulty), Loki BF profiles from all replicates were dominated by two prominent peaks corresponding to the two QTLs. Overall, addition of the second QTL to the simulation model strengthened the linkage signals obtained from Loki as compared to the single-QTL simulated replicates, due to the greater proportion of major-gene variance relative to polygenic and residual variance in the two-QTL samples (Table 1). Nonetheless, Loki was not successful, by and large, in estimating inheritance parameters accurately in the presence of two QTLs (data not shown).

Lastly, lm_markers, applied to the quantitative trait, also performed well in detecting true linkage. It was able to find some evidence for linkage (peak LOD > 1) from most replicates at the generating trait locus. It was able to provide an exceedingly strong linkage signal near QTL2 in replicate E, under both marginal models, whereas SimWalk2 only provided weak evidence of linkage at locations intermediate to the two true trait loci. The increase of LOD score from lm_markers relative to that from SimWalk2 at the true trait loci may be due, in part, to the greater information available from the original quantitative trait relative to the dichotomized trait, and the ability of lm_markers to accommodate polygenic variance in the inheritance model. Moreover, only lm_markers detected either locus in replicate D, although Loki provided weak evidence for linkage near QTL1.

5 Conclusions, Recommendations, and Other Considerations

We have compared four MCMC-based programs from three software packages for model-based multipoint linkage scans of genetic data on complex pedigrees: Loki, MORGAN (lm_bayes, lm_markers), and SimWalk2. Comparisons were done in terms of computational burden, power, and sensitivity to model specification.

The total computational burden was the smallest for Loki. It should be noted, however, that Loki is highly output-intensive during program execution. Running Loki on a shared cluster can cause system-wide slowdowns, as Loki frequently writes output to the user's account over a network. Scans reported here were performed on a shared cluster, but during low-use periods. The LOD-score programs do not share this drawback.

Overall, lm_bayes and SimWalk2 produced very similar results. In both the real and simulated data examples, SimWalk2 required less CPU time to run than lm_bayes, even when the MCMC run for the latter was halved compared to the recommended length. However, it is important to note that there is no built-in convergence diagnostic tools within either of the program, and as such the recommended run length may not be appropriate, as the required number of MCMC iterations to reach convergence may vary among programs with the particulars of the data and of the specified inheritance model. SimWalk2 and lm_markers run on dichotomized traits returned nearly identical results, but lm_markers completed its analysis in about three-fourths the time required by SimWalk2. The similarity of the results from SimWalk2, lm_markers, and lm_bayes, when run on the same binary trait, are not surprising; they share either the sampling method (lm_markers and lm_bayes) or the inference method (SimWalk2 and lm_markers). However, differences can, and are expected to, occur when the sampling space is more difficult to explore thoroughly (due to single-locus reducibility or near-reducibility) and/or when one inference model is more appropriate to capture information contained in the data.

While lm_bayes, lm_markers, and SimWalk2 are all methods that can handle single-gene models (marginal analysis), Loki employs a joint search strategy for the potential existence of multiple (interacting) loci. This feature distinguishes Loki from the rest of the methods compared, as it is more powerful when the underlying genetic model does have multiple loci linked and/or interacting epistatically. This is indeed a very important feature as it is known that a "ghost" QTL between two linked true QTLs may result if a marginal analysis strategy (using single-locus models) is employed. However, Loki has its own limitations, with the greatest being that the significance of the reported Bayes Factors is poorly understood. A simulation-based approach for estimating significance levels for Loki linkage peaks has been developed [23]. The simulation studies conducted there suggested that the logarithm of the Bayes Factor has a roughly linear relationship with the LOD score from variance-components linkage analysis under an additive model. However, the procedure is highly computationally demanding and has not been tested extensively.

By the same token, how to interpret evidence from LOD score summaries is just as important. Although there are many more discussions in the literature on LOD scores, there does not exist a single consensus, partly due to many complicating issues, including multiple testing and heterogeneity. As the goal of this chapter is on comparing and contrasting several MCMC approaches and programs, we did not make any attempt to set a threshold for linkage declaration. Instead, we reported on the general behavior of the LOD score (and BF) curves, and compare the maximum (positive) LOD scores and locations where the maximum occur. In doing so, we hope to guide the reader to a better understanding of the process of interpreting the MCMC outputs.

The availability of convergence diagnostics is another important issue. Loki's MCMC output is more transparent to diagnostic examination of MCMC mixing. Various diagnostic plots may be constructed from the raw MCMC output [52]. MCMC convergence in lm_bayes was assessed in an earlier study by determining run length required to stabilize LOD scores and by examining the sensitivity of LOD scores to different starting conditions [16]. Here, we did examine consistency of the results under a range of running length, but did not thoroughly test sensitivity to initial conditions (data not shown).

From our analyses, it appears that lm_markers is very powerful under the right circumstances: specifically, when continuous trait is analyzed and when the mode of inheritance is correctly specified. lm_markers allows an additive polygenic variance component to be specified in the inheritance model. When such a variance is present, including the polygenic component has the potential of greatly enhancing power through accounting for some of the variability ("noise") beyond that caused by major genes. However, in complex traits, polygenic variance may be difficult to estimate in a segregation analysis, and attributing genetic variance from major genes to a polygenic variance component may cost power due to model misspecification. Note that neither SEGREG nor Loki provides an estimate of the polygenic variance, which is being absorbed by the residual variance. The lower power of the other two LOD-score methods, and their improved robustness to model misspecification, may be partly ascribed to the differences in analyzing a continuous phenotype and a dichotomized version of the same trait.

We started out with quantitative traits for all our data (real or simulated) so that we can bring Loki into the mix as it is only applicable to continuous data. Although lm_markers can analyze both quantitative and binary traits, lm_bayes and SimWalk2 are only amenable to binary ones, and thus the continuous trait data were dichotomized. As we have seen in the outcomes and discussed in the previous paragraph, this process may result in unfair advantage for Loki and lm_markers (when it is applied to the original quantitative trait directly), and thus caution needs to be exercised in interpreting the relative powers of the various programs. In fact, loss of power in analyzing a dichotomized trait derived from a continuous trait is well documented in the literature, and is not a unique phenomenon with MCMC approaches. However, it is also worth pointing out that exceptions do exist, especially when the underlying quantitative trait model may be incorrectly specified, as we saw in the analysis of the real data and as discussed earlier.

Which MCMC program to use is problem- and data-type-dependent, but some basic analysis strategies are still advisable. If the original trait is quantitative, we recommend analyzing it as a quantitative trait using Loki and/or lm_markers, rather than dichotomizing it, to retain greater power. However, if the original data are binary, then any one of the three, lm_markers, lm_bayes, or SimWalk2, can be used, which would likely give similar results for many types of data. And as such, the choice may boil down to ease of program usage and computational intensity. For an analysis using any of these three programs, a disease model needs to be specified. If the trait is quantitative, then Loki segregation analysis may be performed to identify the marginal models. As currently implemented, Loki does not perform any ascertainment correction. Nonetheless, segregation analysis using Loki is still valid for population-based samples such as the Framingham pedigrees. For a binary trait, programs such as SEGREG are needed to obtain an approximation model. As any model obtained from a segregation analysis is only an approximation, sensitivity analysis of the results to changes in the model is advisable. Also, as lm_markers, lm_bayes, and SimWalk2 are currently only applicable to single-locus models, if marginal analysis leads to interesting regions, then an analysis that allows for multiple disease loci clearly needs to be carried out. Finally, it is important to note that if the data can be analyzed with an exact calculation program, then MCMC should not be attempted.

Some comparisons of these MCMC methods have been carried out previously. Loki and lm_bayes were compared in GAW13 [15] using traits similar to TH. While this previous study found stronger evidence from Loki than we did, our inheritance model L2-bin yielded greater maximum LOD scores from lm_bayes. These contrasts could be explained by difference ways in adjusting the phenotype for covariates, and by difference in selection of parameters for the Loki analysis (for instance, τ_β).

Loki and lm_markers reported similar results in a chromosomal region with strong suggestive evidence for linkage to a real-word reading phenotype associated with dyslexia [21].

As an aside, we also attempted an exact LOD-score analysis of the real, bTH, data using SUPERLINK, but the program was unable to complete the calculations when more than two markers were included. When genotypes were supplied for the two markers flanking the LOD-score peak on chromosome 7q, D7S2195 (187.4 cM) and D7S1805 (194.8 cM), SUPERLINK reported a LOD score of 1.48 at D7S2195 (data not shown), slightly lower than the peak LOD scores from lm_bayes and SimWalk2.

Genotyping for genomewide linkage scans has shifted toward panels of single-nucleotide polymorphisms (SNPs), including the Illumina IVb and the Affymetrix Mapping panels, as inexpensive, high-throughput techniques for SNP genotyping have become available. Relatively much less is known as to how these dense marker maps will affect MCMC-based linkage analysis. In the case of Loki, one report has suggested that dense arrays of SNP markers may cause difficulties in MCMC mixing, and in consequence, inconsistent results across analyses [42]. In a comparison between lm_markers and SimWalk2, lm_markers was considerably more efficient and provided more accurate LOD scores, while these differences were largely absent with microsatellite markers [53].

Throughout this chapter, we have mainly focused on comparing MCMC methods for model-based linkage analysis, and the programs selected are those that have been most tested and deemed most reliable. However, this focus is extremely narrow in light of a wealth of other Monte Carlo approaches and programs in linkage analysis. Also, as alluded to earlier, Loki, lm_bayes, and lm_markers all require that the pedigrees be single-locus peelable, which limits their applicability to extremely complex pedigrees. In such cases, even finding a starting point for the Markov chain can be extremely challenging and special algorithms are needed for such a task [34].

Other than Loki, all the other programs tested are applicable only to models that specify the mode of inheritance of a single major gene locus. However, for complex traits, there are usually multiple genes in action, and an explicit multilocus analysis (either a joint or a conditional approach) would be more powerful. Readers who are interested in various MCMC approaches for two-locus analysis are referred to discussions elsewhere [4, 31, 33, 46].

In the same vein, correct specification, or even just good approximation, of the model of a complex trait is usually extremely difficult, and as such, IBD-based methods may be preferred. Monte Carlo methods for IBD-based linkage analysis can be found in the literature (e.g., [43, 49]). Another complication for analyzing complex traits is locus heterogeneity, which is one of the causes for poor replicability of positive linkage results, and has been treated in various ways, including Bayesian approaches (e.g., [3]). For the four programs compared, only SimWalk2 has a provision for detecting and accounting for heterogeneity. Finally, it is worth noting that the above by no means provides an exhaustive list of Monte Carlo approaches to linkage analysis. There are also many Monte Carlo approaches for tackling other topics in genetic mapping related problems, including, but not limited to, population structure inference [37] and association mapping [6, 10].

6 Web Resources

S.A.G.E. http://darwin.cwru.edu/sage
MORGAN and Loki.
http://www.stat.washington.edu/thompson/Genepi/Pangaea
SimWalk2. http://www.genetics.ucla.edu/software/Simwalk

Acknowledgements We thank Joseph Rothstein, Yun Ju Sung, and Ellen Wijsman for their valuable advice. This work was supported in part by NSF grant DMS-0112050 and NIH grants R01-HG002657 and RO1-HG003054. GAW is supported by NIH grant R01 GM031575. Some of the results of this paper were obtained by using the program package S.A.G.E., which is supported by a U.S. Public Health Service Resource Grant (RR03655) from the National Center for Research Resources. We also acknowledge the Framingham Heart Study for their permission to use the GAW 13 data, and we thank the Framingham Heart Study Investigators for their contributions and the NHLBI for collection of the Framingham Study data.

References

1. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101
2. Atwood LD, Heard-Costa NL (2003) Limits of fine-mapping a quantitative trait. Genet Epidemiol 24:99–106
3. Biswas S, Lin S (2006) A Bayesian approach for incorporating variable rates of heterogeneity in linkage analysis. J Am Stat Assoc 101:1341–1351
4. Biswas S, Papachristou C, Irwin MC, Lin S (2003) Linkage analysis of the simulated data – evaluations and comparisons of methods. BMC Genet 31 (Suppl 1):S70
5. Bonney G (1986) Regressive logistic models for familial disease and other binary traits. Biometrics 42:611–625
6. Chen W-M, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81:913–926
7. Cottingham R, Idury RM, Schffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252–263
8. Cupples LA, Yang Q, Demissie S, Copenhafer D, Levy D, FraminghamHeartStudyInvestigators (2003) Desription of the Framingham Heart Study data for Genetic Analysis Workshop 13. BMC Genet 4(Suppl. 1):S2
9. Dietter J, Spiegel A, an Mey D, Pflug H-J, Al-Kateb H, Hoffmann K, Wienker TF, Strauch K (2004) Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia. Eur J Hum. Genet 12:542–550
10. Ding J, Lin S, Liu Y (2006) Monte Carlo pedigree disequilibrium test for markers on the X chromosome. Am J Hum Genet 79:567–573
11. Elston RC, Stewart J (1971) A general model for the analysis of pedigree data. Hum Hered 21:523–542
12. Fishelson M, Geiger D (2002) Exact genetic linkage computations for general pedigrees. Bioinformatics 18:S189–S198
13. George AW, Thompson EA (2002) Multipoint linkage analyses for disease mapping in extended pedigrees: a Markov chain Monte Carlo approach. Technical report no. 405, Department of Statistics, University of Washington, Seattle, WA
14. George AW, Thompson EA (2003) Discovering disease genes: multipoint linkage analysis via a new Markov chain Monte Carlo approach. Stat Sci 18:515–531

15. George AW, Basu S, Li N, Rothstein JH, Sieberts SK, Stewart W, Wijsman EM, Thompson EA (2003) Approaches to mapping genetically correlated complex traits. BMC Genet 4 (Suppl 1):S71
16. George AW, Wijsman EM, Thompson EA (2005) MCMC multilocus lod scores: application of a new approach. Hum Hered 59:98–108
17. Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732
18. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, Gordon DJ, Krauss RM, Savage PJ, Smith SC, Spertus JA, Costa F (2005) Diagnosis and management of the metabolic syndrome. Circulation 112:2735–2752
19. Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760
20. Horne BD, Malhotra A, Camp NJ (2003) Comparison of linkage analysis methods for genome-wide scanning of extended pedigrees, with application to the TG/HDL-C ratio in the Framingham Heart Study. BMC Genet 4(Suppl. 1):S93
21. Igo RP Jr, Chapman NH, Berninger VW, Matsushita M, Brkanac Z, Rothstein JH, Holzman T, Neilsen K, Raskind WH, Wijsman EM (2006) Genomewide scan for real-word reading subphenotypes of dyslexia: novel chromosome 13 locus and genetic complexity. Am J Med Genet (Neuropsychiatr Genet) 141B:15–27
22. Igo RP, Jr, Chapman NH, Wijsman EM (2006) Segregation analysis of a complex quantitative trait: approaches for identifying influential data points. Hum Hered 61:80–86
23. Igo RP, Jr, Wijsman EM (2008) Empirical significance values for linkage analysis: trait simulation using posterior model distributions from MCMC oligogenic segregation analysis. Genet Epidemiol 32:119–131
24. Irwin M, Cox N, Kong A (1994) Sequential imputation for multipoint linkage analysis. Proc Natl Acad Sci USA 91:11684–11688
25. Kass RE, Rafferty AE (1995) Bayes factors. J Am Stat Assoc 90:773–795
26. Kong A, Liu JS, Wong WH (1994) Sequential imputations and Bayesian missing data problems. J Am Stat Assoc 89:278–288
27. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
28. Lander E, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367
29. Lange K, Sobel E (1991) A random walk method for computing genetic location scores. Am J Hum Genet 49:1320–1334
30. Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446
31. Lin S (2000) Monte Carlo methods for linkage analysis of two-locus disease models. Ann Hum Genet 64:519–532
32. Lin S, Skrivanek Z, Irwin M (2003) Haplotyping using SIMPLE: caution on ignoring interference. Genet Epidemiol 25:384–387
33. Luo Y, Lin S, Irwin ME (2001) Two-locus modeling of asthma in a Hutterite pedigree via Markov chain Monte Carlo. Genet Epidemiol 21(Suppl 1):S24–S29
34. Luo Y, Lin S (2003) Finding starting points for Markov chain Monte Carlo analysis of genetic data from large and complex pedigrees. Genet Epidemiol 25:14–24
35. MacCluer JW, Vandeberg JL, Read B, Ryder OA (1986) Pedigree analysis by computer simulation. Zoo Biol 5:147–160
36. O'Connell JR (2001) Rapid multipoint linkage analysis via inheritance vectors in the Elston–Stewart algorithm. Hum Hered 51:226–240
37. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
38. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
39. Robert CP, Casella G (2004). Monte Carlo statistical methods. Springer-Verlag, New York

40. S.A.G.E. (2007) Statistical analysis for genetic epidemiology, version 5.4. http://darwin.cwru.edu/sage/
41. Shearman AM, Ordovas JM, Cupples LA, Schaefer EJ, Harmon MD, Shao Y, Keen JD, DeStefano AL, Joost O, Wilson PWF, Housman DE, Myers RH (2000) Evidence for a gene influencing the TG/HDL-C ratio on chromosome 7q32.3-qter: a genome-wide scan in the Framingham Study. Hum Mol Genet 9:1315–1320
42. Sieh W, Basu S, Fu AQ, Rothstein JH, Scheet PA, Sterward WCL, Sung YJ, Thompson EA, Wijsman EM (2005) Comparison of marker types and map assumptions using Markov chain Monte Carlo-based linkage analysis of COGA data. BMC Genet 6(Suppl 1):S11
43. Skrivanek Z, Lin S, Irwin M (2003) Linkage analysis with sequential imputation. Genet Epidemiol 25:25–35
44. Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323–1337
45. Sobel E, Sengul H, Weeks DE (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. Hum Hered 52:121–131
46. Sung YJ, Thompson EA, Wijsman EM (2007) MCMC-based linkage analysis for complex traits on general pedigrees: multipoint analysis with a two-locus model and polygenic component. Genet Epidemiol 31:103–114
47. Thompson EA (1995) Monte Carlo in genetic analysis. Technical report no. 294, Department of Statistics, University of Washington, Seattle, WA
48. Thompson EA (2000) Statistical inferences from genetic data on pedigrees, vol. 6. IMS, Beachwood, OH
49. Thompson EA (2005) MCMC in the analysis of genetic data on pedigrees. In: Liang F, Wang J-S, Kendall W (eds) Markov Chain Monte Carlo: innovations and applications. World Scientific, Singapore
50. Wijsman EM, Amos CI (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol 14:719–735
51. Thompson EA, Heath SC (1999) Estimation of conditional multilocus gene identity among relatives. In: Seillier-Moiseiwitsch F (ed) Statistics in molecular biology and genetics: selected proceedings of the 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology. Institute of Mathematical Studies, Hayward, CA
52. Wijsman EM, Yu D (2004) Joint oligogenic segregation and linkage analysis using Bayesian Markov chain Monte Carlo methods. Mol Biotechnol 28:205–226
53. Wijsman EM, Rothstein J, Thompson EA (2006) Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. Am J Hum Genet 79:846–858

Population-Based Association Studies

Xiaofeng Zhu and ShuangLin Zhang

Abstract Population-based association studies have been playing a major role in mapping genes affected complex diseases. The advantages of population based association studies include greater efficiency in sample recruitment and more power than family-based studies. However, population-based association mapping may lead to false positive findings if population stratification is not properly considered. In this chapter, we will review population-based association mapping methods that can control false positive rate due to population stratification. These methods include logistic regression, genomic control, structure association, and semi-parametric approaches. We will apply these methods to a simulated data set and illustrate the advantages and limitations of these methods.

1 Introduction

Population-based association studies have been considered more powerful than family-based linkage studies in the genetic dissection of complex diseases [1, 2]. Population association between genotype at a locus and disease can arise in three ways: (1) the genotype at the locus directly causes the disease; (2) the locus itself is not causal, but is in linkage disequilibrium with a causal locus; (3) population stratification or admixture [3]. In genetic epidemiology, we are interested in the association aroused in the first two cases, but try to avoid the third association, which has little scientific interest. When the studied samples come from an admixed population and cases and controls have different ancestry distribution, the third association can be created between a genetic marker and the disease [4, 5]. For example, a well-known study of type II diabetes mellitus and Gm3;5,13,14 in American Indians suggested the association when analysis was performed in whole sample, but no

X. Zhu (✉)
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA
e-mail: xzhu1@darwin.case.edu

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_5, © Springer-Verlag Berlin Heidelberg 2010

association when analysis was restricted to subjects of full Indian heritage, suggesting the confounding of the disparities of the European ancestry in diabetes group and controls [4]. Even in a relatively homogenous population such as European Americans, subtle population stratification can still result false positive findings, as demonstrated by the association of human adult height and a SNP in lactose gene that is due to the admixture of southern and Northern Europeans [6]. The effect of population stratification becomes even more problematic in whole genome association studies in which a large sample size is usually required to have enough power [7]. To overcome such a problem, alternative statistical methods have been suggested, including family-based designs that utilize family members as controls [8,9]. Although it is immune to population stratification, a disadvantage of family-based designs is the difficulty to recruit enough desired families, therefore limits the statistical power. The case–control design is widely used in epidemiology studies and recently this design is also favored in studying the association between disease and a locus. Statistical methods such as comparing the allele frequencies between cases and controls or logistic regression can be applied. However, these methods will also detect the third kind of association arose by population stratification. Thus, methods using a set of unlinked genetic markers typed in the same samples to control for the effect of the population stratification have been proposed. There are three kinds of approaches: (1) The "genomic control" (GC) method [10]; (2) Structure Association [11–13]; (3) Principal component approaches [14–17]. In this chapter, we will use a simulated data to illustrate the three methods.

2 The Data

We will use a simulated data set throughout the chapter. To provide a reasonable framework for the simulations, we accessed a panel of SNPs that are ancestry informative for African American population across the genome reported by Smith et al. [18]. The allele frequencies of the SNPs and the marker map for both the African and European populations were downloaded from www.journals.uchicago.edu. Briefly, at the first generation the marker genotypes of 10,000 unrelated people were simulated according to the marker allele frequencies in African population assuming HWE and independence of the markers. An admixed population was then formed by taking a proportion λ randomly selected from African population to marry with people generated according to the marker allele frequencies in European population, with the remaining proportion $1 - \lambda$ randomly mating among them. We let λ be drawn from a uniform distribution between 0 and 0.08. The number of children produced by each marriage was assumed to follow a Poisson distribution with mean size 2. The number of crossovers between two marker loci at a distance d cM was assumed to follow a Poisson distribution with mean d/100. This process was repeated in the following generations. All the samples were drawn from the fifth generation. To simulate which individuals are affected, we selected a disease marker (SNP1) located at the middle of chromosome 1. The penetrances of

Table 1 Distribution of the simulated two SNPs in cases and controls. SNP1 is the disease locus and SNP2 is associated with disease because of population stratification

SNP1									
Genotype	Cases	Controls	Odd ratio	p-value	Allele	Cases	Controls	Odd ratio	P-value
d/d	148	221			d	542	657		
d/D	246	215	1.71	0.00015	D	458	343	1.62	$1.53{\times}10^{-7}$
D/D	106	64	2.47	1.53×10^{-6}					
Total	500	500	–			1,000	1,000		
SNP2									
0/0	32	47			0	259	312		
0/1	195	218	1.31	0.273	1	741	688	1.30	0.0087
1/1	273	135	1.71	0.029					
Total	500	500	–			1,000	1,000		

the disease genotypes were 0.2, 0.15, and 0.1 for carrying genotype DD, Dd, and dd, respectively. For the association analysis, we also examined an SNP located on a different chromosome from SNP1. We selected 500 cases and controls, respectively. Table 1 presented the genotype and allele frequencies for these two SNPs. In addition, we also simulated 1,000 SNPs using for control population stratification.

2.1 Association of a Genetic Marker and a Disease

Considering a widely used case–control design, assume there are two alleles at the disease locus with risk allele D and normal allele d, and allele frequencies P_D and P_d. Denote $\pi_{DD}^{(1)}$, $\pi_{Dd}^{(1)}$, and $\pi_{dd}^{(1)}$ be the frequencies of genotype DD, Dd, and dd in cases, respectively. Similarly, the frequencies of the genotype in controls are $\pi_{DD}^{(0)}$, $\pi_{Dd}^{(0)}$, and $\pi_{dd}^{(0)}$. The genotype frequencies in cases and controls are summarized in Table 2. Let f_0, f_1, and f_2 denote the penetrances of genotypes dd, Dd, and DD. That is $f_0 = \Pr(\textit{Disease}|dd)$, $f_1 = \Pr(\textit{Disease}|Dd)$, and $f_2 = \Pr(\textit{Disease}|DD)$ with $f_2 \geq f_1 \geq f_0$. The association between the disease and the genotype can be measured in terms of odds ratios [19], which are

$$\theta_{Dd} = \frac{\pi_{Dd}^{(1)}\pi_{dd}^{(0)}}{\pi_{Dd}^{(0)}\pi_{dd}^{(1)}} \text{ and } \theta_{DD} = \frac{\pi_{DD}^{(1)}\pi_{dd}^{(0)}}{\pi_{DD}^{(0)}\pi_{dd}^{(1)}}.$$

These odds ratios are measured relative to genotype dd. When $\theta_{Dd} = \theta_{DD} = 1$, there is no association between the locus and disease. We should notice that each individual carries two alleles. When a population is in Hardy–Weinberg equilibrium (HWE), that is, an individual's two alleles are independent, we can also measure the association through the alleles, as illustrated in Table 3. Allelic association can be estimated by the odd ratio: $\theta_D = \frac{P_D^{(1)}P_d^{(0)}}{P_D^{(0)}P_d^{(1)}}$. As an example of the data simulated

Table 2 Contingence genotype of a case-control study

Genotype	Cases		Controls		Odds ratio
	Freq	Count	Freq	Count	
dd	$\pi_{dd}^{(1)}$	$n_{dd}^{(1)}$	$\pi_{dd}^{(0)}$	$n_{dd}^{(0)}$	
Dd	$\pi_{Dd}^{(1)}$	$n_{Dd}^{(1)}$	$\pi_{Dd}^{(0)}$	$n_{Dd}^{(0)}$	θ_{Dd}
DD	$\pi_{DD}^{(1)}$	$n_{DD}^{(1)}$	$\pi_{DD}^{(0)}$	$n_{DD}^{(0)}$	θ_{DD}
Total	1	$n^{(1)}$	1	$n^{(0)}$	

Table 3 Contingence table of allele frequency table of a case–control study

Allele	Cases		Control		Odds ratio
	Freq	Count	Freq	Count	
d	$P_d^{(1)}$	$2n_{dd}^{(1)} + n_{Dd}^{(1)}$	$P_d^{(0)}$	$2n_{dd}^{(0)} + n_{Dd}^{(0)}$	
D	$P_D^{(1)}$	$2n_{DD}^{(1)} + n_{Dd}^{(1)}$	$P_D^{(0)}$	$2n_{DD}^{(0)} + n_{Dd}^{(0)}$	θ_D
Total	1	$2n^{(1)}$	1	$2n^{(0)}$	

in Table 1 for SNP1, we estimate the odds ratios $\theta_{Dd} = 1.71$, $\theta_{DD} = 2.47$, and $\theta_D = 1.62$, respectively.

When a marker locus M with alleles A and a locates closely to the disease locus, the association between locus M and the disease locus can be aroused through linkage disequilibrium (LD). LD refers to the nonrandom association of alleles between nearby loci and is the basis of association mapping for complex traits [20–22]. Denoting the allele frequencies of A and a in the population as P_A and P_a. Denoting the frequencies of four haplotypes DA, Da, dA, and da as P_{DA}, P_{Da}, P_{dA}, and P_{da}, respectively. One of the commonly used measures of LD is Δ [22]:

$$\Delta = P_{DA} - P_D P_A.$$

For the marker locus M, we can lay a similar Tables 2 and 3 through replacing D and d by A and a, respectively. Similarly, we denote θ_{Aa} and θ_{AA} the odds ratios relative to genotype *aa*, and θ_A the odds ratio of A to a. Association may also be present when the disease and marker loci are in different chromosomes because of population admixture. For example, the simulated SNP2 in Table 1 has $\theta_{Aa} = 1.31$, $\theta_{AA} = 1.71$, and $\theta_A = 1.30$ although it is not on the chromosome SNP1 located. When $\theta_{Aa} = \theta_{AA} = 1$ or $\theta_A = 1$, we say there is no association between locus M and the disease. The odds ratios of marker locus M can be expressed as the function of the odds ratios at disease locus and linkage disequilibrium coefficient Δ. To illustrate this, we have

$$\theta_A - 1 = \frac{P_A^{(1)} P_a^{(0)}}{P_A^{(0)} P_a^{(1)}} - 1 = \frac{P_A^{(1)} - P_A^{(0)}}{P_A^{(0)} P_a^{(1)}}$$

$$= \frac{\Pr(DA|disease) + \Pr(dA|disease) - \Pr(DA|normal) - \Pr(dA|normal)}{P_A^{(0)} P_a^{(1)}}$$

$$\times \frac{1}{P_A^{(0)} P_a^{(1)}} \left[\frac{\Pr(disease|D)\,P(DA) + \Pr(disease|d)\,P(dA)}{P(disease)} \right.$$

$$\left. - \frac{\Pr(normal|D)\,P(DA) + \Pr(normal|d)\,P(dA)}{P(normal)} \right].$$

By using $P(DA) = \Delta + P(D)\,P(A)$ and $P(dA) = -\Delta + P(d)\,P(A)$, we have

$$\begin{aligned}
\theta_A - 1 &= \frac{\Delta}{P_A^{(0)} P_a^{(1)}} \left[\frac{\Pr(D|disease) - \Pr(D|normal)}{P_D} \right. \\
&\qquad \left. - \frac{\Pr(d|disease) - \Pr(d|normal)}{P_d} \right] \\
&= \frac{\Delta}{P_A^{(0)} P_a^{(1)}} \left[\frac{P_D^{(1)} - P_D^{(0)}}{P_D} - \frac{P_d^{(1)} - P_d^{(0)}}{P_d} \right] \\
&= \frac{\Delta \left(P_D^{(1)} - P_D^{(0)} \right)}{P_A^{(0)} P_a^{(1)}} \left[\frac{1}{P_D} + \frac{1}{P_d} \right] \\
&= \Delta(\theta_D - 1) \frac{P_D^{(0)} P_d^{(1)}}{P_A^{(0)} P_a^{(1)} P_D P_d} \\
&= (\theta_D - 1) \frac{\Delta}{\sqrt{P_A^{(0)} P_a^{(1)} P_D P_d}} \frac{P_D^{(0)} P_d^{(1)}}{\sqrt{P_A^{(0)} P_a^{(1)} P_D P_d}} \qquad (1)
\end{aligned}$$

It implies that $\theta_A = 1$ is equivalent to $\Delta = 0$ because of $\theta_D \neq 1$. It also suggests that testing association of contingence Tables 2 or 3 between a marker locus and disease locus is equivalent to testing for the linkage disequilibrium between a marker locus and the disease locus. For a common disease, the middle part of (1) approximates to the correlation between a marker and disease locus.

2.2 Testing for Association When No Population Stratification Is Present

Section 2.1 suggests that we can test the association between a marker and disease using the contingence Tables 2 and 3. However, for the association test to be valid for Table 3, HWE at marker locus is also required [23]. In general, we can view the columns of cases and controls in Tables 2 or 3 as two multinomial distributions. We can then test the association by testing for the equivalence of the two distributions under the null hypothesis. The log-likelihood of Table 2 for the observed genotype counts is

$$L = \sum_{i=1}^{3} n_i^{(1)} \log \pi_i^{(1)} + \sum_{i=1}^{3} n_i^{(0)} \log \pi_i^{(0)},$$

where $\pi_1^{(1)}, \pi_2^{(1)}, \pi_3^{(1)} \left(\pi_1^{(0)}, \pi_2^{(0)}, \pi_3^{(0)}\right)$ and $n_1^{(1)}, n_2^{(1)}, n_3^{(1)} \left(n_1^{(0)}, n_2^{(0)}, n_3^{(0)}\right)$ denote the frequencies and counts of the three genotypes in cases (controls). The null hypothesis of no association is $\pi_i^{(1)} = \pi_i^{(0)}$ for all i, which is equivalent to $\theta_{Aa} = \theta_{AA} = 1$. Either maximum likelihood ratio test or score test (Pearson Chi-square test) can be applied for testing the null hypothesis [19]. The corresponding two test statistics for the null hypothesis are

$$\mathrm{LRT} = 2\sum_i n_i^{(1)} \log\left(n_i^{(1)}/\hat{n}_i^{(1)}\right) + 2\sum_i n_i{}^{(0)} \log\left(n_i^{(0)} \,/\, n_i^{(0)}\right), \tag{2}$$

$$X_2^2 = \sum_i \left[\frac{\left(n_i^{(1)} - \hat{n}_i^{(1)}\right)^2}{\hat{n}_i^{(1)}} + \frac{\left(n_i^{(0)} - \hat{n}_i^{(0)}\right)^2}{\hat{n}_i^{(0)}}\right], \tag{3}$$

where $\hat{n}_i^{(1)}$ and $\hat{n}_i^{(0)}$ are the expected frequencies under the null hypothesis and can be calculated by $\hat{n}_i^{(1)} = \frac{n^{(1)}\left(n_i^{(1)}+n_i^{(0)}\right)}{n^{(1)}+n^{(0)}}$ and $\hat{n}_i^{(0)} = \frac{n^{(0)}\left(n_i^{(1)}+n_i^{(0)}\right)}{n^{(1)}+n^{(0)}}$, respectively. Both test statistics asymptotically follow a chi-square distribution with two degrees of freedom. For the SNP1 in Table 1, we calculated LRT $= 27.11$ and $X_2^2 = 26.90$, corresponding to p-values 1.3×10^{-6} and 1.44×10^{-6}, respectively. Similarly, we calculated LRT $= 6.99$ and $X_2^2 = 6.97$ for SNP2, corresponding to p-values 0.0303 and 0.0306, respectively.

When the genotypes satisfy HWE, we can test the association by testing the independence of Table 2, which is the same as to test the equivalence of two binomial distributions between cases and controls. The one degree of freedom chi-square test for association is

$$X^2 = \frac{2\left(n^{(1)}+n^{(0)}\right)\left[\left(n^{(1)}+n^{(0)}\right)\left(2n_{DD}^{(1)}+n_{Dd}^{(1)}\right) - n^{(1)}\left(2n_{DD}^{(1)} + n_{Dd}^{(1)} + 2n_{DD}^{(0)} + n_{Dd}^{(0)}\right)\right]}{2n^{(1)}n^{(0)}\left(2n_{DD}^{(1)} + n_{Dd}^{(1)} + 2n_{DD}^{(0)} + n_{Dd}^{(0)}\right)\left(2n_{dd}^{(1)} + n_{Dd}^{(1)} + 2n_{dd}^{(0)} + n_{Dd}^{(0)}\right)}. \tag{4}$$

Under the null hypothesis, this statistic approximately equals to the Armitage's trend test statistic obtained from Table 1 [23,24]. For the SNP1 and 2 in Table 1, we first tested the HWE for both SNPs in pooled cases and controls by a chi-square test and did not observe any departure from HWE (SNP1, p $= 0.206$, SNP2, 0.694). We then calculated the X^2 values 27.54 and 6.89, corresponding to p-values 1.54×10^{-7}, and 0.0087 for SNP1 and SNP2, respectively.

2.3 False Positive Can Be Aroused When Population Stratification Is Present

Consider a case–control design, in which each sampled individual is actually a member of one of two subpopulations, but ignored the individual's origin in the analysis. Assume a marker locus with two allele A and a and not associated with a trait. Let r_i, ν_i, and p_{Ai} be the probability of sampling an individual, the disease prevalence, and A allele frequency from ith subpopulation, respectively. The probability of sampling an affected individual from subpopulation i is $\frac{\nu_i r_i}{\nu_1 r_1 + \nu_2 r_2}$. Similarly, the probability of sampling a normal individual from subpopulation i is $\frac{(1-\nu_i) r_i}{1-\nu_1 r_1 - \nu_2 r_2}$. Under the assumption of the independence of the marker locus and disease, the A allele frequency difference between cases and controls is

$$\begin{aligned}
&\Pr(A|Disease) - \Pr(A|Normal) \\
&= \sum_{i=1}^{2} \Big[\Pr(A|subpop\ i) \Pr(subpop\ i|Disease) \\
&\quad - \Pr(A|subpop\ i) \Pr(subpop\ i|Normal) \Big] \\
&= \frac{r_1 r_2 (p_{A1} - p_{A2})(\nu_1 - \nu_2)}{(\nu_1 r_1 + \nu_2 r_1)(1 - \nu_1 r_1 - \nu_2 r)}.
\end{aligned}$$

Thus, as long as $p_{A1} \neq p_{A2}$ and $\nu_1 \neq \nu_2$, the A allele frequency between cases and controls is different, resulting in false positive finding in association studies even the marker is independent of disease. When the sample size is increased, the problem is even worse [7]. In practice, association studies in populations such as African-Americans or Hispanics may be particular to pay attention to population structure. In our simulated data, SNP2 is not associated in the trait in European and African ancestral populations, however, we observed the association in the sample from the admixed population.

3 Genome-Control Approach

One approach to control for population stratification is the genome-control approach (GC) approach proposed by Devlin and Roeder [10] and was further investigated by other investigators [25–28]. This approach requires additional genotypes at M unlinked biallelic markers to control the effect of population stratification. In current whole genome association studies, the markers for controlling population stratification can be selected from the available markers because such kinds of studies will often genotype 100 K or more SNPs. In the presence of population stratification, the test statistics presented so far may not follow chi-square distributions under the null hypothesis of no association. For example, the X^2 test, given by (4), may not follow a chi-square distribution with one degree of freedom when population structure is present. The GC approach simply rescales the statistic by a multiplicative factor λ, and the recalled statistic X^2/λ follows a chi-square distribution with one degree of freedom. The unlinked marker genotype data can be used to estimate the factor λ. Let $X_1^2 \ldots X_M^2$ be the values for the X^2 statistic at M unlinked markers. Devlin and Roeder [10] proposed to use the median of $\{X_1^2 \ldots X_M^2\}/0.456$ as an estimate of λ. Alternatively, Reich and Goldstein [27] proposed to use the mean of $X_1^2 \ldots X_M^2$ to estimate λ. The GC approach is computationally simple, and it allows for a large number of potential subgroups (i.e., it works well with very fine-scale substructures [10]). It can be undertaken with pooled DNA samples, which can be substantially less expensive than the individual genotyping. An investigation of power indicates that the GC approach can generally be more powerful than the TDT when the same sample sizes are the same [25].

We simulated 1,000 unlinked SNPs using the same method in the data section. The mean and median of X^2 statistic values of the 1,000 SNPs is 1.17 and 0.52. The rescaling parameter λ estimated using median and mean is similar. Using the estimate by the mean of the X^2 values, the rescaled X^2 statistic value for SNP1 and SNP2 are 23.6 and 5.9, corresponding to p-values 1.19×10^{-6} and 0.015, respectively.

4 Structured Association Approach

Pritchard et al. [12] proposed an approach called “structured association” (SA), which contrasts with the GC method [11]. SA uses a set of independent genetic markers to estimate the number of subpopulations based on a Markov chain Monte Carlo (MCMC) method and the ancestry probabilities of individuals from putative “unstructured” subpopulations [12]. This information is then used to test for association. When the number of subpopulations is large, the simulation-based test statistic becomes computationally intensive, especially for genome-wide association analysis. Satten et al. [13] extended it by applying latent-class analysis to infer the population structure while simultaneously estimating the model parameters and testing for association. Other similar approaches including the extension to quantitative traits were also discussed by Zhang et al. [28, 29] and Hoggart et al. [30]. There are two approaches to infer the population structure; one is Bayesian approach that uses Markov Chain Monte Carlo (MCMC) to estimate the parameters [12, 30], the other is the maximum likelihood likelihood approach based on mixture models [13, 17, 29]. We introduce STRuctured population Association Test (STRAT) proposed by Pritchard et al. [12] here.

We assume a case–control study and M unlinked markers are genotyped. The STRAT includes two steps. First, a Bayesian clustering method is used to determine both the number of subpopulations and the fraction of the sampled individual’s ancestry in each subpopulation. Specifically, assume that sampled individuals inherit their marker alleles from a pool of K unstructured subpopulations (where K may be unknown). The allele frequencies at each locus within each subpopulation are assumed to be unknown and need to be estimated. Let q_{ik} denote the proportion of the i th individual's genome originated from subpopulation k. Using the genotypes of n sampled individuals at M unlinked markers, Pritchard et al. and Falush et al. [31] proposed an MCMC approach to estimate the parameters: the number of subpopulations K, and $Q = \{q_{ik} : i = 1, \ldots, M; k = 1, \ldots, K\}$, the allele frequencies at each locus within each subpopulation. The method can be applied to most of the commonly used genetic markers, including microsatellite markers and SNPs, and can produce accurate results using modest numbers of markers. The accuracy of the inference depends on the sample size, the number of markers used, and the magnitude of allele-frequency differences between the subpopulations. In the second step, a likelihood ratio test, based upon the detailed population structure, is used to test a null hypothesis H_0 that subpopulation allele frequencies at the candidate locus are independent of phenotype, against an alternative hypothesis H_1; where the subpopulation allele frequencies at the candidate locus are associated with phenotype. Let G denote the list of genotypes of all sampled individuals at the candidate locus, and P_0 and P_1 denote subpopulation allele frequencies at the candidate locus under H_0 and H_1, respectively. The statistic is the likelihood ratio

$$\Lambda\left(G\right) = \frac{\mathrm{Pr}_1\left(G; \hat{P}_1, \hat{Q}\right)}{\mathrm{Pr}_0\left(G; \hat{P}_0, \hat{Q}\right)},$$

where $\mathrm{Pr}_0\left(G; \hat{P}_0, \hat{Q}\right)$ and $\mathrm{Pr}_1\left(G; \hat{P}_1, \hat{Q}\right)$ are the distributions of G under H_0 and H_1, respectively, and $\hat{P}_0$, $\hat{P}_1$, and $\hat{Q}$ are the estimates of P_0, P_1, and Q, and the values of $\hat{P}_0$, $\hat{P}_1$, and $\hat{Q}$ can be obtained from the MCMC procedure in the first step.

The p-value of this test is evaluated by the following simulation procedure: Generate new genotypes at candidate locus under H_0 for each individual as independent random draws from $\mathrm{Pr}_0\left(\cdot|; \hat{P}_0, \hat{Q}\right)$. Repeat this procedure B times, and obtain genotype data set $G^{(1)}, \ldots, G^{(B)}$. The empirical p-value is given by

$$p - value = \frac{1}{B}\#\{b : \Lambda\left(G^{(b)}\right) > \Lambda\left(G\right)\}$$

where $\#A$ denotes the number of members in set A. Simulation studies show that this method can control the population stratification provided the number of unlinked markers is large enough, and it is more powerful than TDT test in most of the cases. One of the difficult problems is the estimation of the number of subpopulations K; especially when there are a large number of potential subgroups [11].

We performed the SA analysis for the data in Table 1. We first estimated the number of subpopulations using the simulated 1,000 SNPs. We chose 100,000 replications after burnin length 100,000 in analysis using STRUCTURE. We obtained the log probabilities of the data given numbers of subpopulations 1, 2, and 3 are -1735367.3, -1534082, and -1535688.0, respectively. Thus, the posterior probability of 2 subpopulations model is close to 1 assuming a uniform prior on 1, 2, and 3 subpopulations, suggesting a model with 2 subpopulations has the best fit. Using the information obtained from STRUCTURE we performed the association test for both SNP1 and 2, and we obtained the corresponding p-values $<10^{-6}$ and 0.048 based on 1000,000 permutations, respectively.

5 Methods Based on Principal Components (PC)

The principal components of genetic marker data have been used for characterizing population differences [32] and recently have been applied to correcting population stratification in genetic association studies. The central idea is that an individual's genetic background can be represented by his/her genetic markers, which can be summarized using the principal components of marker data. The principal components can then be incorporated into a statistical model in analysis. For instance, we can model the association between a trait Y and genetic marker g using a generalized linear model

$$f\left(E\left(Y\right)\right) = \beta g + \mu\left(T\right) + \varepsilon,$$

where $f\left(E\left(Y\right)\right)$ is a link function, $\mu\left(T\right)$ a function of principal components obtained from marker data, and ε is a random error. Several methods have been proposed to control for population stratification by using the principal components obtained from a set of unlinked markers. Zhu et al. [17] modeled $\mu\left(T\right)$ in a mixture model incorporated in the logistic regression, while Chen et al. [14] and Zhang et al. [16] proposed semi-parametric approach using kernel smoothers, and Price et al. [15] directly considered a linear function of T, which is a special case of Chen et al. [14] and Zhang et al. [16]. Consider a case control study with $n^{(1)}$ cases and $n^{(0)}$ controls. The M unlinked markers for controlling population stratification are represented by a matrix $X = \left(X_1, \ldots, X_{n^{(0)}+n^{(1)}}\right)^T$, where $X_i = \left(x_{i1}, \ldots x_{iM}\right)^T$, $i = 1, \ldots, n^{(0)} + n^{(1)}$, x_{im} is the genotype value of the mth marker for the ith individual and its values are 0, 1, and 2 for marker genotypes 11, 12, and 22, respectively and the superscript T represents a transpose of a vector or matrix. The sample covariance matrix of marker data is $\Sigma = Cov\left(X\right) = \sum_{i=1}^{n^{(0)}+n^{(1)}} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T$, which is an $M \times M$ matrix. Let e_j be the jth eigenvector corresponding to the jth largest eigenvalue of Σ. The corresponding PC for the ith individual is $t_{ij} = \left(X_i - \bar{X}\right)^T e_j$. Let $T_i = \left(t_{i1}, \ldots, t_{is}\right)$ be the first few PCs, where $s \leq M$. All the PC-based approaches are based on T_i. In the following sections, we discuss the PC-based approaches in detail.

5.1 Mixture Model

Consider the case–control design mentioned earlier. Let y_i represent the disease value for the ith individual, with $y_i = 1$ if affected and 0 if unaffected, and g_i the genotypic value for a candidate locus which we wish to test for association with the trait. Zhu et al. [17] suggested using T_i to infer an individual's genetic background and correct the effect of population stratification. For example, considering a data set consisting of samples simulated from two different populations. Figure 1 demonstrates the histograms of the first and second principal components. It can be seen that the individuals from two different populations are clustered into two groups. When individuals are sampled from the simulation we mentioned at the beginning, where an individual is sampled from an admixed population simulated by continuous gene flow model, the distribution of the first two principal components are shown in Figure 2. The first principal component can fit an admixture distribution of two normal distributions well, suggesting each individual is admixed from the gene pools of two ancestral populations. The correlation between the first principal component and the true ancestry is extremely high, reaching to 0.987. But the second principal component still follows a normal distribution (p value of the two-sample Komogorov–Smirnov test is 0.33) and the correlation with the true ancestry is 0.025. Thus, Zhu et al. suggested using a mixture model of the principal components of marker data to infer an individual's genetic background. If there are K subpopulations, the principal components are assumed to follow approximately a mixture of K normal distributions. Because the principal components are independent, conditional on the kth subpopulation we can assume that the distribution of T_i is the product of M normal distributions: $f\left(T_i|k\right) = \prod_{j=1}^{s} N\left(t_{ij}|\mu_{kj}, \sigma_{kj}^2\right)$, where $N\left(t_{ij}|\mu_{kj}, \sigma_{kj}^2\right) = \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} \exp\left[-\frac{(t_{ij}-\mu_{kj})^2}{2\sigma_{kj}^2}\right]$ is a normal density function. The distribution of T_i is

$$f\left(T_i\right) = \sum_{k=1}^{K} \lambda_k \prod_{j=1}^{s} N\left(t_{ij}|\mu_{kj}, \sigma_{kj}^2\right),$$

where λ_k is the probability that an individual originates from the kth subpopulation, with the restriction $\sum \lambda_k = 1$.

Given an individual is from the kth subpopulation, a logistic regression model is applied to model the association between a candidate gene and a trait:

$$\log\left[\frac{\Pr\left[y_i = 1|g_i,k\right]}{\Pr\left[y_i = 0|g_i,k\right]}\right] = \mu + \beta g_i + \delta_k, \tag{5}$$

where δ_k indicates the effect of the kth subpopulation subject to the restriction that $\delta_K = 0$, and β represents the effect of the candidate gene. It is assumed that the effect of the candidate gene is the same across subpopulations, but this is not a

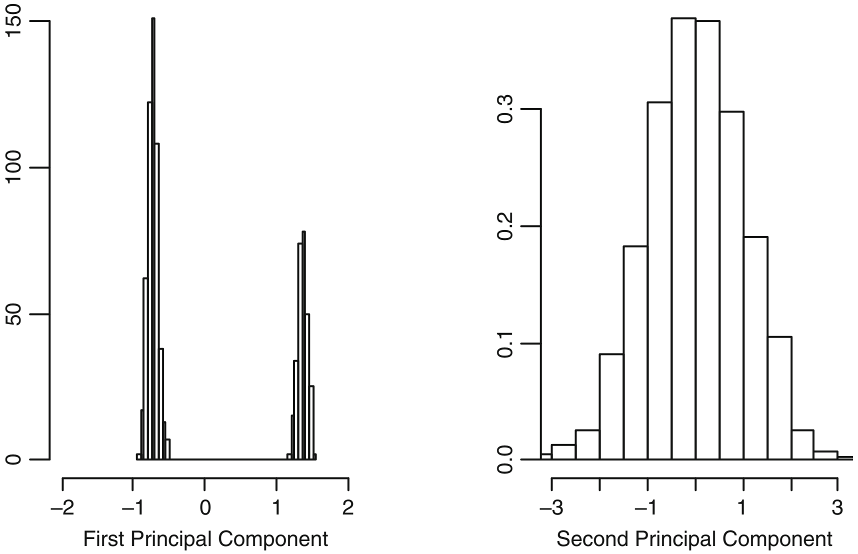


Fig. 1 The histograms of the first two principal components when data consisting of samples simulated from two different populations

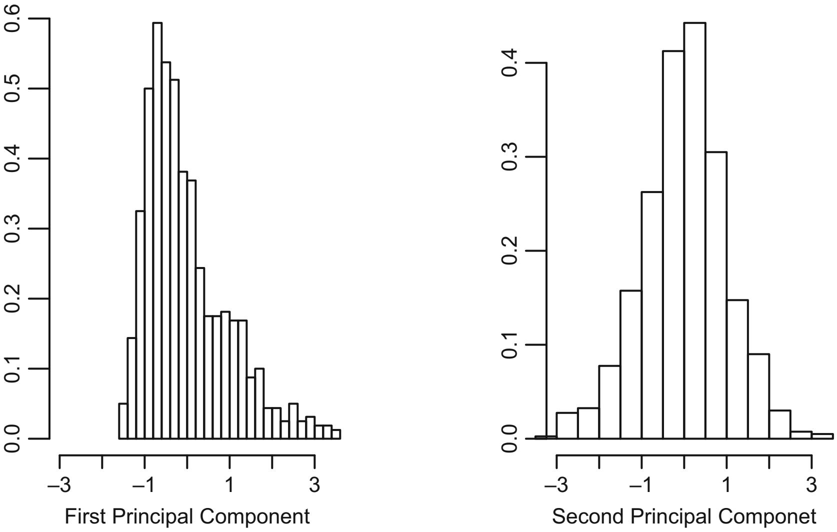


Fig. 2 The histograms of the first two principal components when data consisting of samples simulated from an admixed population with two parental populations

necessary assumption because we can use a population specific β. The likelihood of the observed data is then

$$L_{LOGISTIC} = \prod_{i=1}^{N} \sum_{k=1}^{K} \lambda_k \Pr\left[y_i | g_i, k\right] f\left(T_i | k\right) / f\left(T_i\right). \tag{6}$$

To test the null hypothesis $\beta = 0$, a score statistic can be applied. To infer the number of subpopulations, Zhu et al. [17] suggested use Bayesian Information Criterion (BIC) to estimate the number of subpopulations. To estimate the BIC value given K, the E-M algorithm for a mixture of normal distributions as described by Celeus and Govaert [33] is used. For the simulated data in Table 1 with 1,000 SNPs adjusting for population structure, the BIC values corresponding to $\mathrm{K} = 1, 2, 3,$ and 4 are 9695.1, 9531.8, 9570.3, and 9585.5, suggesting 2 subpopulations best fits the data. Further association test resulted $\mathrm{p} = 9.75 \times 10^{-7}$ and 0.027 for SNPs 1 and 2, respectively.

A question that arose from the principal components analysis is how many principal components should be used in controlling for population stratification. This question can also be modified as which principal component will contribute ancestry information. From the analysis of the second principal component, we observed that a principal component will distribute as a normal distribution if it does not contribute any ancestry information. We can thus test weather a principal component deviates from a normal distribution by the Kolmogorov–Smirnov test. In the example of the simulated data, we only identified that the first principal component is significantly deviated from normal distribution $\left(\mathrm{p} = 9.2 \times 10^{-8}\right)$. The p-values of the rest of the principal components are greater than 0.34. This result suggests that only the first principal component should be used in controlling for population stratification. However, it should also be cautioned that such a test is not very powerful, and a likelihood ratio test of a mixture model against a normal distribution may be favored.

5.2 Semi-Parametric Approach

We consider a case–control design and use the notation given in the previous section. The ith individual's principal components T_i is also called the genetic background value of the ith individual. Chen et al. [14] proposed a QualSPT method to model the relationship between a trait and candidate gene loci and genetic background by a semi-parametric logistic model:

$$\log \frac{P\left(y_i = 1 | X_i, T_i\right)}{1 - P\left(y_i = 1 | X_i, T_i\right)} = X_i^T \beta + \mu\left(T_i\right),$$

where $\mu\left(T\right)$ is an unknown smoothing function of genetic background variable T and is not parameterized. The advantage of this model is that the effect of population stratification to a phenotype is assumed to be taken care by the function $\mu(T)$ in the logistic regression model. $\mu(T)$ can either be a linear or nonlinear function, therefore, with great flexibility in modeling the relationship between genetic background and phenotype. Under this model, the association test is to test the null hypothesis $H_0 : \beta = 0$. The log-likelihood function is

$$\begin{aligned} L(\beta, \mu) &= \sum_{i=1}^{n} l(\beta, \mu(T_i); X_i, y_i) \\ &= \sum_{i=1}^{n} \{y_i \left[X_i^T \beta + \mu(T_i)\right] - \log\left[1 + \exp\left(X_i^T \beta + \mu(T_i)\right)\right]\}. \end{aligned}$$

The QualSPT statistic is a likelihood ratio test statistic

$$\Lambda = \frac{L\left(\hat{\beta}, \hat{\mu}_1(T_i)\right)}{L(0, \hat{\mu}_0(T_i))}$$

where $\hat{\mu}_0(T_i)$ and $\hat{\mu}_1(T_i)$ are the maximum likelihood estimators (MLE) of $\mu(\cdot)$ under the null and alternative, respectively, and $\hat{\beta}$ is the MLE of β under alternative. Under the null hypothesis H_0, the QualSPT statistic follows a chi-squared distribution with degrees of freedom equals to the dimension of β. The estimation of parameter β and nonparameter function $\mu(T)$ under semi-parametric logistic models has been developed [34–36]. Chen et al. [14] proposed to follow the local likelihood approach [35]. For a known smoothing parameter h and a given kernel function K $(\cdot)$; the estimation method is an iterative procedure that follows two steps:

Step 1. For a given β, η is obtained by solving the following equation:

$$\sum_{i=1}^{n^{(0)}+n^{(1)}} K\left(\frac{T_i - T}{h}\right) \frac{\partial}{\partial \eta} l(\beta_m, \eta, X_i, y_i) = 0.$$

Denote $\hat{\mu}_m(T_1)$, $\hat{\mu}_m(T_2), \ldots, \hat{\mu}_m(T_n)$ be the solutions of η for $T = T_1$, $T = T_2, \ldots, T = T_n$, respectively. Here, β_m is the current estimated value of β.

Step 2. Solving the equation for β by $\sum_{i=1}^{n^{(0)}+n^{(1)}} \frac{\partial}{\partial \beta} l(\beta, \hat{\mu}_m(T_i), X_i, y_i) = 0$ results the updated parameter estimate β_{m+1}.

We then repeat this two-step process until convergence occurs.

To choose Smoothing Parameter h, Chen et al. [14] and Zhang et al. [16] suggested to choose h that minimizes a Kolmorgorov test statistic. Specifically, for a given h; we perform QualSPT to all the M unlinked markers and obtain the p-values $p_1, \ldots, p_M$. These p-values follow a uniform distribution if population stratification is well controlled. Let F_n be the empirical distribution function of the p-values $p_1, \ldots, p_M$ and F be the uniform distribution function. To test the null hypothesis

H_0: p-values $p_1, \ldots, p_M$ follow a uniform distribution, the test statistic of the Kolmorgorov test is $L(h) = \max_x |F_n(x) - F(x)|$, and we reject the null hypothesis when $L(h)$ is large. The Kolmorgorov test statistic $L(h)$ is a function of h. Zhang et al. and Chen et al. [14, 16] proposed to choose h such that

$$h^* = \arg\min_h L(h).$$

This procedure also provides a method to check if the population stratification has been well controlled by the set of unlinked markers. If the p-value of the Kolmorgorov test $(h = h^*)$ is greater than a prespecified significance level, e.g. 0.05, we may consider that the population stratification has been well controlled. Otherwise, these M unlinked markers cannot well control the population stratification, and additional markers might be required.

As an application of SPT approach to the data, we first check if the markers are enough to control the population stratification. With 1,000 SNPs, we found that population can be well controlled (Komogorrov test statistics $= 0.785$). We then performed association tests and found SNP1 is significant and 2 is not based on 1000,000 permutations (SNP1, p $= 4 \times 10^{-6}$ SNP2, p $= 0.485$).

5.3 Linear Model Approach

Using the principal components of genetic marker data to account for the population stratification has also been further proposed by Price et al. [15], although the method is conceptually the same as the methods described before. This method first performs a regression analysis by regressing both the phenotype and marker genotype values on the principal components for unrelated data. Association between the phenotype and marker is then tested using the residual correlation. This approach is simple and easily applied to the data with testing a large amount of markers. In detail, this method first calculates the residual after regressing the first L principal components for both trait value and genotypic value by

$$y_i = \beta_0 + \beta_1 t_{i1} + \ldots + \beta_L t_{iL} + \varepsilon_i$$

and

$$g_i = \alpha_0 + \alpha_1 t_{i1} + \ldots + \alpha_L t_{iL} + \tau_i,$$

where ε_i and τ_i are random errors. Let $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_L$, $\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_L$ be the least-squares estimators of $\beta_0, \beta_1, \ldots \beta_L$, $\alpha_0, \alpha_1, \ldots, \alpha_L$, respectively. Since the principal components are orthogonal, $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_L$, $\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_L$ can be easily calculated by

$$\hat{\beta}_l = \frac{\sum\limits_{i=1}^{N} y_i t_{il}}{\sum\limits_{i=1}^{N} t_{il}^2} \quad \text{and} \quad \hat{\alpha}_l = \frac{\sum\limits_{i=1}^{N} g_i t_{il}}{\sum\limits_{i=1}^{N} t_{il}^2}.$$

The phenotype and genotype residuals for each individual are calculated by

$$y_i^* = y_i - \hat{\beta}_0 - \hat{\beta}_1 t_{i1} - \ldots - \hat{\beta}_L t_{iL}$$

and

$$g_i^* = g_i - \hat{\alpha}_0 - \hat{\alpha}_1 t_{i1} - \ldots - \hat{\alpha}_L t_{iL}.$$

The test statistic of testing association between the phenotype and marker is defined by $T = (N - L - 1)\, r^2$, where r is the correlation between y_i^* and g_i^*, which follows a chi-square distribution with one degree of freedom. Because of its simplification and easily programming, this method has also been extended to combine family and unrelated samples [37]. Intuitively, y_i^* and g_i^* can be viewed as the trait and marker values after removing the effect of population structure. In other words, we can consider y_i^* and g_i^* as if they are from a homogenous population. Any association test based on y_i^* and g_i^* will not be affected by population structure, including testing gene–gene interaction. This method can be applied to both quantitative and qualitative traits. We used the first ten principal components to control for population stratification in the simulated data and obtained the p-values 1.29×10^{-6} and 0.115 for SNP1 and 2, respectively.

6 Discussion

Since the traditional linkage analysis has low power to detect common variants with small odds ratio, association studies have been waged to search for the genetic variants of complex traits [1]. With the rapid technological advances, large scale testing of thousands of SNPs across the genome in large sample size will become routine. The first genome-wide association studies have been published recently and several new genetic variants have been detected to be associated with macular degeneration, obesity, types 1 and 2 diabetes, prostate cancer and multiple sclerosis, among others, suggest promise for association studies [38–46]. However, false positive findings have also been concerned because of a large scale tests. In addition, population stratification is another source contributing false positive in association studies and the problem can worsen when sample size is increased [1]. Genome-control approach is a computationally simple and fast method, and therefore is appareling in the genome-wide association study when thousands of SNPs are tested. With a large number of SNPs available, the performance of GC can be very well, although the power of GC is dependent on the markers chosen for controlling population structure. For example, GC will reduce the power if only ancestry informative markers are used because such kinds of markers are more likely to be associated with a trait in admixed populations. Therefore, the factor to adjust for population may be overestimated. In comparison, Structure Association method based on MCMC approach is computationally intense. When a large number of markers are available such as in whole genome association studies, it is almost impossible for STRUCTURE to use all the markers unless a subset of ancestry informative markers (AIMs) are selected. One way to search the AIMs is by selecting the unlinked markers with large F_{st} values or large allele frequency differences between ancestral populations. However, the selection of AIMs may be biased to known population structure, resulting inadequately treatment of unknown subtle population structure. The principal component methods can be computational fast and is favorable in current genome-wide association studies. The principal component methods work well with a large number of random markers and even with some markers in linkage disequilibrium, and are superior to STRUCTURE. To account for multiple comparisons, Bonferroni correction is often too conservative because the linkage disequilibrium among markers. Permutation test by randomly shifting the disease status accounts for the LD among SNPs, therefore, leads to a much accurate estimate of type I error rate with the cost of more computation. Since the axes of principal components of the marker data are unchanged when the disease status is shifted, the principal components-based approaches are still fast to calculate the empirical p-value through permutations. However, a challenge of principal components-based methods is how many principal components we should use in order to well control the population stratification. It has been suggested that the first ten principal components are usually adequate for samples from most populations [15]. However, it may be not enough for samples coming from a population admixed by many subpopulations. Including too many principal components in a model will result in loss of statistical power and complexity of the statistical model. We suggest to test if a principal component fits a normal distribution vs. a mixture of normal distributions. In practice, we can also select principal components by stepwise selection in the regression analysis. When population stratification is caused by the additive effect of both natural selection and admixture process acted locally, using the genome-wide markers may not well eliminate the effect of population stratification. How to eliminate the effect of population stratification of a local region and maintain the statistical power needs further investigations.

Web Resources

http://pritch.bsd.uchicago.edu/software.html (for Structure 2.1 and STRAT programs)
http://bioinformatics.med.yale.edu (for semi-parametric program)

The software of mixture model approach and principal component method are available upon request to xzhu1@darwin.case.edu

Acknowledgements This work was supported by grant from National Human Genome Research Institute (HG003054).

References

1. Risch N, Merikangas K, (9–13–1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
2. Risch NJ, (6–15–2000) Searching for genetic determinants in the new millennium. Nature 405:847–856
3. Clayton, (2007) Population association. In: Balding DJ, Bishop M and Cannings C (eds) Handbook of statistical genetics. pp. 939–959
4. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43: 520–526
5. Lander ES, Schork NJ, (9–30–1994) Genetic dissection of complex traits. Science 265: 2037–2048
6. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37:868–872
7. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512–517
8. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516
9. Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458
10. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
11. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228
12. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181
13. Satten GA, Flanders WD, 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
14. Chen HS, Zhu X, Zhao H, Zhang S (2003) Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 67:250–264
15. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909
16. Zhang S, Zhu X, Zhao H (2003) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 24:44–56
17. Zhu X, Zhang S, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23:181–196
18. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, De Jager PL, Mignault AA, Yi Z, De TG, Essex M, Sankale JL, Moore JH, Poku K, Phair JP, Goedert JJ, Vlahov D, Williams SM, Tishkoff SA, Winkler CA, De L, V, Woodage T, Sninsky JJ, Hafler DA, Altshuler D, Gilbert DA, O'Brien SJ, Reich D (2004) A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet 74:1001–1013
19. Agresti A (2002) Categorical data analysis. Second edition, John Wiley & Son I
20. Chakravarti A, Buetow KH, Antonarakis SE, Waber PG, Boehm CD, Kazazian HH (1984) Nonuniform recombination within the human beta-globin gene cluster. Am J Hum Genet 36:1239–1258
21. Lewontin RC, Kojima KI (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472
22. Lewontin RC (1964) The interaction of selection and linkage. I. General Considerations; Heterotic Models. Genetics 49:49–67

23. Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53: 1253–1261
24. Armitage P (1955) Test for linear trend in proportions and frequencies. Biometrics 11:375–386
25. Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944
26. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166
27. Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4–16
28. Zheng G, Freidlin B, Gastwirth JL (2006) Robust genomic control for association studies. Am J Hum Genet 78:350–356
29. Zhang S, Zhao H (2001) Quantitative similarity-based association tests using population samples. Am J Hum Genet 69:601–614
30. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504
31. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587
32. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press
33. Celeux G, Govaert G (1995) Gaussian parsimonious clustering model. Pattern Recognition 28:781–793
34. Severini TA, Wong W (1992) Profile likelihood and conditionally parametric models. Ann Stat 20:1768–1802
35. Severini TA, Staniswalis JG (1994) Quasi-likelihood estimation in semiparametric models. J Amer Stat Assoc 89:501–511
36. Simonoff JS (1996) Smoothing Methods in Statistics
37. Zhu X, Li S, Cooper RS, Elston RC (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet In Press.
38. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
39. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (5–11–2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894
40. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, bers-Akkers MT, Godino-Ivan MJ, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39:631–637
41. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PI, Gabriel SB, Mirel DB, Ivinson AJ, Pericak-Vance MA, Gregory SG, Rioux JD, McCauley JL, Haines JL, Barcellos LF, Cree B, Oksenberg JR, Hauser SL (8–30–2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357:851–862
42. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, Christman MF (4–14–2006) A common genetic variant is associated with adult and childhood obesity. Science 312:279–283

43. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson BK, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S (6–1–2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336
44. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (6–1–2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345
45. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38:617–619
46. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'Shea AM, Zogopoulos G, Cotterchio M, Newcomb P, McLaughlin J, Younghusband B, Green R, Green J, Porteous ME, Campbell H, Blanche H, Sahbatou M, Tubacher E, Bonaiti-Pellie C, Buecher B, Riboli E, Kury S, Chanock SJ, Potter J, Thomas G, Gallinger S, Hudson TJ, Dunlop MG (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 39:989–994

Family-Based Association Studies

Kui Zhang and Hongyu Zhao

Abstract Over the past decade, association studies based on linkage disequilibrium have become increasingly popular for detecting genetic variations underlying complex human diseases because association-based methods have been shown to have more power than traditional linkage-based methods in theoretical and empirical studies. There are two general designs in association studies: family-based designs that use pedigrees and population-based designs that use unrelated individuals. Although population-based designs are generally more powerful than family-based designs, and the recruitment of unrelated individuals is easier than the recruitment of families, they are subject to bias in the presence of population stratification. As a compromise between linkage studies and population-based association studies, family-based association designs can have similar power with population-based designs and are robust in the presence of population stratification. Therefore, family-based association designs have received great attention recently. In this chapter, we first review methods that can analyze the simplest family-based association design with one affected offspring with its two parents, all genotyped at a bi-allelic marker locus. We then discuss its various extensions that can increase power and utilize multi-allelic markers, families with multiple siblings, families with incomplete parental genotypes, quantitative traits, and multiple tightly linked markers. The association methods using family-based designs can be broadly classified into two groups: nonparametric methods based on the allele counting and parametric methods based on the likelihood function. Although these methods result in similar test statistics for the simplest family-based association design with one affected offspring with its two parents, their extensions on more complex situations vary greatly. Further developments of statistical methods to utilize general pedigrees and to detect gene–environment interactions are also discussed. Finally, we conclude this review by listing the available software packages that can carry out the analysis of family-based association designs and illustrating some of them based on a real data set.

K. Zhang (✉)
Section on Statistical Genetics, Department of Biostatistics University of Alabama at Birmingham, Birmingham, AL 35294, USA
e-mail: kzhang@ms.soph.uab.edu

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_7, © Springer-Verlag Berlin Heidelberg 2010

1 Introduction

Over the past decade, many complex human diseases such as hypertension, diabetes, and obesity [1–3], have increased in incidence both in the United Sates and in developed countries, and pose a striking threat to human health. During this period, considerable efforts were expended to dissect the genetic etiology of such diseases to help us better understand their pathogenesis with the intent of yielding improved strategies for prevention, diagnosis, and treatment [4]. The first step toward this ultimate goal is to determine genetic variations underlying these diseases. Linkage analysis and association analysis are the two main strategies used by researchers in this context and both have been successfully applied to dissect genes responsible for simple Mendelian diseases in which only one or two genes have a major impact, including Huntington diseases [5], cystic fibrosis [6], Fanconi anemia [7], and breast cancer [8, 9]. However, both strategies have been less successful so far until very recently for identifying genes responsible for complex diseases, including hypertension, diabetes, obesity, schizophrenia, alcoholism, etc., that likely originate from the small effects of many genes, as well as gene–gene and gene–environment interactions.

The underlying principle is the same for linkage analysis and association analysis: both are based on LD, which refers to the nonrandom association between alleles at different tightly linked loci in the population [10]. In linkage analysis data are collected on pedigrees enriched with affected members. If a marker is close to the disease locus, the alleles at the marker locus and the disease locus will tend to co-segregate together. The genomic regions between them will be shared among many affected individuals within a pedigree. However, because only a small number of observed recombinants occur within a pedigree, genes cannot be localized by this approach to a small interval, generally on the order of one megabases to several megabases, making identification of the causal genetic variants difficult [11, 12]. A commonly used population-based association design is the case–control design, in which data are collected on unrelated, affected, and unaffected individuals. Because of the very large number of recombination events over the past generations, the genomic regions shared by the unrelated, affected individuals will be much shorter than those shared by affected individuals in an extended pedigree. Thus, the interval containing the causal genetic variants can be narrowed up to several kilobases [6, 13]. Therefore, association studies for both genome-wide mapping and fine mapping based on LD have become increasingly popular as they offer a potentially more cost effective and powerful approach for gene mapping than linkage analysis [14–18]; Botstein and Risch, 2003.

Recently, genome wide association (GWA) studies, which aim to genotype hundreds of thousands of single nucleotide polymorphisms (SNPs) across the human genome for a large number of samples, have been proposed and offer great promise to detect genes underlying complex human diseases. GWA studies based on a large number of unrelated individuals have already shown great success and there are more and more genes found to be association with several complex human diseases [19–21]. However, one of the major limitations of case–control association studies is that they are subject to bias in the presence of population stratification. In case–control studies, population stratification arises when samples are selected from several genetically different populations with different proportions in cases and controls. Population stratification can generate suspicious association between markers and the disease susceptibility locus (DSL). That means a positive association can occur even neither is the allele itself a cause of the disease nor is the allele in linkage disequilibrium with a susceptible allele at the disease gene in the presence of population stratification. Thus, it is always a concern for association studies in heterogeneous populations, such as populations in major cities in the United States. In contrast, appropriate analyses of family-based association studies are not affected by population stratification. In addition, significant findings in family-based association studies indicate that the marker locus is not only associated but also linked with the DSL. Although the family-based design can be less powerful than the case–control design, the power difference between these two types of designs is generally small, especially when case–parent trios are used [17, 22, 23]. Therefore, family-based designs will continue to play important roles in association studies.

In this chapter, we review methods for the case–parent design and discuss its various extensions that can increase power and utilize families with various structures (e.g., families with multiple siblings, families with incomplete parental genotypes, general pedigrees), quantitative traits, and multiple tightly linked markers. We also list available software packages that can carry out such analysis and illustrate some of them using the Oxford Angiotensin converting enzyme (ACE) data set [24].

2 Basic Notations

In this review, we will mainly focus on nuclear families with a pair of parents and one or more offspring, and introduce the following notations. We assume a total of n nuclear families are collected. In the ith family, there are n_i offspring. The genotype at a multi-allelic locus of the jth offspring in the ith family is denoted by: $g_{ij}\,(i = 1, \ldots, n; j = 1, \ldots, n_i)$. The corresponding genotypes of the mother and the father in the ith family are denoted by g_{im} and g_{if}, respectively. The genotypes of offspring and parents in the ith family are denoted by $g_{io} = (g_{i1}, \ldots, g_{in_i})$ and $g_{ip} = (g_{im}, g_{if})$, respectively. For a multi-allelic marker with k alleles, the alleles are denoted by $A_1, \ldots, A_k$. Sometimes, we need to work on coded genotypes and denote them as $X_{io} = (X_{i1}, \ldots, X_{in_i})$ and $X_{ip} = (X_{im}, X_{if})$, which are functions of g_{io} and g_{ip}. The definition of X depends on the context. For example, X_{ij} can be defined as the number of copies of allele A_1 at a maker locus. We further use $Y_{io} = (Y_{i1}, \ldots, Y_{in_i})$ and $Y_{ip} = (Y_{im}, Y_{if})$ to denote the phenotypic values of offspring and parents in the ith family. Y can either be qualitative or quantitative phenotypes. For qualitative phenotypes with two status of affected and unaffected, we use $Y = 1$ to represent the affected individual and $Y = 0$ to represent the unaffected individual. If we do not specifically refer to the ith family in some of formulas, we omit the subscript i for simplification.

3 Qualitative Traits, Trios, Bi-Allelic Markers

The simplest family-based design for association studies is the case–parent design, in which an affected individual and its parents are collected and genotyped at bi-allelic markers. The alleles transmitted from parents to the affected offspring and the alleles of not transmitted can be determined based on the observed genotype data. Thus, a 2 by 2 transmission/nontransmission table for a bi-allelic marker with alleles A_1 and A_2 from n case–parent trios can be constructed:

	Nontransmitted		
Transmitted	A_1	A_2	Total
A_1	t_{11}	t_{12}	t_{1+}
A_2	t_{21}	t_{22}	t_{2+}
Total	t_{+1}	t_{+2}	$4n$

In this table, t_{ij} $(i = 1, 2; j = 1, 2)$ represents the number of parents who have genotype A_iA_j and transmit allele A_i to the affected offspring. An appropriate analysis for the data presented in this table is the McNemar test, named (in this context) the Transmission/Disequilibrium Test (TDT) by [25]: $\mathrm{TDT} = (t_{12} - t_{21})^2 / (t_{12} + t_{21})$, which compares the number of A_1 alleles transmitted to the offspring from theirs parents and the number of A_1 alleles not transmitted.

The TDT has several advantages. First, it only assumes the first Mendel's law of inheritance. The specification of the disease model and the distribution of the disease in the general population are not required and will not affect its validity. Thus, the TDT is not only robust to the population stratification but also robust to any misspecification of the disease model and the distribution of the disease. Second, the TDT test statistic has an asymptotic chi-square distribution with one degree of freedom if either $\theta = \frac{1}{2}$ or $\delta = 0$, where θ and δ are the recombination fraction and the linkage disequilibrium between the marker locus and the DSL, respectively. Therefore, it is clear that the TDT is a test for both linkage and association [26]. Initially, the TDT was proposed by Spielman et al. [25] to test the linkage between a marker locus and the DSL in the presence of association. Instead, it is now more often used to test the association in the presence of linkage. Actually, we can have three types of null hypotheses: no linkage and no association, no linkage in the presence of association, and no association in the presence of linkage, but only one alternative hypothesis: the maker is in both linkage and association with the DSL. In most situations, it is not important to distinguish these null hypotheses because the TDT statistic and many of its extensions are valid under any one of these three null hypotheses. However, some extensions of the original TDT are only valid under the null hypothesis of no linkage between the marker locus and the DSL [26], because the distribution of those test statistics is derived under the null hypothesis of no linkage. In such situation, it is important to explicitly state the null hypothesis that will be tested.

3.1 Qualitative Traits, Trios, Multi-Allelic Markers

In this subsection, we first describe the extensions of the original TDT for multi-allelic markers. Similar with the transmission/nontransmission table from a bi-allelic marker, we can construct an m by m transmission/nontransmission table for a multi-allelic marker with k alleles $A_1, \ldots, A_m$ from n case–parent trios:

	Nontransmitted			
Transmitted	A_1	$\cdots$	A_k	Total
A_1	t_{11}	$\cdots$	t_{1k}	t_{1+}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_k	t_{21}	$\cdots$	t_{kk}	t_{k+}
Total	t_{+1}	$\cdots$	t_{+k}	$4n$

In this table, t_{ij} $(i = 1, \ldots, k; j = 1, \ldots, k)$ represents the number of parents who have genotype A_iA_j and transmit allele A_i to the affected offspring. Several test statistics have been constructed from this table. A direct generalization of the TDT is to test if the table of transmission/nontransmission is symmetry: $TDT_c = \sum_{i<j} \left(t_{ij} - t_{ji}\right)^2 / \left(t_{ij} + t_{ji}\right)$, which has an asymptotic chi-square distribution with $k\left(k-1\right)/2$ degrees of freedom [27–29]. Clerget-Darpoux [27] proposed a statistic to test the marginal homogeneity: $TDT_m = \sum_{i=1}^{k} \left(t_{i+} - t_{+i}\right)^2 / \left(t_{i+} + t_{+i}\right)$. Spielman and Ewens [30] proposed a similar statistic to test the marginal homogeneity: $TDT_{SE} = \frac{k-1}{k} \sum_{i=1}^{k} \left(t_{i+} - t_{+i}\right)^2 / \left(t_{i+} + t_{+i} - t_{ii}\right)$. As noted by Sham [31] and Schaid [32], TDT_m and TDT_{SE} may not have a chi-square distribution with $k - 1$ degrees of freedom and tend to be anticonservative. Cleves et al. [33] derived the exact calculation of p-values for these statistics. In addition, the permutation procedure can be and has been widely used to assess their significance appropriately [34, 35].

In addition to the nonparametric tests based on the contingency table, retrospective likelihood-based methods have been developed to analyze multi-allelic markers. For a case–parent design, the conditional probability of parental genotypes and offspring's genotype, g_m, g_f, and g_o given the disease status of offspring, D, is

$$L^{(F)} = P\left(g_m, g_f, g_o | D\right) = P\left(g_m, g_f | D\right) P\left(g_o | g_m, g_f, D\right) = L^{(P)} L^{(O)}.$$

In this formula, $L^{(P)}$ is the conditional probability of parental genotypes given the disease status of offspring and $L^{(O)}$ is the conditional probability of offspring's genotype given the parental genotypes and the disease status of offspring [36]. Although the use of the full conditional likelihood $\left(L^{(F)} = L^{(P)} L^{(O)}\right)$ may be more powerful, they may yield biased results in the presence of population stratification [32, 36]. Therefore, most of likelihood-based methods model the probability of an offspring's genotype conditioning on parental genotypes and the disease status of offspring $\left(L^{(O)}\right)$. Since the probability of offspring's genotype is conditioning on parental genotypes, these methods are robust to population stratification. Sham and Curtis [29] proposed a conditional logistic regression model with the likelihood function of $\log(L) = \sum_{i<j} [n_{ij} \log(p_{ij}) + n_{ji} \log(p_{ji})]$, where p_{ij} is the probability of transmitting allele A_i given that the parent has the genotyping A_iA_j and n_{ij} is the number of parents who have the genotype A_iA_j but transmit the allele A_i to the offspring. They assumed that the logarithm of the odds has the following form: $\log \frac{p_{ij}}{p_{ji}} = b_i - b_j$, where b_i and b_j are parameters associated with alleles A_i and A_j. Then, the likelihood ratio test can be constructed to test for linkage and association. Conditional logistic models and their extensions have also been discussed by many other researchers [37–40]. Sinsheimer et al. [41] developed the gamete-competition model that is applicable to general pedigrees, and their method is identical to the Sham and Curtis model for case–parent trios.

Instead of modeling the probability that a particular marker allele is transmitted by a heterozygous parent, Weinberg et al. [42] developed a log-linear model for the probability of mating types. Their log likelihood function has the form of $L = \sum_{(O,M,F)} n_{O,M,F} \log\left(p_{(O,M,F)|D}\right)$, where (O, M, F) is the mating type of case–parent trios defined by Weinberg et al. [42], $n_{O,M,F}$ is the number of trios with the mating type (O, M, F) and $p_{(O,M,F)|D}$ is the conditional probability of mating type (O, M, F) given the disease status of offspring. Assuming mating symmetry, Weinberg et al. [42] defined six mating types and modeled $p_{(O,M,F)|D}$ as: $\log\left(p_{(O,M,F)|D}\right) = \gamma_{(O,M,F)} + \beta_O + \alpha_{M,F}$, where β and α are parameters associated with the number of copies of A_1 allele for offspring's and parental genotypes. The model can be easily modified to accommodate different disease models and parent-of-origin effects [43].

Schaid [32] proposed another conditional logistic regression approach, where the probability of offspring genotype is calculated conditional on the parental genotypes and offspring disease status as follows by assuming $P(D|g_o, g_m, g_f) = P(D|g_o)$:

$$P(g_o|g_m, g_f, D) = \frac{P(D|g_o)\, P(g_o|g_m, g_f)}{\sum_{g_o^* \in G} P(D|g_o^*)\, P(g_o^*|g_m, g_f)} = \frac{r(g_o)}{\sum_{g_o^* \in G} r(g_o^*)}.$$

In the above expression, D represents the affection status of the offspring, g_o^* one of the four possible genotypes of the offspring conditional on parental genotypes, and $r(g)$ is the relative risk of disease for genotype g, which consists of two alleles A_iA_j. As pointed out by Schaid [32] and Baksh et al. [44], other than $P(D|g_o)$, only $r(g)$ can be directly estimated from the log likelihood function. The logistic model presented by Schaid [32] provides a general framework to test the association and linkage, this model and its extensions and their similarities and differences between the score test and the TDT have been widely discussed and used (e.g., [39, 40, 45, 46]).

In general, $r(g)$ can be modeled as follows according to the disease model: $r(g) = X'\beta$, where X is the coded vector for the observed genotype g. For example, $r(g)$ can be written as $\log[r(g)] = \log[r(A_iA_j)] = \beta_i + \beta_j$ for the simple multiplicative model [32]. Clayton and Jones [47] proposed a more general multiplicative model with the following form: $h[r(g)] = h[r(A_i, A_j)] = \beta_i + \beta_j =$

$\frac{1}{2}\left\{h\left[r\left(A_i, A_i\right)\right]+h\left[r\left(A_j, A_j\right)\right]\right\}$, where h is an unspecified monotone increasing function. In all these aforementioned models, the null hypothesis of no association, i.e., $\beta = 0$, can be tested using the likelihood ratio test.

Other than the conditional likelihood proposed by Clayton and Jones [47] and Schaid [32], Lunetta et al. [48] used the prospective likelihood of the phenotype given offspring's genotype to test the association models for an arbitrary phenotype and a maker locus. Lunetta et al. [48] assumed the mean value of Y_{ij}, $\mu_{ij} = E\left(Y_{ij}\right)$ has a linear relationship with X_{ij}, the coded genotypes of offspring. Specifically, through a link function l used in the generalized linear model, we have $l_{ij} = l\left(\mu_{ij}\right) = \beta_0 + \beta_1 X_{ij}$. With the dichotomous phenotype, the natural link function is the logit function, $l_{ij} = l\left(\mu_{ij}\right) = \operatorname{logit}\left(\mu_{ij}\right) = \log\left[\mu_{ij}/\left(1-\mu_{ij}\right)\right] = \beta_0 + \beta_1 X_{ij}$. For a continuous phenotype with the normal distribution, the link function is $l_{ij} = l\left(\mu_{ij}\right) = \mu_{ij} = \beta_0 + \beta_1 X_{ij}$. Lunetta et al. [48] then computed the prospective likelihood of phenotype Y_{ij} conditioning on the genotype X_{ij} assuming that all the offspring are independent given their genotypes: $\log\left[L\left(\beta_0, \beta_1\right)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[Y_{ij}l_{ij} - a\left(l_{ij}\right)\right)]$, where $l_{ij} = l\left(\mu_{ij}\right) = \beta_0 + \beta_1 X_{ij}$ and $a\left(l_{ij}\right)$ is a function of l_{ij} with the property of $\partial a\left(l_{ij}\right)/\partial l_{ij} = \mu_{ij}$. Then, the score statistic U is $U = \sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{ij}\left(Y_{ij} - \mu\right)$, where μ is an nuisance parameter that only affects the power but not the validity of S. To adjust for the population stratification, Lunetta et al. [48] proposed to use appropriate permutation distributions for the offspring allele values and computed the distribution of Uas a function of offspring's genotypes, conditioning on parental genotypes and trait values for offspring and parents. For a bi-allelic marker with alleles A_1 and A_2, when only one affected offspring is used and X_{ij} is defined as the number of copies of allele A_1 for genotype g_{ij}, the score statistic proposed by Lunetta et al. [48] is $U = \sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{ij}\left(Y_{ij} - \mu\right) = t_A$, which is just the total number of A_1 alleles transmitted to the affected offspring.

There are at least two advantages for general likelihood-based methods proposed by Schaid [32] and Lunetta et al. [48]. First, these methods tend to be more powerful if the disease model is appropriately specified while the misspecification of the disease model generally does not affect the validity of tests. Second, these methods are easily extended to handle families with multiple affected and unaffected offspring, families with missing parental genotypes, and complex phenotypes. The inclusion of environmental covariates and the detection of interaction between genes and environmental covariates can also be incorporated into the model without much difficulty. We will introduce these extensions in the corresponding sections.

Other than methods based on allele counting and likelihood function, Rabinowitz and Laird [49] provided a general framework to test the linage and association between the maker and the DSL with arbitrary pedigree structure and arbitrary missing marker information based on the correlation of phenotypes and marker genotypes. The proposed family-based association test (FBAT) statistic, has a similar formula with the statistic proposed by Lunetta et al. [48]: $U = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu\right)\left(X_{ij} - E\left(X_{ij}|S_i\right)\right)$. The parameter μ is a pre-specified constant that depends on the nature of the trait. The choice of μ generally does not affect the validity but will affect the power of the test [49, 50]. The conditional expected value of

X_{ij}, $E\left(X_{ij}|S_i\right)$, is computed conditional on the sufficient statistic S_i [49] under the null hypothesis. For the FBAT statistic, the genotype X_{ij} is treated as random conditioning on the sufficient statistic S_i, but the trait Y_{ij} is treated as fixed. Under the null hypothesis of no linkage or no association, X_{ij} is centered around $E\left(X_{ij}|S_i\right)$, thus the FBAT statistic, U, has an expected value of 0. If the variance of FBAT statistic can be computed, $Z = \frac{U}{\sqrt{Var(U)}}$ or $Z^2 = \frac{U^2}{Var(U)}$, can be used as the test statistic. For large samples, Z is approximately distributed as the standard normal distribution and Z^2 is approximately distributed as the chi-square distribution with one degree of freedom. If it is difficult to compute the variance of U, an empirical variance can be estimated from the data [51–53]. The FBAT statistic provides a general framework to test the association for an arbitrary phenotype Y and maker loci. For case–parent trios genotyped at a bi-allelic locus, the sufficient statistic is the parental genotypes, Y_{ij} is equal to 1, μ is taken to 0, and X_{ij} is the number of copies in the offspring genotype for an allele. In this situation, Z^2 is the same as the TDT statistic proposed by Spielman et al. (1993). Actually, many extensions of TDT-based methods have similar formula as the FBAT statistic (e.g., [50,54]). For a multi-allelic maker with k alleles, U becomes a vector with k elements and the test statistic $U^T Cov\left(U\right) U$ has an approximate chi-square distribution with the degree of freedom equal to the rank of *Cov(U)*, which is the covariance matrix of U. In the FBAT statistic, it does not need to specify the disease model, thus it is robust for the misspecification of the distribution of Y_{ij}. It is also easy to extend the FBAT statistic to handle families with multiple offspring, complex phenotypes, and environmental covariates, which will be described in the corresponding sections.

4 Family with Multiple Siblings

When multiple affected and/or unaffected offspring are available within a nuclear family, each case–parent can be considered independently if there is no linkage between the maker locus and the DSL, because the transmission of alleles to an offspring is independent of the transmission of alleles to another offspring in the absence of linkage between the marker locus and the DSL. Therefore, the TDT and its extensions for multi-allelic markers are still valid for testing linkage in the presence of association. However, the TDT is not a valid test for testing association in the presence of linkage, because transmissions of an allele from a parent to the affected offspring are correlated. One strategy is to randomly choose one affected offspring from each family and then perform the TDT. However, this strategy sacrifices the available data and tends to be less powerful. Martin et al. [54] proposed tests that can use all affected offspring and multi-allelic markers. They also developed methods that can analyze families with an arbitrary number of affected offspring together. Here, we only outline their test based on the affected sib pairs and bi-allelic markers and point out its relationship with the TDT statistic. Denote t_j as the number of heterozygous parents with genotype A_1A_2 who transmit allele A_1 to j affected offspring, the test statistic for sib pairs is $T_{sp} = \frac{(t_0-t_2)^2}{t_0+t_2}$, which has an approximate chi-square distribution with one degree of freedom. Using their notations, the TDT statistic for affected sib pairs is $\mathrm{TDT} = \frac{(t_0-t_1)^2}{(t_0+t_1+t_2)/2}$. The two statistics, the TDT and T_{sp} have the following relationship: $\mathrm{TDT} = T_{sp}\frac{t_0+t_2}{(t_0+t_1+t_2)/2}$. Wicks [55] argued that the TDT is more powerful than T_{sp} to test for linkage because the TDT utilizes excess sharing, that is, the tendency for $t_0 + t_2$ to exceed t_1 in the presence of linkage. Based on this observation, Wicks [55] proposed a family of TDT-like statistics for affected pairs to test linkage in the absence of association: $TDT\,(\alpha) = \frac{(t_0-t_2)^2}{(1-\alpha)(t_0+t_2)+\alpha t_1}$. Under the null hypothesis of no linkage, $TDT\,(\alpha)$ has a chi-square destitution with one degree of freedom, and $TDT\,(\alpha = 1)$ is the most powerful test in this class. However, as we have pointed out, the TDT and $TDT\,(\alpha)$ may not be used to test association in the presence of linkage.

For a bi-allelic marker with two alleles A_1 and A_2, heterozygous parents who are more likely to transmit A_1 to affected offspring, are also more likely to transmit A_2 to unaffected offspring when there is linkage and association between the marker and the DSL. Therefore, unaffected offspring contain information about linkage and association, and can be included in the analysis to increase power. Guo et al. [56] developed a method, called the informative-transmission disequilibrium test (i-TDT), which can utilize transmission information from heterozygous parents to their affected offspring as well as the unaffected offspring from families with at least one affected offspring. As for TDT, only the heterozygous parents are included in the analysis. Denote $t_{i1}^m\left(t_{i1}^f\right)$, $t_{i2}^m\left(t_{i2}^f\right)$, $s_{i1}^m\left(s_{i1}^f\right)$, and $s_{i2}^m\left(s_{i2}^f\right)$ as the number of affected offspring who inherit A_1 but not A_2, the number of affected offspring who inherit A_2 but not A_1, the number of unaffected offspring who inherit A_1 but not A_2, and the number of unaffected offspring who inherit A_2 but not A_1 from a heterozygous mother (father) in the ith family, respectively. Define $d_i^m = (t_{i1}^m + s_{i2}^m) - (t_{i2}^m + s_{i1}^m)$ and $d_i^f = \left(t_{i1}^f + s_{i2}^f\right) - \left(t_{i2}^f + s_{i1}^f\right)$, then $d_i^m > 0$ (or $d_i^f > 0$) indicates the mother (or father) in the ith family are informative for allele A_1 and $d_i^m < 0$ (or $d_i^f < 0$) indicates the mother (or father) in the ith family are informative for allele A_2. The i-TDT statistic of Guo et al. [56] has the following formula: $\mathrm{i} - \mathrm{TDT} = \left[\sum_{i=1}^n \left(d_i^m + d_i^f\right)\right]^2 / \left[\sum_{i=1}^n (\ (d_i^m)^2 + \left(d_i^f\right)^2\right]$. i-TDT has an approximate chi-square distribution with one degree of freedom under the null hypothesis of no linkage or no association. When all families contain only one affected offspring, i-TDT is identical to the TDT of Spielman et al. [25]. Guo et al. [56] demonstrated that i-TDT can increase power when all offspring are included in the analysis through simulation studies.

As we have mentioned in Sect. 3.1, the conditional likelihood-based methods, the FBAT method and those methods similar to FBAT can be easily extended to accommodate families with multiple affected and unaffected offspring. Many methods have been developed under these frameworks. The FBAT statistic can be written as $U = \sum_{i=1}^n \sum_{j=1}^{n_i} (Y_{ij} - \mu)\,(X_{ij} - E\,(X_{ij}|S_i))$, where the sufficient statistic S_i is the parental genotypes, g_{im} and g_{if} for family i. Under the null of hypothesis of no linkage or no association, the expected value of U is equal to zero. Because the variance of U can be difficult to calculate in the presence of linkage, it was suggested to use its empirical variance to construct the test [52]. For each nuclear family i, $E\left(U_i\right) = 0$, thus $\sum_{i=1}^{n}\left[\left(Y_{ij} - \mu\right)\left(X_{ij} - E\left(X_{ij}|S_i\right)\right)\right]^2 = \sum_{i=1}^{n} U_i^2$ is an unbiased estimation of $Var\left(U\right)$. The test $Z = \frac{U}{\sqrt{Var(U)}}$ has an approximate standard normal distribution for large samples. If we let $Y_{ij} = 1$ to represent the affected offspring, $Y_{ij} = 0$ to represent the unaffected offspring, S_i and X_{ij} to be the number of copies of allele A_1 carried by this offspring, the statistic U is identical to the statistic derived by Lunetta et al. [48] based on the prospective likelihood and the generalized linear model: $U = \left(1 - \mu\right) t_A - \mu t_U$, where t_A is the total number of A_1 alleles transmitted to the affected offspring and t_U is the total number of A_1 alleles transmitted to the unaffected offspring. If μ is set to be 0, then only the affected offspring are used, and it is more powerful for the rare disease because most information is contained in the genotypes of affected individuals. On the other hand, including the unaffected offspring may increase the power when the disease is common. Many methods have the identical or similar formula with the statistic U. For example, if μ is set to the population prevalence of the disease, it is identical to the test statistics proposed by Whittaker and Lewis [57].

When there are multiple offspring, the conditional probability of offspring's genotypes given parental genotypes and offspring's traits for the ith family with n_i offspring is

$$\begin{aligned} &P\left(g_{i1}, \ldots, g_{in_i} \,|\, g_{im}, g_{if}, Y_{i1}, \ldots, Y_{in_i}\right) \\ &= \frac{P\left(Y_{i1}, \ldots, Y_{in_i} \,|\, g_{i1}, \ldots, g_{in_i}, g_{im} \cdot g_{if}\right) P\left(g_{i1}, \ldots, g_{in_i} \,|\, g_{im}, g_{if}\right)}{\sum_{g_{io}^*} P\left(Y_{i1}, \ldots, Y_{in_i} \,|\, g_{io}^*, g_{im}, g_{if}\right) P\left(g_{io}^* \,|\, g_{im}, g_{if}\right)}, \end{aligned}$$

where g_{io}^* is one possible genotypes of the offspring conditional on the parental genotypes. When there is no linkage between the marker and the DSL, the conditional probability of disease given offspring's genotypes does not depend on the parental genotypes, and is independent of their sibling's disease status and marker genotype. Therefore, the conditional probability $P\big(Y_{i1}, \ldots, Y_{in_i} \,|\, g_{i1}, \ldots, g_{in_i}, g_m, g_f\big)$ can be computed as $\prod_{j=1}^{n_i} P\left(Y_{ij} \,|\, g_{ij}\right)$. The likelihood function can be written as $L = \frac{\prod_{j=1}^{n_i} P(Y_{ij} \,|\, g_{ij})}{\sum_{g_{io}^*} \prod_{j=1}^{n_i} P\left(Y_{ij} \,|\, g_{ij}^*\right)} = \frac{\prod_{j=1}^{n_i} r(g_{ij})}{\sum_{g_{io}^*} \prod_{j=1}^{n_i} r\left(g_{ij}^*\right)}$, where $r\left(g\right)$ is the relative risk of disease for genotype g. In the presence of linkage, the joint probability of disease status of offspring, $P\left(Y_1, \ldots, Y_{n_i} \,|\, g_1, \ldots, g_{n_i}, g_m, g_f\right)$, will vary depending on the number of marker alleles shared identical by descent (IBD) among the sibling pairs. In this situation, Siegmund and Gauderman [58] proposed to calculate the conditional probability of the disease status given parental genotypes, offspring's genotypes, and IBD matrix among siblings. As a result, this conditional probability of the disease status of any offspring given its genotype is independent of their siblings' disease status and genotypes. Denote the IBD matrix among offspring as π, the likelihood function reduces to $P\left(g_{i1}, \ldots, g_{in_i} \,|\, g_{im}, g_{if}, Y_{i1}, \ldots, Y_{in_i}, \pi\right) = \frac{\prod_{j=1}^{n_i} r(g_{ij})}{\sum_{g_{io}^* \,|\, \pi} \prod_{j=1}^{n_i} r\left(g_{ij}^*\right)}$, where g_{io}^* is all possible genotypes of the offspring that are compatible with parental genotypes and IBD matrix for the ith family. Then, the likelihood ratio test and the score test can be derived. Martin et al. [59] used a similar method to calculate the conditional probability of reconstructed parental genotypes when the parental genotypes are missing. However, one drawback of the method proposed by Siegmund and Guaderman (2001) [58] is that the implementation is difficult using standard statistical packages such as SAS or R, due to the calculation of IBD matrix. Therefore, Cordell and Clayton [60] assumed that the conditional probability of disease given offspring's genotypes does not depend on the parental genotypes, and is independent of their sibling's disease status and marker genotypes, even in the presence of linkage. Then, the likelihood can be easily calculated as $L = \frac{\prod_{j=1}^{n_i} r(g_{ij})}{\sum_{g^*_{io}} \prod_{j=1}^{n_i} r\left(g^*_{ij}\right)}$. Instead of the likelihood ratio test, the Wald score test should be used [37, 60–62]. A robust "information-sandwich" estimate of the variance/covariance matrix (e.g., [63]) can be obtained to account for the correlation in offspring's disease status. Simulation results showed the Wald test can give the correct type I error rate [62]. Similarly, Zou [64] discussed how to use the retrospective logistic regression with the sandwich variance estimator to analyze family-based association studies.

Millstein et al. [65] developed another conditional logistic regression model for jointly testing linkage and association for families with two affected offspring. Their model contains two covariates, one is used to quantify association and the other is used to quantify linkage between the marker and the DSL. Specifically, the likelihood of the genotypes g_1 and g_2 for two affected offspring, conditioning on their phenotypes, Y_1 and Y_2 and their parental genotypes, g_m gand g_f, has the following formula: $P(g_1 \,|\, g_f, g_m, Y_1, Y_2) = P(g_1 \,|\, g_f, g_m, Y_1) P(g_2 \,|\, g_1, g_f, g_m, Y_1, Y_2)$. As usual, the first part can be modeled using a conditional logistic likelihood with parameter β(e.g., [32]) $P(g_1 \,|\, g_m, g_f, Y_1) = \exp(\beta g_1)/[\sum_{g_1^*} \exp(\beta g_1^*)]$, where g_1^* is one of four possible genotypes of offspring 1 conditional on the parental genotypes. Under some reasonable assumptions, the second part can be modeled as a conditional logistic likelihood with two parameters β and γ: $P(g_2 \,|\, g_1, g_m, g_f, Y_1, Y_2) = \exp(\beta g_2 + \gamma \pi_{12})/[\sum_{g_2^*} \exp(\beta g_2^* + \gamma \pi_{12}^*)]$, where g_2^* is one of four possible genotypes of offspring 1 conditional on the parental genotypes, π_{12} the number of alleles shared IBD between g_1 and g_2, and π_{12}^* is the number of alleles shared IBD between g_1 and g_2^*. Thus, this method can be easily implemented with a standard conditional logistic regression approach using available statistical packages (e.g., SAS or R). The simulations showed that their method can be more powerful than some standard tests for linkage and association.

5 Families with Missing Parental Genotypes

Unobservable parental genotypes present difficulties for the TDT and are indeed common in studying diseases that have a late onset age. In this section, we review methods that can analyze data from two different types of nuclear families: nuclear families with only one parent missing and nuclear families with both parents missing but with multiple siblings. For both types of families, although parental genotypes are unavailable, information about their genotypes may be contained in the genotypes of their offspring and available spouse. Therefore, parental genotypes may be constructed from their offspring's genotypes for some families. In the context of the TDT, we may treat families with reconstructed genotypes as they have been genotyped and only include these families in the analysis, as suggested by Speilman and Ewens (1996, 1999). However, Curtis, Curtis and Sham, Spielman and Ewens, and Knapp [66–69] noticed that this procedure can introduce bias and correcting such bias may require the knowledge of population frequency of marker alleles. Knapp [69] proposed a statistical procedure called RC-TDT (reconstruction combined TDT) to analyze four types of families together: (1) families with both parents genotyped; (2) families with one or two missing parental genotypes and missing parental genotypes can be constructed; and (3) families with missing parental genotypes that cannot be reconstructed but the condition for the sib TDT [70] is satisfied. Knapp [69] provided necessary and sufficient conditions for the observed genotypes in the offspring to allow for the reconstruction of parental genotypes and mating types and derived the appropriate mean and variance of the test statistic conditioning on parental mating types under the null hypothesis. Curtis [66] also presented similar conditions but used them for a slightly different purpose. The simulation studies showed that the RC-TDT has the correct type error rate and is more powerful than the sib TDT proposed by Spielman and Ewens [70] for the test of linkage [69, 71].

The TDT type of methods based on the allele counting have been developed for nuclear families with both parental genotypes missing and with genotypes of multiple siblings. These methods do not use the parental genotypes but require that families must contain at least one affected sibling and one unaffected sibling. For a sibling pair consisting of one affected sib and one unaffected sib, Curtis and Sham [29] proposed to compare each marker allele in the affected individual and in the unaffected sibling and use the following statistics for a bi-allelic marker: $Z_c = \left[t_{12} - \left(\frac{s_1}{2} + s_2\right)\right] / \sqrt{\frac{s_1}{4} + s_2}$, where $s_i\ (i = 1, 2)$ is the number of sibships that increase the test statistic of t_{12} or t_{21} by allele A_i and t_{ij} is increased by 1/2 only if maker allele A_i in the affected sibling and marker allele A_j in the unaffected sibling. Under the null hypothesis, Z_c has an approximate standard normal distribution. If there are multiple siblings within a family, Curtis and Sham [29] proposed to randomly select one affected offspring and then select one unaffected offspring whose marker genotype is maximally different from that of the affected offspring. This approach is unbiased although the procedure selects the most different unaffected sibling and the test statistic is therefore valid to test the association in the presence of linkage. Curtis [66] extended their method to handle multi-allelic makers using a likelihood model similar to that of Sham and Curtis [29], but its performance may be poor [72].

For multi-allelic markers, Boehnke and Langefeld [73] developed several family-based tests of association that can only use a pair of one affected sibling and one unaffected sibling. For a marker with k alleles, the data can be arranged in a $2 \times k$ contingency table in which the rows represent the disease status and the columns represent marker alleles. There are many ways to construct the contingency table, and here we outline one counting scheme that is most powerful. For an allele, it is counted only if it is not in both the affected sibling and the unaffected sibling. Then, the discordant sib pair (DSP) test statistic based from the table is $T_{DSP} = \sum_{j=1}^{k} \frac{(t_{1j}-t_{2j})^2}{t_{1j}+t_{2j}}$, where t_{1j} is the number of counted allele A_j in the affected siblings and t_{2j} is the number of counted allele A_j in the unaffected siblings. The permutation procedure that randomly permuted the affection statuses of the sibs was proposed to evaluate the significance level of T_{DSP}.

Spielman and Ewens [70] developed the sib TDT (S-TDT) method to analyze multiple affected and unaffected siblings if they satisfy two criteria: (1) there are at least one affected sibling and one unaffected sibling within each family and (2) the siblings must not have the same genotype. For each marker allele A_i, the S-TDT statistic is defined as $S - TDT_i = \frac{t_i - E(t_i)}{\sqrt{Var(T_i)}}$, where t_i is the number of allele A_i presented in the affected siblings from all families and $E\left(t_i\right)$ and $Var\left(t_i\right)$ are the mean and variance of t_i respectively. For a bi-allelic marker, the statistic $S - TDT_1$ can be used, whereas for a multi-allelic marker with k alleles, the test statistic $S - TDT_{\max} = \max\left|S - TDT_i\right|$ can be used. In addition, due to the calculation of $E\left(t_i\right)$ and $Var\left(t_i\right)$ discussed by Spielman and Ewens [70], the S-TDT statistic has an approximate standard normal distribution under the null hypothesis of no linkage. More importantly, the S-TDT statistic can be combined with the TDT to analyze different types of families [70]. Schaid and Rowland [74] noted that the S-TDT is equivalent to the conditional likelihood having log-additive effects of the marker alleles. The similarities and differences between the S-TDT and the Mantel –Haenszel test were discussed by Laird et al. [75] and Ewens and Spielman (1998).

For families with multiple affected and unaffected siblings, some methods discussed earlier can only use one pair of affected and unaffected sibling from each family while others are only valid to test linkage in the presence of association. To include all available siblings from the same family, Horvath and Laird [76] developed sibship disequilibrium test (SDT), for testing association in the presence of linkage. Their test procedure can be outlined as follows. For a multi-allelic marker with k alleles and a set of siblings, denote t_A and t_U as the number of affected siblings and unaffected siblings and define:

$t_{iA} = \big(\text{Total number of allele } A_i \text{ among the affected}\big) / \, t_A$ and
$t_{iU} = \left(\text{Total number of allele } A_i \text{ among the unaffected}\right) / \, t_U$.

Let $d_i = t_{iA} - t_{iU}$, b_i be the number of sibships for which $d_i > 0$ and c_i be the number if sibships for which $d_i < 0$. Then for a bi-allelic marker, the SDT statistic can be defined as $\mathrm{SDT} = \frac{(b_1-c_1)^2}{b_1+c_1}$. The extension of SDT to multi-allelic markers and the combined analysis of families with or without parental genotypes have been discussed in the literature (Curtis et al., 1999; [76]). Guo et al. [56] extended their i-TDT method to handle families with multiple affected and unaffected offspring in the presence of missing parental genotypes.

To make use of case–parent trios with only one available parent, Sun et al. [50] proposed two unbiased test statistics for linkage and association based on an accurate method of estimating the risk ratio for case–parental control design studies [77]. Suppose there are two alleles A_1 and A_2 at the marker locus. Let t_{ij} $(i, j = 0, 1, 2)$ be the number of affected offspring whose genotype has i copies of allele A_1 and whose one available parent has j copies of allele A_1, then the first test statistic, called 1-TDT [50] is $1 - \mathrm{TDT} = \frac{(t_{01}+t_{12}) - (t_{10}+t_{21})}{\sqrt{(t_{01}+t_{12}) + (t_{10}+t_{21})}}$. From 1-TDT, we can see that it only uses the offspring–parent with genotypes (A_1A_2, A_1A_1), (A_1A_2, A_2A_2), (A_1A_1, A_1A_2), and (A_2A_2, A_1A_2). In other words, one of genotypes has to be heterozygous and the other is homozygous. The 1-TDT has an approximate standard normal distribution if either of the two following assumptions holds: (1) males and females with the same genotype have the same mating preference and (2) father and mother are missing with the same probability given that one of them is missing. Sample sizes required to detect the association for the S-TDT and 1-TDT were investigated by Wang and Sun [78]. Under a variety of genetic models, the sample size needed for the 1-TDT is roughly the same as that needed for the S-TDT with one affected and one unaffected sibs, and is about twice of that needed for the TDT.

Sun et al. [50] also proposed a second test statistic which is valid even when both assumptions fail. The same approach has been extended to analyze quantitative traits and families with multiple siblings [79]. Denote the coded genotype $X_{ij} = 1$ if the offspring–parent genotype is (A_1A_2, A_1A_1) or (A_2A_2, A_1A_2) and $X_{ij} = -1$ if the offspring–parent genotype is (A_1A_2, A_2A_2) or (A_1A_1, A_1A_2), then the test statistic is $1 - \mathrm{TDT} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n_i}(Y_{ij}-\mu)X_{ij}}{\sum_{i=1}^{n}\left[\sum_{k=1}^{n_i}(Y_{ij}-\mu)X_{ij}\right]^2}$, where μ is a constant that can affect its power but not its validity. This test statistic is valid for testing association in the presence of linkage and has an approximate standard normal distribution if either of the above two assumptions holds.

In additional to the nonparametric tests based on allele counting, likelihood-based methods have been developed to analyze families with missing parental genotypes. For nuclear family i with n_i offspring, the conditional probability of parental genotypes and offspring's genotypes, g_{if}, g_{im}, and $g_{i0} = (g_{i1}, \ldots, g_{in_i})$, given the phenotypes of offspring, $Y_{io} = (Y_{i1}, \ldots, Y_{in_i})$, can be written as

$$L_i = P\left(g_{im}, g_{if}, g_{io} | Y_{io}\right) = P\left(g_{im}, g_{if} | Y_{io}\right) P\left(g_{io} | g_{im}, g_{if}, Y_{io}\right) = L_i^{(P)} L_i^{(O)}.$$

When both parental genotypes are available, only $L_i^{(O)} = P\left(g_{io} | g_{im}, g_{if}, Y_{io}\right)$ are needed and tests based $L^{(O)}$ are robust to the population stratification [32, 36]. When parental genotypes are missing, $L_i^{(P)}$ should be used, depending on the missing genotypes. In this situation, the likelihood function is summed over all possible genotypes of parents that are compatible with the observed parental and offspring's genotypes and it has the following form:

$$L_i = \sum\nolimits_{g_{im}, g_{if}} P\left(g_{im}, g_{if}, g_{io} | Y_{io}\right) = \sum\nolimits_{g_{im}, g_{if}} P\left(g_{im}, g_{if} | Y_{io}\right) \times P\left(g_{io} | g_{im}, g_{if}, Y_{io}\right).$$

Many methods have been proposed using the full likelihood function, L_i, to analyze families with missing parental genotypes. One should be cautious when applying these methods. Because the calculation of $L_i^{(P)} = P\left(g_{im}, g_{if}|Y_{io}\right)$ generally depends on population models and assumptions such as Hardy–Weinberg equilibrium, the proposed tests may lead to biased results in the presence of population stratification.

Schaid and Li [80] extended the likelihood-based method of Schaid [32] to allow for missing parental genotypes and multiple siblings. The likelihood function for family i is

$$L = \sum\nolimits_{g_{im}, g_{if}} I\left(g_{im}, g_{if}\right)$$
$$\frac{P\left(g_{im}, g_{if}\right) P\left(Y_{io}|g_{io}, g_{im}, g_{if}\right) P\left(g_{io}|g_{im}, g_{if}\right)}{\sum\nolimits_{g_{im}, g_{if}} I\left(g_{im}, g_{if}\right) P\left(g_{im}, g_{if}\right) \sum\nolimits_{g_{io}^*} P\left(Y_{io}|g_{io}^*, g_{im}, g_{if}\right) P\left(g_{io}^*|g_{im}, g_{if}\right)},$$

where $I\left(g_{im}, g_{if}\right)$ is an indicator function and takes the values of 1 if the parental genotypes g_{im} and g_{if} are compatible with observed parental and offspring's genotypes and 0 otherwise. Schaid and Li [80] proposed to use the Expectation-Maximization (EM) algorithm to evaluate the likelihood function and estimate genotype relative risks. Since they assumed random mating and Hardy–Weinberg equilibrium when evaluating $P\left(g_{im}, g_{if}\right)$, this method is biased for stratified populations. In addition, they assumed the conditional probability of the trait of any offspring given its genotype is independent of their siblings' traits and genotypes. Thus, the proposed test is only valid for testing linkage in the presence of association. Such method and its extensions that can handle missing parental genotypes have also been discussed by Cordell et al. [37].

Martin et al. [81] developed a parental genotype reconstruction (PGR) method to analyze families with missing parental genotypes. In the first step of PGR, the posterior probability of each possible genotypes of parents is calculated based on the observed parental and offspring's genotypes with the assumption of linkage disequilibrium. Specifically, the following posterior probability was evaluated: $P\left(g_{im}^*, g_{if}^*|g_{io}, g_{im}, g_{if}\right)$, where g_{im}^* and g_{if}^* are hypothetical genotypes of parents and $\left(g_{io}, g_{im}, g_{if}\right)$ are observed parental and offspring's genotypes. Since this posterior probability is not conditioning on the traits of offspring, the calculation can be greatly simplified. In the second step of PGR, the possible configurations of parental genotypes and their posterior probabilities can be used to analyze the genotype or haplotype relative risk. Again, this method assumed random mating and Hardy–Weinberg equilibrium and may be biased for stratified populations.

Weinberg [82] generalized the log-linear model developed in their earlier work [42] to analyze families with missing parental genotypes. Instead of modeling the posterior probability of parental genotypes, Weinberg [82] proposed to estimate the posterior probability of parental mating types using the EM algorithm. Denote $(O, M, F)_i$ as the mating type of family i, the likelihood function for this family is summed over all possible mating types that are compatible with the observed parental and offspring genotypes: $L_i = \sum_{(O,M,F)_i} p\left[(O,M,F)_i\right] p_{(O,M,F)_i|D}$, where $p\left[(O,M,F)_i\right]$ is the posterior probability of mating type to be estimated by the EM algorithm and $p_{(O,M,F)|D}$ is the conditional probability of mating type given the disease status of offspring. As described in Weinberg et al. [42], $\log\left(p_{(O,M,F)}\right) = \gamma_{(O,M,F)} + \beta_O + \alpha_{M,F}$, where β and α are parameters associated with the number of copies of A_1 alleles for offspring's and parental genotypes. Since the model does not assume random mating and Hardy–Weinberg equilibrium, it is robust to population stratification.

Whittemore and Tu [83] generalized the likelihood-based score statistic of Schaid (1996) and Schaid and Li [80] to combine data from case–control and family-based designs. The score statistic has two components: the nonfounder statistic that evaluates disequilibrium in the transmission of marker alleles from parents to offspring and the founder statistic that compares the observed or inferred founder genotypes with those of controls or those of some reference population. This method is further generalized by Shih and Whittemore [84] to (1) accommodate other types of phenotypes, such as censored times to failure and quantitative traits; (2) account for within family correlation in phenotypes; and (3) allow for missing parental genotypes. Although their method can analyze data with families with arbitrary structures and arbitrary patterns of missing data, the inappropriate specifications on the distribution of founder genotypes can lower the power and introduce bias for stratified populations.

Instead of using the conditional likelihood proposed by Schaid and Clayton [32, 36], Siegmund et al. [62] proposed to use the conditional probability of traits given the siblings' genotypes to test the association in the presence of linkage. The conditional likelihood function for n sibships is $L = \prod_{i=1}^{n} \frac{\prod_{j=1}^{n_i} r(g_{ij})}{\sum_{g_o^*} \prod_{j=1}^{n_i} r\left(g_{ij}^*\right)}$, where $r\left(g_{ij}\right)$ is the conditional probability of trait given its genotype and g_o^* is all possible genotypes of the offspring. It can be seen that they assumed that the conditional probability of traits given offspring's genotypes does not depend on the parental genotypes, and is independent of their siblings' trait and genotypes. Thus, this method can use all siblings and does not require the specification of exact correlation between siblings. The test can be easily performed using available statistical packages such as SAS or R, and a robust variance estimate can be used to compute a Wald test that is still valid for testing association in the presence of linkage. This model has been further discussed by Cordell et al. [37] for more complex family structures. Similarly, Jonasdottir et al. [85] proposed a mixed effects model to test association in the presence of linkage for families with missing parental genotypes. In their model, the population level association is modeled using a fixed effect and the correlation of traits among siblings is characterized using log-gamma random effects. Therefore, their method is still valid in the presence of linkage.

Although statistical methods from traditional statistical packages such as SAS or R that can analyze family-based association studies have been proposed (e.g., Cordell and Clayton (2002); [37]), it is difficult for these packages to handle missing data. Therefore, Croiseau et al. [86] proposed a multiple imputation method as a solution. Their procedure can be outlined as follows. In the first step, the complete phase-known data are generated based current parameter values for population haplotype frequencies and genotype frequencies using a multiple imputation via data augmentation. The initial haplotype frequencies can be estimated from the EM algorithm. In the second step, the population haplotype frequencies are updated by sampling them from their posterior distribution given the current complete data. These two steps are repeated for a large number of times to reach a stationary distribution. In the last step, m data sets at intervals (e.g., every 1,000 repeats) are selected to perform the analysis. Croiseau et al. [86] showed that their method is more powerful than the method that only uses the families without missing data and is robust in the presence of population stratification with moderate amount of missing data (missing rate $\leq 30\%$).

The aforementioned likelihood-based methods that can handle missing parental genotypes are valid only when the parental genotypes are missing at random, i.e., the probability of having missing parental genotypes does not depend on the phenotypes of their offspring, their genotypes, and genotypes of their offspring. Allen et al. [87] illustrated that parental missingness can depend on parental genotypes in some situations, i.e., the missingness is informative. Even with a slightly informative missingness for parental genotypes, the methods based on missing at random can perform very poorly [87]. To account for the informative missingness of parental genotypes, Allen et al. [88] proposed a testing procedure based on the conditional likelihood of observed parental and offspring genotypes given the offspring's phenotypes and parental missing patterns. Specifically, the conditional likelihood function for family i is $L_i = P\left(g_{io}, g_{im,}g_{if}|Y_i, R_i\right) = P\left(g_{io}|g_{im,}g_{if}, Y_i\right) P\left(g_{im,}g_{if}|Y_i, R_i\right)$ under the assumption of that the parental missingness is independent of the offspring's genotypes given the parental genotypes and the offspring's phenotype across all populations, where R_i is the parental missing pattern. For example, we can use $R_i = (1, 0)$ to represent the case where only father's genotype is missing whereas $R_i = (0, 0)$ for the case when both parental genotypes are missing. In general, the term $P\left(g_{im,}g_{if}|Y_i, R_i\right)$ involves models on parental missing pattern R and on parental mating types. However, correct specification of missingness model is not so straightforward and misspecification of the missingness model can lead to bias for stratified populations. Chen [88] proposed new family-based association test that are robust for stratified populations in the presence of informative missing. Chen [88] used the conditional probability of offspring's genotypes given the observed parental genotypes, parental missing pattern, and offspring's phenotypes. For complete data, this method is identical to the conditional likelihood of Schaid and Sommer [89]. For families with missing parental genotypes, this method depends on parental missing patterns and mating types. However, Chen [88] treated them as nuisance parameters and did not require to directly modeling them. Therefore, such method is easier to be implemented and more robust.

Sebastiani et al. [90] proposed a nonparametric method, the robust TDT (rTDT) that does not assume any missing pattern, to handle missing parental genotypes in case–parent designs. In rTDT, the possible genotypes of missing parents are first constructed based on genotypes of offspring and available parents. Then, the usual TDT statistics are obtained on each of possible genotypes of missing parents. In the final step, the minimum TDT statistic is used to test the association and linkage between the marker and the DSL. Sebastiani et al. [90] derived an efficient algorithm to calculate rTDT, since the simple enumeration calculation over all combinations of possible genotypes over all missing parents would be very time consuming. Their simulation results showed that rTDT can achieve higher power and greater significance than the popular TDT method in some situations.

6 Quantitative Phenotypes

We have so far focused on the analysis of qualitative phenotypes, especially binary phenotypes. However, many phenotypes are measured quantitatively and quantitative phenotypes generally contain more information than qualitative phenotypes. For family-based designs with binary phenotypes, families having at least one affected offspring are generally collected. For quantitative phenotypes, although collecting families with extreme values of quantitative phenotypes can increase statistical power, random families are commonly recruited in genetic studies. Therefore, analysis methods that handle quantitative phenotypes have similarities and differences to those that analyze qualitative phenotypes.

Many approaches have been developed in the last several years to analyze quantitative phenotypes using family-based designs. Allison [91] developed five statistics to analyze quantitative traits using family-based designs, and TDT_{Q5} was found to be the most flexible and most powerful under a variety of genetic models in his simulation studies. Here, we describe TDT_{Q5} as follows. In TDT_{Q5}, the quantitative phenotype is regressed on offspring genotypes while controlling for parental mating types, which is determined by parental genotypes. For a bi-allelic maker with alleles A_1 and A_2, there are only three informative mating types ($A_1A_1 \times A_1A_2$, $A_1A_2 \times A_1A_2$, and $A_1A_2 \times A_2A_2$). Then they are coded as two dummy variables and entered into the regression model. A F-test with two degrees of freedom can be used to assess the significance if the marker locus is in linkage and association with the DSL. Therefore, TDT_{Q5} corresponds to the ordinary regression analysis. Thus, the analysis using TDT_{Q5} can be carried out using commonly available and well-tested statistical software such SAS or R. In addition, it can be easily generalized to analyze data with multi-allelic loci and families with multiple siblings. Allison and Neale [92] discussed how TDT_{Q5} can be generalized to more complex situations. For a multi-allelic locus, more dummy variables can be used to code the parental mating types and entered into the regression model. If one can use additional dummy variables to indicate if the father is heterozygous or the mother is heterozygous, TDT_{Q5} can be generalized to test imprinting effects. For families with multiple siblings, the weighted generalized least-square regression or mixed model can be used to estimate the residual correlation among siblings [93]. To analyze quantitative phenotypes using families with missing parental genotypes, Allison et al. [94] developed a mixed model coupling with the permutation procedure to test the null hypothesis of no linkage between the marker and the trait locus.

The mixed model has the following form: $Y_{ij} = \mu + \alpha_{g_{ij}} g_{ij} + \beta_i + (\alpha\beta)_{g_{ij}i} + \varepsilon_{g_{ij}i}$, where α_g is the fixed effect for genotypes, β_i is the random effect to model the correlation between the sibship i, and the interaction effect of α and β is modeled as the random effect. The mixed effects model allows the straightforward inclusion of covariates and other genes. Based on the mixed model, Allison et al. [93] proposed a permutation procedure to test the null hypothesis of no linkage. Simulation results showed that the permutation procedure generally has greater power and, furthermore, it has the advantage of being distribution free.

Since it is relatively easy to carry out the regression analysis, many regression-based methods have been developed to analyze quantitative phenotypes using family based designs after the development of TDT_{Q5}. In George et al. [94], the phenotype, Y, was assumed to be continuous and as the dependent variable. The transmission status of the associated allele, X, was considered as the primary independent variable. Other covariates C were considered independent variables and incorporated into the regression model. The correlations among family members were also incorporated. For an individual j in the ith family, George et al. [94] defined their regression model as $Y_{ij} = \beta_0 + \beta_c C_{ij} + \beta_X X_{ij} + \varepsilon_{ij}$, where X_{ij} takes the value of 1 if the allele A_1 is transmitted from a heterozygous parent and 0 otherwise. George et al. [94] showed that testing the null hypothesis $\beta_X = 0$ is equivalent to test the null hypothesis of no linkage or no association. Zhu and Elston [95] proposed an alternative regression model in which the transmission status was defined in a different way. Simulation studies have shown that a variant of Zhu and Elston's method is more powerful in most situations [96]. Zhu et al. [97] further extended this parametric method to accommodate data with missing parental genotypes. With the assumption of a random sample of individuals, Yang et al. [98] developed a similar regression model with additional regressors.

Many researchers have proposed generalized linear models to analyze qualitative and quantitative phenotypes (Cordell and Clayton (2002); [44, 99]). One of the advantages of generalized linear models is that the environmental covariates and gene–environment interactions can be easily incorporated into the analysis. In the generalized linear model, it is assumed that the transformation of the mean value of Y_{ij}, $\mu_{ij} = E(Y_{ij})$ has a linear relationship with X_{ij}, which is coded genotype of g_{ij} conditional on the parental genotypes g_{im} and g_{if}. Specifically, through a link function l used in the generalized linear model, we have $l_{ij} = l(\mu_{ij}) = \beta_0 + \beta_1 X_{ij}$. Under the assumption of normality, the conditional probability of offspring's genotypes, $g_{io} = (g_{i1}, \ldots, g_{in_i})$, given offspring's phenotypes, $Y_{io} = (Y_{i1}, \ldots, Y_{in_i})$, and parental genotypes, (g_{im}, g_{if}), for family i is $L_i = P(g_{io}|g_{im}, g_{if}, Y_{io}) = \frac{P(Y_{io}|g_{io}, g_{im} \cdot g_{if}) P(g_{i1}, \ldots, g_{in_i}|g_{im}, g_{if})}{\sum_{g^*_{io}} P(Y_1, \ldots, Y_n|g^*_{io}, g_{im}, g_{if}) P(g^*_{io}|g_{im}, g_{if})}$. Liu et al. [99] proposed a score test for the association between the marker locus and the trait locus. To construct the score statistic, some unknown parameters must be specified or estimated under the null hypothesis. For quantitative phenotypes, it involves the estimation of parameter β_0, which cannot be estimated from the likelihood function under the null hypothesis of no association. Therefore, Liu et al. [99] suggested to use the mean of phenotypes as an estimate of β_0 and further adjusted its bias in the ascertainment by the addition of an offset term that is assumed to be known. Although this procedure will not affect the validity of the score test, it may reduce the power of the test. Baksh et al. [44] modeled the joint probability of offspring genotypes and phenotypes conditioning on the parental genotypes and the ascertainment in family-based designs. In this situation, the nuisance parameters (e.g., β_0) can be estimated directly from the likelihood function and a likelihood ratio test statistic can be constructed.

Another group of methods to analyze quantitative phenotypes is based on the likelihood framework. Clayton and Jones [47] assumed that the trait, Y_{ij}, for a given individual has a normal distribution conditional on his/her genotypes, g_{ij}, i.e., $Y_{ij} \sim n\left(\mu_{g_{ij}}, \sigma^2\right)$. Then, the conditional probability of genotype g_{ij} given the trait value and the parental genotypes is $P\left(g_{ij}|Y_{ij}, g_m, g_{if}\right) = \frac{\phi\left(\left(Y_{ij}-\mu_{g_{ij}}\right)/\sigma\right)}{\sum_g \phi((Y_{ij}-\mu_g)/\sigma)}$, where ϕ is the probability density function of standard normal distribution and the sum in the denominator is over all possible transmissions from the parents to the offspring. If we re-parameterize the mean value $\mu_{g=A_iA_j}$ for the genotype A_iA_j as $h\left(\mu_{g=A_iA_j}\right) = h\left(\mu_0\right) + \beta_i + \beta_j$, then a score statistic can be used to test the null hypothesis of $\beta = 0$, which is equivalent to test the null hypothesis of no linkage or no association between the marker locus and the trait locus.

Kistner and Weinberg [100] extended the log-linear model of Weinburg et al. (1998) for quantitative phenotypes. Denote (O, M, F) as the mating type of case–parent trios, Kistner and Weinberg [100] modeled the probability of O conditioning on M, F, and the phenotype, Y, based on a multinomial distribution. Specifically, $p_{(O|M,F,Y)} = \exp\left(\beta_C Y + \alpha_{OMF}\right)$. The null hypothesis of no linkage or no association between the DSL and the phenotype can be tested by setting $\beta_O = 0$ for all O. Kistner and Weinberg [101] further extended this model to allow for missing parental genotypes. If we allow different β_O for different mating types, the model can be used to test parent-of-origin effects.

To allow for a simultaneous test of allelic association and linkage using only siblings, Fulker et al. [102] generalized variance components models in quantitative trait (QTL) mapping and Cardon [103] developed a regression based method which is an extension of Fulker's method. Fulker's method has further been generalized to handle families with an arbitrary number of siblings, general pedigrees, and genome wide association studies [104–106]. In the variance components model, the association is partitioned into between- and within- siblings components, and a robust test is constructed only on the basis of the within siblings component. Specifically, Abecasis et al. [104] assumed that the mean of Y_{ij} satisfies: $E\left(Y_{ij}\right) = \mu + \beta_b b_i + \beta_w w_{ij}$, where β_b and β_w are between and within siblings effects and b_i and w_{ij} are orthogonal between- and within-family components of genotype g_{ij}. For a marker locus with two alleles A_1 and A_2, we define g_{ij} as the number copies of allele A_1 minus one, then $b_i = \left(\sum_j g_{ij}\right)/n_i$ if parental genotypes are unknown and $b_i = \left(g_{im} + g_{if}\right)/2$ if parental genotypes are available and $w_{ij} = g_{ij} - b_i$. Thus, b_i is the expectation of each g_{ij} conditional on family data and w_{ij} is the deviation from this expectation for offspring j. For each family, the $n_i \times n_i$ covariance matrix, Ω_i, has the elements: $\Omega_{ijk} = \sigma_a^2 + \sigma_g^2 + \sigma_e^2$ for $j = k$ and $\Omega_{ijk} = \pi_{ijk}\sigma_a^2 + 2\varphi_{ijk}\sigma_g^2$ for $j \neq k$, where σ_a^2 is the additive genetic variance of the QTL, σ_g^2 is the variance attributable to polygenes, σ_e^2 is the residual environmental variance, and π_{ijk} is the proportion of alleles shared IBD at the marker locus between individuals j and k in family i. Under the assumption of normality, the likelihood of the data for family i is $L_i = (2\pi)^{-n_i/2} |\Omega_i|^{-1/2} \exp\left[-\frac{1}{2}(Y_i - \mu_i)' |\Omega_i|^{-1/2} (Y_i - \mu_i)\right]$. Abecasis et al. [104] showed that the test of null hypothesis of $\beta_w = 0$ is equivalent to test the association between the maker locus and the trait locus. In this situation, the log likelihood ratio test statistic, $2\log\left(\left(\prod_i L_i\right) / \left(\prod_i L_i(\beta_w = 0)\right)\right)$ has an asymptotic chi-square distribution with one degree of freedom.

Purcell et al. [107] generalized the methods of Fulker et al. [102] and Abecasis et al. (2000a) to incorporate parental phenotypes. For quantitative phenotypes and offspring's phenotypes, the mean of Y_{ij} satisfies: $E(Y_{ij}) = \mu + \beta_b b_i + \beta_w w_{ij}$, which is exactly the same as the model of Abecasis et al. [104] . For the parental phenotypes, the mean of Y_{im} and Y_{if} satisfies: $E(Y_{im}) = \mu_p + \beta_{pb} b_i + \beta_{pw} w_{im}$ and $E(Y_{if}) = \mu_p + \beta_{pb} b_i + \beta_{pw} w_{if}$ where $w_{im} = g_{im} - b_i$ (or $w_{if} = g_{if} - b_i$) is the deviation from the expectation of offspring's genotypes for the parents. For each family, the $(n_i + 2) \times (n_i + 2)$ covariance matrix, Ω_i, can also be constructed. Therefore, up to four main parameters, $(\beta_b, \beta_w, \beta_{pb}, \beta_{pw})$ with other nuisance parameters define the association between the marker and the phenotype and the corresponding likelihood ratio tests can be constructed. Simulation studies have shown that the incorporation of parental phenotypes can be considerably more powerful than equivalent quantitative tests that do not use parental phenotypes.

For analysis methods of quantitative based on generalized linear models and likelihood function, it is generally assumed that the quantitative phenotype of interest has a normal distribution. A departure from normality can inflate their type I error rates [108]. A general solution is to transform the trait values using methods such as Box–Cox transformation [109]. However, it is difficult to identity an appropriate transformation and different transformations can generate conflicting results. Inappropriate transformation can also affect the type I error rates and power. Diao and Lin [108] extended previous analysis methods for quantitative traits such as QTDT [104] to allow for a completely unspecified transformation function for the trait values. Their generalized model of Abecasis et al. [104] assumes the mean of transformed trait value satisfies: $E(T(Y_{ij})) = \mu + \beta_b b_i + \beta_w w_{ij}$, where β_b and β_w are between- and within-siblings effects and b_i and w_{ij} are orthogonal between- and within-family components of genotype g_{ij}. Based on this assumption, Diao and Lin [108] constructed a nonparametric likelihood function and proposed a method to estimate the transformation function T by assuming that T is step-wise function from the data. Since the estimates of parameters are based on the rank of Y_{ij}, their method is robust to outliers. Their simulation results showed that their method had the appropriate type I error rate and was more powerful than the existing methods.

Other than the parametric methods based on generalized regression and likelihood function, nonparametric methods have been proposed to analyze quantitative using family data. One advantage of such methods is that no assumption is made about the distribution of phenotypes. The tests are therefore valid for any type of sampling schemes based on the phenotypes of the individuals. Xiong et al. [110] compared the average trait values of offspring inheriting one allele versus the other to test the linkage for the marker locus and the trait locus. For a maker locus with k alleles $A_1, \ldots, A_k$, denote $\bar{Y}_{i.}$ as the average trait value of offspring inheriting allele A_i from the parental genotype containing the allele A_i, $\bar{Y}_{.i}$ as the average trait value of offspring not inheriting allele A_i from the parental genotype containing the allele A_i, and V_i^2 as the estimated variance of $\bar{Y}_{i.} - \bar{Y}_{.i}$. Under the null hypothesis of no linkage or no association, $\bar{Y}_{i.} - \bar{Y}_{.i}$ has the expected value of 0. Therefore, Xiong et al. [110] proposed the statistic $TDT_Q = \frac{k-1}{k} \sum_{i=1}^{k} \frac{\left(\bar{Y}_{i.} - \bar{Y}_{.i}\right)^2}{V_i^2}$ and TDT_Q asymptotically follows a chi-square distribution with $k-1$ degrees of freedom under the null hypothesis of no linkage or no association. It is worth emphasizing that TDT_Q is not a valid test for association in the presence of linkage due to the way that V_i^2 was estimated [110]. This method has been further generalized by Fan et al. [111] for multi-allelic markers.

Rabinowitz [112] proposed to assess the correlation between the offspring's phenotypes and genotypes conditioning on the parental genotypes. The method has been generalized to test the association with arbitrary pedigree structure and arbitrary missing marker information and implemented in the software package FBAT [49]. As we have mentioned, the test statistic of Rabinowitz [112] can be expressed as $Z = U/\sqrt{Var\,(U)}$, where $U = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(Y_{ij} - \mu\right)\left(X_{ij} - E\left(X_{ij}|S_i\right)\right)$ and Z has an asymptotic standard normal distribution under the null hypothesis of no linkage or no association. The constant μ affects the power of Z but not its validity. The variance of U, can be conveniently estimated by $Var\,(U) = \sum_{i=1}^{n} \left[\sum_{j=1}^{n_i} \left(Y_{ij} - \mu\right)\left(X_{ij} - E\left(X_{ij}|S_i\right)\right)\right]^2$. Such approach has also been generalized to include families with missing parental information by Sun et al. [79] and Monks and Kaplan [113]. To use information from all available offspring, Monks and Kaplan [113] proposed three statistical tests for quantitative traits: T_{QP}, T_{QS}, and T_{QPS}, where T_{QP} requires the parental genotypes and is identical to the test proposed by Rabinowitz [112], T_{QS} is based on siblings, and T_{QPS} is a combination of T_{QP} and T_{QS}.

7 Joint Analysis of Multiple Markers

In association studies, many tightly linked markers other than a single marker are generally genotyped. One strategy is to analyze each marker separately and then adjust for multiple comparisons by the Bonferroni correction. Such analysis ignores the linkage disequilibrium among markers and tends to be conservative. Methods based multiple tightly linked markers may provide more power than methods based on single markers because the former exploits LD information with the DSL from multiple markers. The methods using multiple markers can be generally classified into two groups: those based on unphased genotypes and those based on known or inferred haplotypes. Both types of methods have shown to be more powerful than single marker-based methods in simulation and empirical studies [114–119]. As we can see from the formula of the TDT statistic, only the parents with heterozygous genotypes will contribute to the TDT statistic. These parents are called informative parents. Thus, haplotypes across several markers have an addition advantage over single markers for family based association studies: haplotypes can increase the heterozygosity of parents and provide more informative parents in most situations. In addition, haplotype-based methods can be more powerful when multiple disease-susceptibility alleles occur within a single gene [120] and can potentially capture cis-interactions between two or more causal variants. However, it is also worth noting that the methods based on multiple markers may not always be more powerful than single marker-based methods, and the haplotype-based methods are not always more powerful than the methods based on genotypes at either a single marker locus or multiple marker loci [115–117]. Actually, relative efficiencies of these methods depend on many factors and it is still not clear which method is optimal in many situations.

Liang et al. [121] developed a multipoint, parametric approach for gene mapping using case–parent trios. Their test is based on an expression of expected preferential-allele-transmission statistics for transmission. Suppose L bi-allelic markers are genotyped at $t_1, t_2, \ldots, t_L$ and the position of DSL is at τ along the chromosome. For the mother in the ith family, we define its preferential-allele-transmission at the locus $l\,(l = 1, \ldots, L)\,M_{il}$, as 1 if she transmits allele A_1 but not A_2 to the affected offspring, -1 if she transmits allele A_2 but not A_1, and 0 otherwise. Similarly, the preferential-allele-transmission for the father, F_{il}, can also be defined. Under the null hypothesis of no linkage or no association between the marker and the DSL, M_{il} and F_{il} have the expected value of 0 conditioning on the disease status of offspring. Liang et al. [121] derived that the expected M_{il} and F_{il} conditional on the disease status of offspring, which is a function of marker position, t_l, the DSL location, τ, and other parameters: $E\left(M_{il}\right) = E\left(F_{il}\right) = \mu\left(t_l; \tau, C, N, \pi_l\right)$. Given observed t_l, M_{il}, and $F_{il}\;(i = 1, \ldots, n; l = 1, \ldots, L)$, Liang et al. [121] proposed to use the generalized-estimating-equation (GEE) [122] approach to jointly analyze M_{il} and F_{il} and proposed a chi-square test with one degree of freedom to estimate the DSL location, τ. Comparing with the TDT, this method has two advantages. First, families with homozygous parents are included in the analysis, which may increase the power. Second, a chi-square test with one degree of freedom was proposed to avoid the correction of multiple testing problems for multiple markers. Hsu et al. [123] further generalized this method for quantitative phenotypes.

Fan and Xiong [124] extended the regression model for QTL mapping to jointly test linkage and association using multiple markers. Denote the coded genotypes of individual j in the ith family at marker A and B as $X_{ij,A}$, $Z_{ij,A}$, $X_{ij,B}$, and $Z_{ij,B}$, respectively. Fan and Xiong [124] proposed the following regression model: $Y_{ij} = \beta + w_{ij}\gamma + X_{ij,A}\alpha_A + X_{ij,B}\alpha_B + Z_{ij,A}\beta_A + Z_{ij,B}\beta_B + e_i$, where w_{ij} are other covariates. For each family, the covariance matrix is same as that defined by Abecasis et al. [104]. The advantages of this method include the joint analysis of two tightly markers, the combined analysis of population and family data, and the use of parental phenotypes. The method has been further extended to analyze sibship and general pedigrees [125, 126]. However, the construction of coded genotypes, $X_{ij,A}$, $Z_{ij,A}$, $X_{ij,B}$, and $Z_{ij,B}$, is solely based their own genotypes without the use of their parental genotypes [124] (Fan et al., 2005b). Therefore, this method may be biased in the presence of population stratification.

To simultaneously analyze genotypes or haplotypes across multiple markers, Fan et al. [127] generalized the Hotelling's T^2 test for population data [117] to case–parent trios. Their method for genotypes across L bi-allelic markers and n case–parent trios can be outlined as following. Denote $X_{io,l}$ and $X_{ip,l}$ as the number of copies of an allele at the lth marker $(l = 1, \ldots, L)$ for offspring and parents, respectively. Define $\bar{X}_{o,l}$ and $\bar{X}_{p,l}$ as the average of $X_{io,l}$ and $X_{ip,l}$ across n trios, respectively. For a single marker locus, Fan et al. (2005a) used the statistic: $T_l = \sqrt{n}\left(\bar{X}_{o,l} - \bar{X}_{p,l}\right)/\sqrt{V_l}$, where $V_l = \sum_{i=1}^{n}\left[(X_{io,l} - X_{ip,l}) - \left(\bar{X}_{o,l} - \bar{X}_{p,l}\right)\right]/(n-1)$. T_l has an approximate $t-$distribution under the null hypothesis for the larger sample size. For L tightly linked markers, $T_l\,(l = 1, \ldots, L)$ are not independent but the T^2 statistic can be constructed: $T^2 = n\left(\bar{X}_{o,1} - \bar{X}_{p,1}, \ldots, \bar{X}_{o,1} - \bar{X}_{p,1}\right)^{\tau} V^{-1}\left(\bar{X}_{o,L} - \bar{X}_{p,L}, \ldots, \bar{X}_{o,L} - \bar{X}_{p,L}\right)$, where V is the covariance matrix of $\bar{X}_{o,l} - \bar{X}_{p,l}\,(l = 1, \ldots, L), \ldots, \ldots, \bar{X}_{o,L} - \bar{X}_{p,L}$) and can be empirically estimated from $X_{io,l}$ and $X_{ip,l}$ $(i = 1, \ldots, n; l = 1, \ldots, L)$ [127]. Under the null hypothesis of no association, T^2 has an approximate chi-square distribution with L degrees of freedom. Fan et al. [127] further extended this method to handle multi-allelic markers and haplotypes and genotypes across several haplotype blocks. They applied the T^2 statistic to a real data and obtained the smaller p-values than those obtained from the single marker analysis, indicating the proposed T^2 statistic tests are potentially more powerful.

Xu et al. [128, 129] proposed a multi-marker family-based association test that linearly combines the single-marker test statistics using weights. Suppose there are L markers and the FBAT statistic obtained from each marker locus is denoted by $z_l\,(l = 1, \ldots, L)$, Xu et al. [128, 129] proposed a class of test statistics that linearly combine the single-marker statistics: $S = W^T Z$, where $W = (w_1, \ldots, w_L)$ are weights and $Z = (z_1, \ldots, z_L)$. If W is fixed with regard to Z, then S has an approximate multinomial distribution, $N\left(0, W^T \Sigma W\right)$, under the null hypothesis of no linkage or no association, where $\sum$ is the estimated covariance matrix of Z. To obtain the weights, W, Xu et al. [129] proposed to use the statistics obtained from the "conditional mean model" [130]. For the lth marker of the jth offspring in the ith family, the conditional mean model uses $E\left(Y_{ij}\right) = \alpha + \beta E\left(X_{ij,l}\right)$, where $E\left(X_{ij,l}\right)$ is the expected genotypic value for the ith family. Let $\hat{z}_{l\beta}$ denote the standardized least-square estimator of β in the conditional mean model β in (4), then $\hat{z}_{l\beta}$ can be used as the weight to construct the following global test over L markers: $Z_{LC} = \left(\hat{Z}_{\beta}^{T} Z\right)/\sqrt{\hat{Z}_{\beta}^{T} \Sigma \hat{Z}_{\beta}}$, where $\hat{Z}_{\beta}^{T} = (\hat{z}_{1\beta}, \ldots, \hat{z}_{L\beta})$ and Z_{LC} has an approximate standard normal distribution under the null hypothesis of no linkage or no association. There are several advantages of this approach. First, it uses the same data set for the FBAT test. Second, $\hat{z}_{l\beta}$ is independent to z_k under the null hypothesis. Third, $\hat{z}_{l\beta}$ is positively correlated with z_k under the alternative hypothesis. Similar idea has been used to develop a two-stage approach that performs the screening and association tests using the same sample [131]. Simulation studies have shown that their method using the data-driven weights has the valid type I error rate and is more powerful than the Hotelling-T^2 test.

Rakovski et al. [132] extended the genotypic analysis of Chapman et al. [115] for multiple markers for case–control studies to family-based studies. The proposed method has formula similar to the T^2 test proposed by Fan et al. [127] and is identical with FBAT statistic for the single marker case [49]. Let L denote the number of markers and $U_{i,k} = \sum_{j=1}^{n_i} \left(Y_{ij} - \mu\right) \left(X_{ij,k} - E\left(X_{ij,k} | S_{i,k}\right)\right)$ denote the statistic for the ith family at the lth marker $(l = 1, \ldots, L)$. We denote $Var\left(U_{i,k}\right) = \sum_{j,k} \left(Y_{ij} - \mu\right) \left(X_{ij,k} - E\left(X_{ij,k} | S_{i,k}\right)\right) \left(Y_{il} - \mu\right) \left(X_{il,k} - E\left(X_{il,k} | S_{i,k}\right)\right)$ as the variance of $U_{i,k}$, which is calculated under the null hypothesis of no linkage or no association. Then, FBAT score across L markers can be written as $FBAT_{MM} = \left(\sum_{i=1}^{n} U_{i,1}, \ldots, \sum_{i=1}^{n} U_{i,L}\right)^{\tau} V \left(\sum_{i=1}^{n} U_{i,1}, \ldots, \sum_{i=1}^{n} U_{i,L}\right)$, where V is the covariance matrix of $\sum_{i=1}^{n} U_{i,l} \left(l = 1, \ldots, L\right)$ and can be empirically estimated from $\sum_{i=1}^{n} U_{i,l} \left(l = 1, \ldots, L\right)$ (Rakovski et al., 2007). For a large number of families, $FBAT_{MM}$ has an approximate chi-square distribution with the degrees of freedom equal to the rank of V. Similar to FBAT, $FBAT_{MM}$ can handle quantitative traits, arbitrary family structure, and arbitrary missing patterns. Rakovski et al. (2007) showed that $FBAT_{MM}$ can be more powerful than the single marker-based FBAT in their simulation studies.

Other than methods based on genotypes of multiple markers, many methods based on known haplotypes or inferred haplotypes across several markers have also been developed. For tightly linked markers, recombinants are unlikely events. Thus, it is reasonable to assume the haplotype configurations obtained do not contain recombinants in most situations. Then, haplotypes can be considered as alleles at a multi-allelic marker and the transmission of haplotypes follows the Mendelian Law of inheritance. Thus, virtually all the extended TDT methods that can handle multi-allelic markers can be directly applied to phase known haplotype data. Clayton and Jones [47] discussed the generalization of the TDT to detect association between haplotypes and the DSL based on a generalized haplotype risk model. When there are many haplotypes, the proposed test to have low power due to the large degrees of freedom. Therefore, Clayton and Jones [47] assumed that similar haplotypes tend to have similar effects and modeled the haplotype effects with a multivariate normal distribution with variance-covariance matrix of *vS*, where S is a known matrix expressing haplotype "similarity" between pairs of haplotypes and v is a single parameter that represents the haplotype association. A natural way to determine the element s_{ij} in S is to classify two haplotypes h_i and h_j either as similar $(s_{ij} = 1)$ or dissimilar $(s_{ij} = 1)$. Another nature measure of the similarity between two haplotypes is $S_{h_i,h_j}\left(l\right)$, the length of the contiguous region over which the two haplotypes, h_i and h_j, are identical by state, as was also used by Van der Meulen and te Meerman [133] and Bourgain et al. [134] .

If haplotypes are known, one can also consider transmitted haplotypes as case haplotypes and the un-transmitted haplotypes as control haplotypes and directly apply haplotypes-based methods for case–control studies to family data. Both Van der Meulen and te Meerman [133] and Bourgain et al. [134] proposed haplotype sharing-based methods. The idea behind this type of methods is that two haplotypes containing the disease allele are more closely related than two haplotypes without the disease allele and two haplotypes with one having the disease allele and one not having the disease allele. Therefore, if the marker is in the region flanking the disease gene, we expect to observe an excess length of shared haplotypes. Suppose there are n transmitted haplotypes (case haplotypes) and n un-transmitted haplotypes (control haplotypes) and that L tightly linked SNP markers are genotyped in a region of interest. The transmitted haplotypes are denoted as $h_1, \ldots, h_n$, the un-transmitted haplotypes are denoted $h_{n+1}, \ldots, h_{2n}$, and $S_{h_i,h_j}(l)$ is the length of the contiguous region around the lth marker over which the two haplotypes, h_i and h_j, are identical by state. Van der Meulen and te Meerman [133] proposed the HHS at the lth marker for the transmitted haplotypes:

$$HSS\,(l) = \sqrt{\frac{\sum_{i \neq j}\left(S_{h_i,h_j}(l)\right)^2 - \left(\sum_{i \neq j} S_{h_i,h_j}(l)\right)^2 / \left(n\,(n-1)\right)}{n\,(n-1) - 1}}.$$

They used a randomization procedure to estimate its mean and variance under the null hypothesis of no association. Under the assumption of normality, the $p-$value can be calculated for each $HHS\,(l)$. Bourgain et al. (2000) defined $A\,(l) = \frac{2}{n(n-1)} \sum\limits_{i=1}^{n} \sum\limits_{j=i+1}^{n} \mathrm{S}_{h_\mathrm{i},h_j}(l)$ and $U\,(l) = \frac{2}{m(m-1)} \sum\limits_{i=n+1}^{n+m} \sum\limits_{j=i+1}^{n+m} \mathrm{S}_{h_\mathrm{i},h_j}(l)$ for transmitted haplotypes and un-transmitted haplotypes and used $ILCS = \max_{1 \leq l \leq L} (A\,(l) - U\,(l))$ as the test statistic. Its significance can be assessed by a simple permutation test [134]. The simulation studies showed that this method is more powerful than the TDT for tingly linked markers [134, 135]. Bourgain et al. [136] further extended the MILC method to handle missing data and ambiguous haplotypes. For haplotype data, Lange and Boehnke [137] developed the Haplotype Runs Test (HRT) based on haplotype sharing. In contrast to the other proposed haplotype sharing methods, the HRT test statistic is only based on transmitted haplotypes and $S_{h_i,h_j}(l)$ is weighted by allele frequencies at makers. For missing data and ambiguous haplotypes, Lange and Boehnke [137] used the same strategies proposed by Bourgain et al. [136] to define $S_{h_i,h_j}(l)$. Their simulations results showed the HRT is more powerful than the method of Bourgain et al. [134, 136].

In most situations, the haplotypes are unknown and they cannot be completely determined even using family data. One strategy is to identify the most likely haplotype configuration without recombinants and treat them as phase known data. But such strategy ignores the uncertainty in haplotype inference, which can result in a loss of power. Another way is to identify all compatible haplotype configurations with their posterior probabilities and analyze them together. However, the identification of the most likely haplotype configuration and all compatible haplotype configurations often involves the estimate of haplotype frequencies based on likelihood function, which may not be robust for stratified populations. Therefore, many methods that can use multiple haplotype configurations to account for the uncertainty in haplotype inference but are still rubout for stratified populations have been proposed.

Several methods have been proposed based the full or the partial of the following likelihood function ([37, 138]; Clayton, 1999; Seltman et al., 2001) to construct a likelihood-ratio test. When there are ambiguities:

$$L^{(F)} = P\left(g_m, g_f, g_o | D\right) = P\left(g_m, g_f | D\right) P\left(g_o | g_m, g_f, D\right) = L^{(P)} L^{(O)},$$

where D represents the affected status of the offspring, $g_o,\ g_m$, and g_f are the genotypes (or paired haplotypes if they are known) of the offspring, mother, and father, respectively. The first component, $L^{(P)}$, depends on the haplotype relative risks and haplotype frequencies in the population. The second component, $L^{(O)}$, only depends on the haplotype relative risks. Both $L^{(F)}$ and $L^{(O)}$ can be used to test if the haplotype relative risks equal one. The inclusion of $L^{(P)}$ can increase the power but introduce bias for stratified populations. Clayton [36] proposed a score test based on the partial-likelihood to reduce the influence of population stratification as much as possible. The test statistic is derived either based on $L^{(F)}$ for those trios with ambiguous haplotypes or $L^{(O)}$ for those trios with determined haplotypes. Cordell et al. (2002) proposed a unified stepwise regression procedure for association mapping using family data based on $L^{(O)}$ only:$L^{(O)} = P\left(G_c | G_m, G_f, D\right) = \frac{P(D|G_c,G_m,G_f)}{\sum_{G*\in G} P(D|G*,G_m,G_f)P(G*|G_m,G_f)P(G_m,G_f)}$, where $P\left(G^* | G_m, G_f\right)$ are functions of underlying population haplotype frequencies and recombination fractions and generally take different values. Seltman et al. [139] adapted the partial likelihood approach [36] to simultaneously test linkage and association between haplotypes and the disease. To reduce the number of degrees of freedom required for testing a large number of haplotypes and improve power, Seltman (2001) proposed the "evolutionary tree" (ET)–TDT. Simulation results have shown that the ET–TDT can be more powerful than other proposed methods under reasonable conditions [139]. Cordell et al. (2002) proposed to use some event ξ in the family such that $P\left(G * | \xi\right)$ will not depend on the underlying haplotype frequencies and the recombination fractions. Then, $L^{(O)}$ can be expressed in a conditional logistic regression framework. They proposed several strategies to choose the event ξ. For example, we can use all families that the parental haplotypes can be unambiguously deduced without recombinants or all families that the haplotypes are "inferable" (Cordell et al., 2002). However, this method may discard many families and results in loss of power. Dudbridge [138] used the full likelihood function $L^{(F)}$ and the EM algorithm to maximize $L^{(F)}$ under the null hypothesis as well as under the alternative hypothesis in the presence of ambiguous haplotypes. Although they assumed the one single population with HWE, population stratification will only lead to a conservative test [138].

Other than likelihood-based methods, Zhao et al. [119] developed a method to construct the transmission/nontransmission table based on the estimated haplotype frequencies and proposed a test that is similar to TDT_{SE}. Specifically, the contribution of a compatible haplotype assignment $\{G_m, G_f, G_O\}$ of a trios to the transmission/nontransmission table is weighted by $P\left(G_m\right) P\left(G_f\right) /$

$\sum_{\left\{G_m^*, G_f^*\right\}} P\left(G_m^*\right) P(G_f^*)$, where $\left\{G_m^*, G_f^*\right\}$ are all possible haplotype assignments for this trios. Under the null hypothesis of no linkage, Zhao et al. [119] proved the table is symmetric and such symmetry is not affected by the choice of haplotype frequencies. Thus, the test proposed by Zhao et al. [119] does not suffer from inflated type I error rates due to population stratification. Knapp and Becker [140] observed that the symmetry of the transmission/nontransmission table is not essential for the validity of the test proposed by Zhao et al. [119]. Based on this observation, Knapp and Becker [140] proposed several intuitive modifications for Zhao's method that can potentially increase its power but not affect its validity, and extended Zhao's method to handle general nuclear families with arbitrary number of affected and unaffected offspring.

Lin et al. [141] proposed the exhaustive allelic transmission disequilibrium test (EATDT) to test the linkage and association from case–parent data. The underlying idea of EATDT is that either methods based on single markers or methods based on haplotypes are optimal in all situations and it is difficult to know which method is optimal in a given data. But we may gain additional power by exhaustively searching all alleles including single markers as well as all haplotypes with lengths less than a pre-specified threshold within a window. For n families and any given window containing L bi-allelic markers, a transmission/nontransmission table is constructed and then the corresponding p-value is calculated based on either a specific allele at a single makers or a specific haplotype across multiple markers within the window in EATDT. The p-values are calculated over all combinations of markers within the window and the minimum p-value among them is chosen as the p-value for the window. Instead of using the Bonferroni correction, Lin et al. [141] proposed a permutation procedure to adjust its significance in EATDT. Simulation results showed that EATDT can detect both common and rare DSLs in GWA studies.

Zhang et al. [53] proposed a HS-TDT method to evaluate linkage and association between a maker and the DSL by assessing the correlation between the trait value and the difference of haplotype-sharing scores between the parental haplotypes that are transmitted and not transmitted to their offspring. The MILC method proposed by Bourgain et al. [134–136] is a special case of the HS-TDT. Let $X_{ij}\left(l\right)$ denote the difference of the haplotype-sharing scores between the parental haplotypes that are transmitted and not transmitted to the jth offspring in the ith family at the lth marker [53], then the proposed score for the ith family can be expressed as: $U_i\left(l\right) = \sum_{j=1}^{n_i}\left(Y_{ij} - \mu\right) X_{ij}\left(l\right)$. Under the null hypothesis of no linkage or no association, the trait values are independent of $X_{ij}\left(l\right)$, thus $E\left(U_i\left(l\right)\right) = 0$. The test statistic across all L markers is defined as $U = \max_{l=1,\ldots,L}\left(U\left(l\right)\right)$, where $U\left(l\right) = \sum_{i=1}^{n} w_i U_i\left(l\right)$ and the summation is over all families and w_i is a constant. The significance of U can be assessed using a permutation procedure. For phase unknown data, the haplotype configurations with their corresponding posterior probabilities can be estimated using the EM algorithm with the assumption of a single population with HWE and $X_{ij}\left(l\right)$ can be defined accordingly. Zhang et al. [53] proved that the choice of μ in $U_i\left(l\right)$, w_i in the $U\left(l\right)$, and the haplotype frequencies will only affect the power but not the validity of HS-TDT. Thus, HS-TDT is still valid in the presence of population stratification. The optimal values of μ and

w_i to achieve the maximum power is unclear. In practice, μ can be set as 0 for the qualitative trait and as mean over all children in all families for the quantitative trait. w_i can be set as $1/n_i$ according to the number of offspring in the family [113] or 1 to all families [50]. HS-TDT is applicable to arbitrary nuclear families and can handle both qualitative and quantitative traits. The simulation results also showed that HS-TDT is more powerful than existing single-marker TDTs and haplotype-based TDTs. However, the haplotype sharing based methods are not robust to genotyping errors, missing data, and recent marker mutations. Knapp and Becker [142] found that the haplotype sharing used in HS-TDT [53] can result in an inflated type I error rate in the presence of genotyping errors.

Several methods have been proposed to define robust haplotype sharing based statistics with genotyping errors and missing data (Bourgain et al., 2002; [143–145]), readers are referred to these papers for more details. In this book, Epstein and Kwee [146] reviewed two haplotype based approaches in details. One is the approach of Horvath et al. [147], which is the extension of the FBAT approach to haplotypes and is applicable for arbitrary family structures and missing parental genotypes. The other is a robust approach recently developed by Allen and Satten [148], which can perform haplotype and haplotype–environment interaction analysis within case–parent triads. Readers can refer to this chapter for more details about these two methods. In addition, the development of haplotype-based methods is an active research area and many different methods have been proposed ([149–157]; Onkamo et al., 2002).

8 Other Association Methods Using Family-Based Designs

8.1 General Pedigrees

Although some aforementioned methods, such as FBAT, have been generalized to handle large pedigrees with multiple generations, most of them are only applicable to nuclear families. A large number of large pedigrees with multiple generations have been collected and are continually being collected in practice. Thus, it is important to develop methods that can analyze data from general pedigrees. Martin et al. [51] developed a valid test of association using general pedigrees and Zhang et al. [156] generalized this method for quantitative phenotypes. Abecasis et al. [105] also proposed a similar test tests for qualitative phenotypes using general pedigrees. Martin et al. [157] further extended such allele based tests to genotype-based tests that are applicable to general pedigrees. The basic idea of these methods is to collect all informative nuclear families and all informative sibships in a single pedigree as a unit in the test statistic. We can at least consider three types of informative nuclear families within a pedigree: (1) the parental genotypes are completely available; (2) only one parent has the genotype and the family satisfies the condition specified by Sun et al. [50] ; (3) the parental genotypes are not available but the family has at least one affected sibling and one unaffected sibling. For these three types of families, we can construct three statistics U_1, U_2, and U_3, which have the expected value of 0 under the null hypothesis of no linkage or no association [50, 51, 54]. Denote $U = U_1 + U_2 + U_3$ as the summary statistic, the estimated variance of U is just U^2 under the null hypothesis and it is unbiased in the presence of linkage. More generally, we can use a weighted statistic $U = w_1 U_1 + w_2 U_2 + w_3 U_3$, where $w_i\ (i = 1, 2, 3)$ is the weight for U_i. In general, the weights are calculated according to the number of informative nuclear families and Martin et al. [158] provided a weighting scheme to avoid potential bias in their test. Liu and Gordon [159] considered a set of weights that can achieve more power. However, the optimal weights that can achieve the maximum power are still unclear. Suppose we have n pedigrees and U_i is the summary statistic for pedigree i, then $Z = \frac{\sum_i U_i}{\sqrt{\sum_i U_i^2}}$ has an asymptotic standard normal distribution.

Cantor et al. [160] proposed a likelihood-based association test in a linked region using large pedigree, which is an extension of method by Xiong and Jin [161] . For a pedigree with n individuals, the prospective likelihood function can be written as: $L = \sum_{g_1,\ldots,g_n} \prod_i \text{Pentrance}\,(Y_i|g_i) \prod_j \text{Prior}\,(g_j) \prod_{(k,l,m)} \text{Transmission}\,(g_m| g_k, g_l)$. Here, person i has the phenotype Y_i and possible genotype g_i, the product on j is over all founders, and the product on (k, l, m) is over all parent–offspring trios. The likelihood can be parameterized as a function of recombination rate and linkage disequilibrium between the marker locus and trait locus. Thus, the likelihood ratio test can be constructed to test the association and linkage. However, the calculation of the prior probability of founder's genotypes requires the good estimate of haplotype frequencies, this method may be biased for structured populations.

Chen and Abecasis [106] developed tests that are applicable to general pedigrees and GWA studies. Due to the high cost in GWA studies, not all samples with phenotypes will be genotyped. In this situation, Chen and Abecasis [106] developed a method to impute genotypes of ungenotyped individuals based on genotypes of their genotyped relatives. The model used in Chen and Abecasis [106] is similar to the model used in QTDT [105]. Specifically, they assumed that the mean of Y_{ij} satisfies: $E\,(Y_{ij}) = \mu + \beta_g g_{ij} + \beta_c C_{ij}$, where β_g and β_c are genetic effect and covariate effect, and g_{ij} is the number of copies of allele A_1 for biallelic markers. For each family, the $n_i \times n_i$ covariance matrix, Ω_i, has the identical elements with the covariance matrix in QTDT. For individuals without genotypes, the genotypic score, g_{ij}, is replaced by $\bar{g}_{ij}$, the estimated genotypic score based on its genotyped relatives. Chen and Abecasis [106] showed that their methods have appropriate type I error rates and more power when phenotype information for ungenotyped individuals is included in analysis. However, their methods did not distinguish between and within family effects thus may not be robust in the presence of population stratification.

8.2 Gene–Gene $(G \times G)$ interaction and Gene–Environment $(G \times E)$ Interaction

It is generally believed that gene–gene interactions and gene–environment interactions play an important role in many human complex diseases. Many methods have been developed to detect statistical interactions. Statistical interactions between genes and environmental variables may not correspond to biological interactions. They generally mean that the joint effects of genetic and environmental variables can not be added, if the additive model is assumed, or cannot be multiplied, if a multiplicative model is assumed. Umbach and Weinberg [162] adopted such statistical definition of interaction and extended the log-likelihood model to study the interaction of gene and a binary environmental variable. Denote (O, M, F) as the mating type, E as the binary environmental variable, and $p_{(O,M,F,E)|D}$ is the conditional probability of mating type (O, M, F) and E given the disease status of offspring, then the model proposed by Umbach and Weinberg [162] is

$$\begin{aligned}\log\left(p_{(O,M,F,E)|D}\right) = \mu_{(M,F)} + \delta_{(M,F)} I_{\{E=1\}} + \beta_O + \eta_O I_{\{E=1\}} \\ + \log(2)\, I_{\{(O,M,F)=(1,1,1)\}}.\end{aligned}$$

The null hypothesis of no interaction between O and E can be tested by setting $\eta_1 = \eta_2 = 0$. This test is equivalent to a test studied by Schaid [46]. In this model, one implicit assumption is that, conditional on the parental genotypes, the offspring's environmental status is independent of its genotype at the candidate locus. Cordell et al. [37] extended the conditional logistic regression to model gene–gene interactions and gene–environment interactions.

Hsu et al. [163] extended the multipoint approach of by Liang et al. [120] and developed a new method to detect gene–gene interactions in two unlinked regions. The method is based on the preferential-allele-transmission statistic and can be outlined as follows. Suppose two regions I and II are unlinked and there is no more than one DSL in each region. Suppose that are L_1 markers in region I and L_2 markers in region 2, Hus et al. (2003) defined preferential-allele-transmission statistic for these two region as M_{il}^1 and F_{il}^1 and M_{il}^2 and F_{il}^2, respectively [121]. Liang et al. [121] have derived that the expected M_{il}^1 and F_{il}^1 (or M_{il}^2 and F_{il}^2) conditional on the disease status of offspring, is a function of marker position in region I, t_l^1 (or in region II, t_l^2), the DSL location in region I, τ_1 (or in region II, τ_2), and other parameters: $E\left(M_{il}^1\right) = E\left(F_{il}^1\right) = \mu_1\left(t_l^1; \tau_1, C_1, N_1, \pi_l^1\right)$ $\left(E\left(M_{il}^2\right) = E\left(F_{il}^2\right) = \mu_2\left(t_l^2; \tau_2, C_2, N_2, \pi_l^2\right)\right)$, Hus et al. (2003) further derived that the expected M_{il}^2 and F_{il}^2 conditioning on M_{il}^1 and F_{il}^1 and the disease status of offspring, which is a function of marker position in region II, t_l^2, the DSL location in region 2, τ_2, and other parameters: $E\left(M_{il}^2 | M_{il}^1\right) = E\left(F_{il}^2 | F_l^1\right) = \mu_3\left(t_l^2; \tau_2, C_3, N_3, \pi_l^2\right)$. Based on these equations, GEE approach can be used to estimate all parameters and test gene-gene interaction in two unlinked regions. The method can utilize multiple markers into the analysis and only assumes that the two regions are completely unlinked and that there is no more than one DSL in each region but does not assume any particular mode of inheritance.

Lake and Laird [164] proposed a method to study gene–environment interactions based on correlation of genotypes and environmental covariates and parental mating types. Specifically, the proposed FBAT-I statistic has the following formula: $T = \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_{OMF}\right) \left(E_{ij} - \bar{E}_{OMF}\right)$, where X_{ij} and E_{ij} are the genotype and the measure of environmental covariate for the jth offspring in the ith family, $\bar{X}_{OMF}$ and $\bar{E}_{OMF}$ are the parental mating type specific mean of the genotype and the environmental covariate for the mating type (O, M, F). Under the null hypothesis of no gene–environment interactions, T has an expected value of 0. However, the usual method for calculating the variance of T in standard FBAT methods is not applicable due to the possible main effects of the gene under the null hypothesis. Thus, Lake and Laird [164] proposed to permute the residuals of $X_{ij} - \bar{X}_{OMF}$ and $E_{ij} - \bar{E}_{OMF}$ within each mating type to obtain the empirical distribution of T and calculate its empirical p-value. Since FBAT-I is based on the parental mating type, the method is robust to population stratification and can easily extended to test parent-of-origin effects.

Martin et al. [165] proposed the MDR-TDT, which combines the multifactor dimensionality reduction (MDR) [166] and the genotype-TDT [157], to detect gene–gene interaction based on family data. The MDR method was initially developed to detect gene–gene interactions based on case–control samples, while the genotype-PDT was developed to test for association between the DSL and genotypes at a locus or multiple loci. In the MDR-TDT, the genotype-PDT statistic other than the test statistic of multiple marker loci from case–control samples is used to identify high-risk multilocus genotypes.

9 Software Packages and Power Consideration

Due to the complexity of implementing analysis methods for family-based association designs, there is a generally need for special software to perform such analysis. Fortunately, a large number of software packages have been developed by the original authors of the methods and many of them have been widely used in the analysis of family-based association studies. We have compiled a list of several commonly used programs and described their functions in Table 1. We also include several packages that can calculate power and sample sizes for family-based association designs because these software packages are helpful for designing family-based association studies. We also suggest users to choose their program upon their analysis compatibility of pedigree structures, phenotypes, and genotypes. For a detailed description of these software packages, please refer to corresponding manuals. For a more complete list of software packages that can analyze family-based association studies, please refer to http://linkage.rockefeller.edu/soft/.

Table 1 Available software packages for analysis of family-based association studies

Software	Pedigree	Phenotype	Genotype	Link	References
FBAT	General pedigree	Binary trait, quantitative trait, ordinal trait, time-on-set, $G \times G$ and $G \times E$ interaction	Single marker, haplotype, genotype of Multiple markers	http://www.biostat.harvard.edu/∼fbat/fbat.htm	Horvath et al., 2000; Laird and Lange, 2006; [49, 130, 167, 168]
GASSOC	Trio	Binary trait	Single marker	http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm	[32]
Genetic Power Calculator	Trio	Binary trait, quantitative trait	Single marker	http://pngu.mgh.harvard.edu/∼purcell/gpc/	[169, 170]
HS-TDT	Nuclear family	Binary trait, threshold-selected quantitative trait	Haplotype	http://www.math.mtu.edu/∼shuzhang/software.html	[53]
Multiple TDT	Trio	Binary trait	Haplotype	http://bioinformatics.med.yale.edu/group/software.html	[119]
PBAT	General pedigree	Binary trait, quantitative trait, ranked trait, time-on-set, $G \times G$ and $G \times E$ interaction	Single marker, haplotype, genotype of multiple markers	http://www.biostat.harvard.edu/∼clange/default.htm	[168, 171–174]
PDT	General pedigree	Binary	Single marker	http://www.chg.duke.edu/software/pdt.html	[51]
PEDGENIE	General pedigree	Binary trait, quantitative trait	Single marker, haplotype	http://bioinformatics.med.utah.edu/PedGenie/index.html	[149]
QTDT	General pedigree	Quantitative trait	Single marker	http://www.sph.umich.edu/csg/abecasis/QTDT/	Abecasis et al. (2000a)
QUANTO	Trio, Sib pair	$G \times G$ and $G \times E$ interaction	Single marker	http://hydra.usc.edu/GxE	[175–177]

(*continued*)

Table 1 (continued)

Software	Pedigree	Phenotype	Genotype	Link	References
TDTASP	Nuclear family, sib pair	Binary trait	Single marker	http://biostatistics.mdanderson.org/SoftwareDownload/	[178]
TDT-AE	Trios	Binary trait	Single marker	ftp://linkage.rockefeller.edu/software/tdtae2	[176, 177]
TDT/S-TDT	Nuclear family	Binary trait	Single marker	http://genomics.med.upenn.edu/spielman/TDT.htm	[25, 70]
TRANSMIT	Nuclear family	Binary traits	Haplotype	http://www-gene.cimr.cam.ac.uk/clayton/software/	[36]

Table 2 The description of the original Oxford ACE data and three testing data sets

	Data set			
	Original data	First testing data	Second testing data	Third testing data
Family types	General pedigrees	General pedigrees	Nuclear family	Trios
# of families	83	69	57	41
# of individuals	666	553	280	123
# of individuals without genotype	111	91	14	9
# of founders without genotype	107	87	14	9
# of individuals without phenotype	261	148	33	23

To illustrate these software packages, we use the Oxford ACE data set. The Oxford ACE data is from a study of the functional mutation in the angiotensin-I converting enzyme (ACE) gene [24]. This data has been used to illustrate several newly developed methods for TDT for quantitative traits (Abecasis et al., (2000b); [126]) and for haplotype inference from general pedigrees (O'Connell 2000; [179]). The Oxford ACE data contains 666 individuals from 83 extended pedigrees. Pedigrees range in size from two to three generations, including 4–18 individuals each. Genotypes are available for 555 individuals at ten bi-allelic markers in strong LD, spanning a very short region (26 kb) within the ACE gene. The overall percentage of missing data is about 20.0%. The phenotype, circulating ACE level is available for only 405 individuals. We remove 14 families without circulating ACE levels and use genotype and phenotype data of 553 individuals from 69 families as our first testing data. We also created pseudo disease status for each individual based on his/her circulating ACE level. Specifically, we assign the affected status to an individual if his/her circulating ACE level is greater than the median of observed circulating ACE levels. Since some software packages are only applicable to nuclear families or case–parent trios, we choose one nuclear family and one case–parent trios from each extended family to form our second and third testing data sets. A brief description of the original data set and three testing data sets can be found in Table 2.

We apply several commonly used software packages to these testing data sets. First, we apply FBAT and QPDT (see Table 1 for more details about these software) to the first and second testing data sets with the circulating ACE level. For both methods, we conduct the analysis based on single markers. For FBAT, we also conduct haplotype analysis based on two adjacent makers. The results are listed in Table 3. We can see that all makers at the ACE locus are strongly linked to ACE levels and evidence for association is strongest at the I/D marker locus because the p-values based on single maker analysis from both FBAT and QPDT are smallest.

Table 3 The testing results from FBAT and QPDT for the Oxford ACE data. The circulating ACE level of each individual is used. All p-values are not adjusted for multiple testing. The p-values from the haplotype analysis are the minimum p-value among all haplotypes tested

Marker name	Marker position (in base pair)	p-values					
		FBAT (single marker analysis)		QPDT (Single marker analysis)		FBAT (Two-locus haplotype analysis)	
		First data	Second data	First data	Second data	First data	Second data
T-5491C	$-2,851$	4.1×10^{-11}	5.0×10^{-7}	4×10^{-12}	1×10^{-8}	8.8×10^{-11}	8.2×10^{-2}
A-5466C	$-2,826$	1.4×10^{-11}	8.0×10^{-7}	N/A	N/A	1.6×10^{-11}	3.8×10^{-6}
T-3892C	$-1,252$	7.1×10^{-13}	5.3×10^{-7}	N/A	N/A	7.7×10^{-12}	4.0×10^{-6}
A-240T	$2,400$	4.1×10^{-12}	1.6×10^{-6}	4×10^{-13}	2×10^{-7}	6.6×10^{-11}	6.0×10^{-6}
T-93C	$2,547$	4.0×10^{-12}	4.4×10^{-7}	2×10^{-13}	7×10^{-8}	2.6×10^{-12}	8.3×10^{-6}
T-1237C	$8,128$	3.9×10^{-12}	4.4×10^{-6}	4×10^{-14}	2×10^{-8}	4.6×10^{-14}	7.0×10^{-7}
G2215A	$12,257$	2.0×10^{-15}	4.5×10^{-7}	2×10^{-17}	3×10^{-9}	1.9×10^{-14}	2.0×10^{-6}
I/D	$14,094$	1.4×10^{-15}	4.0×10^{-8}	2×10^{-18}	5×10^{-11}	7.9×10^{-13}	9.7×10^{-7}
G2350A	$14,521$	2.0×10^{-13}	3.2×10^{-7}	1×10^{-17}	3×10^{-10}	4.1×10^{-13}	6.0×10^{-7}
4656(CT)3/2	$23,945$	1.5×10^{-14}	1.9×10^{-7}	9×10^{-18}	9×10^{-11}	N/A	N/A

We then test the association between the ACE locus and the pseudo disease status using FBAT, TRANSMIT, and MultipleTDT (see Table 1 for more details about these software) and the results listed in Table 4. For FBAT, we perform both single marker and haplotype based analysis. For TRANSMIT and MultipleTDT, we only conduct haplotype based analysis. All haplotype analysis is based on twp adjacent markers.

10 Discussion

Mapping genes underlying complex human diseases presents great challenges for human geneticists. Theoretical and empirical studies have shown linkage analysis as a tool for mapping disease genes is less powerful than association based analysis. On the other hand, traditional case–control association designs using unrelated samples may be biased in the presence of population stratification. Family-based association designs, which are robust to the population stratification, provide a compromise between the above two approaches. In addition, family-based association studies offer a solution to detect genomic imprinting. Imprinting, also known as "parent-of-origin effects," are referred to different effects of an allele on the offspring that depend on the parental source of that allele. Parent-of-origin effects have been found in many genes and diseases (e.g., [182, 183]). Several methods have been proposed to detect parent-of-origin effects based on family-based association designs. Weinberg et al. [42] and Weinberg (1999) extended their log-likelihood method to detect genomic imprinting. Cordell et al. [37] described how to use a generalized linear model to detect genomic imprinting. Whittaker et al. [184] illustrated how to use simple linear models to estimate parent-of-origin effects for quantitative phenotypes. Hu et al. [185] extended 1-TDT method of Sun et al. [50] and their method can incorporate families with only parent available to detect imprinting.

Recently, GWA studies, which aim to genotype hundreds of thousands SNPs across the human genome for a large number of samples, have proved to be a powerful approach to detect genes underlying complex human diseases (e.g., [19–21]). Since hundreds of thousands of markers are genotyped in GWA studies, the statistical power of such studies can be diluted due to the correction of multiple-testing problem. To avoid this problem, multiple-stage designs have been proposed in GWA studies (e.g., [186]). In such designs, multiple independent sets of samples are collected. One of them is genotyped at all SNPs and used to select a small set of candidate SNP markers. Other sets of samples are only genotyped and analyzed at this small set of SNP markers. Since only a small set of SNP markers is tested in the final stage, the number of association tests is reduced and the correction of multiple-testing problem is less severe. However, such deigns need to collect multiple sets of samples. Family-based association designs potentially provide a solution to achieve the power of multiple-stage designs using a single set of samples. Van Steen et al. [132] proposed a two-stage approach that performs the screening and association tests using the same sample. In the first stage, the markers are selected based on

Table 4 The testing results from FBAT, TRANSMIT, and MultipleTDT for the Oxford ACE data. The pseudo disease status each individual is used. All p-values are not adjusted for multiple testing. The p-values from the haplotype analysis of FBAT are the minimum p-value among all haplotypes tested. The p-values from the haplotype analysis of TRANSMIT are global p-values abd are calculated based on approximate chi-square distribution. The p-values from the haplotype analysis of MultipleTDT are global$p$-values and are obtained based on 1,000 permutations

Marker name	Marker position (in base pair)	p-values					
		FBAT (Single marker analysis)		FBAT (Two-locus haplotype analysis)		TRANSMIT	MultipleTDT
		Second data	Third data	Second data	Third data	Third data	Third data
T-5491C	$-2,851$	1.4×10^{-2}	1.6×10^{-3}	4×10^{-12}	1.8×10^{-3}	2.8×10^{-3}	< 0.001
A-5466C	$-2,26$	5.6×10^{-2}	5.3×10^{-3}	N/A	1.8×10^{-3}	1.1×10^{-2}	<0.001
T-3892C	$-1,252$	3.3×10^{-2}	6.5×10^{-3}	N/A	2.7×10^{-3}	1.4×10^{-2}	<0.001
10^{-3}**A-240T**	$-2,400$	4.4×10^{-2}	7.1×10^{-3}	4×10^{-13}	2.7×10^{-3}	4.4×10^{-2}	0.001
T-93C	$-2,547$	3.3×10^{-2}	5.3×10^{-3}	2×10^{-13}	6.0×10^{-3}	1.3×10^{-2}	0.009
T-1237C	$-8,128$	2.9×10^{-1}	5.0×10^{-2}	4×10^{-14}	4.7×10^{-3}	3.8×10^{-2}	N/A
G2215A	$-12,257$	3.3×10^{-2}	4.3×10^{-3}	2×10^{-17}	4.3×10^{-3}	1.1×10^{-2}	N/A
I/D	$-14,094$	3.1×10^{-2}	1.1×10^{-2}	2×10^{-18}	5.3×10^{-3}	1.1×10^{-2}	0.003
G2350A	$-14,521$	6.1×10^{-2}	5.3×10^{-3}	1×10^{-17}	5.3×10^{-3}	2.4×10^{-2}	N/A
4656(CT)3/2	$23,945$	8.8×10^{-2}	1.3×10^{-2}	9×10^{-18}	N/A	N/A	N/A

phenotypes of offspring and between-family genotype scores. In the second stage, the selected markers are tested using phenotypes of offspring and within-family genotype scores. The idea of Van Steen's method is that the screening test in the first stage is statistically independent of the association test in the second stage. In addition, the method is robust to population stratification, since only FBAT statistic is used in the final stage [132]. Feng et al. [187] further extended this approach that can utilize general pedigree with an arbitrary structure and phenotypes of founders and parents in families.

In this chapter, we have reviewed methods that can analyze data from family-based association studies. We first focused on methods that can analyze the simplest family based association design with one affected offspring with its two parents, all genotyped at a bi-allelic marker locus. We then discussed its various extensions that can increase power and utilize multi-allelic markers, families with multiple siblings, families with incomplete parental genotypes, quantitative traits, and multiple tightly linked markers. There are many other available methods not reviewed in this book chapter and it is beyond the capacity to review all available methods in a single book chapter. Readers are referred to other review papers for more details (e.g., Laird and Lange, 2006; [188–190]) However, we still would like to mention several types of analysis methods for family-based association designs. Ho and Bailey-Wilson [191] extended the TDT methods to test for linkage between X-linked markers and diseases that affect either males only or both genders. Similarly, Horvath et al. [192] proposed two procedures: the XS-TDT and the XRC-TDT that extended the S-TDT [70] and RC-TDT (Knapp, 1999a), respectively. For age of disease onset, Ghosh and Reich [193] extended the S-TDT [70] with the adjustment for age of onset. The FBAT methods have been extended to incorporate age of onset information with various phenotype coding [194, 195]. Both unrelated case and controls samples and family samples can be available from a single study or multiple studies. Joint analysis of them can improve the power and increase the opportunity to detect the DSL. Several methods have been proposed to jointly analyze data from families and unrelated samples [63, 196–198]. Simulation results showed these methods can provide more power than methods using only one type of data. Readers are referred to those papers for more details. In GWA studies for complex human diseases, a set of phenotypes would be used to characterize diseases and measured at the same time. These phenotypes may be correlated due to the same pathway or shared environmental factors. In such situation, joint analysis of multiple phenotypes tends to be more powerful than the analysis of all phenotypes individually with corrections of multiple-testing problem. In addition, the association tests for different phenotypes will be correlated, thus the correction of multiple-testing problem will tend to be conservative. Lange et al. [199] generalized the FBAT to tests all phenotypes simultaneously. The proposed test, FBAT-GEE test, is flexible to test phenotypes from different types (e.g., continuous phenotypes, discrete phenotypes, etc.). Another challenge to analyze multiple phenotypes is how to choose phenotypes that should be included in the analysis. One way is to use the variance component analysis but results can be difficult to interpret. Thus, Lange et al. [200] proposed a two-stage approach to select phenotypes that will be tested in the final stage to reduce the burden of multiple-testing problem. However, developments of more advanced methods to analyze multiple phenotypes are still warranted.

Given that many statistical methods have been proposed in the last several years for family-based association studies, the performance of these methods is of great interest to human geneticists who study complex traits. Many studies have been done to compare various methods (e.g., [201–204]) Lange et al., 2002. In a recent study, Nicodemus et al. [204] assessed the type I error rate and compared the performance of several commonly used methods for family-based association designs, including FBAT [44,52], PDT [46], SDT [74], TDT (Spielman et al., 1993), TRANSMIT [37], and several other methods. Through extensive simulations, they found that nearly every method can maintain appropriate type I error rates under all conditions. Although no single method is uniformly more powerful than the other methods, their power varied greatly and the difference in terms of power between the most powerful method and the least powerful method can be as large as 50%. Nicodemus et al. [204] found that the relative performance of different methods clearly depends on many factors, including pedigree structure, missing patterns of parental genotypes, population structures, and genetic models. Because the mode of inheritance for complex diseases is usually unknown, methods that perform well under a wide range of models are certainly desirable. As more and more approaches are introduced in the literature, systematic comparisons are always needed to give guidelines to human geneticists.

Although we have outlined several attractive features that make family-based association useful in GWA studies of common human diseases, they have been criticized for several limitations. First, GWA studies require to genotype a large number of samples at hundred of thousands of markers. Compared with case–control studies, family-based association studies need to recruit a large number of samples with their relatives, which is more difficult than recruiting a large number of unrelated individuals. Second, case–control association study designs are more powerful than family-based association designs, especially for common human diseases. It is a general belief that gene–environment interactions, as well as gene–gene interactions play an important role in many complex human diseases. In terms of power, case–control designs are superior to family-based association tests for detecting gene–gene and gene–environment interactions. Case–control association designs have often been criticized for inducing false positives due to population stratification. Several methods have been proposed to use genomic markers to control population stratification in the analysis of case–control data (Devlin and Roeder, 1999; [205–207]). Third, family-based association designs are more sensitive genotype errors. In case–control studies, random genotyping errors will only make tests conservative under the null hypothesis. But with family-based association designs, random genotyping error can lead to inflated type I error rates [141, 179, 208, 209]. For this reason, several methods have been proposed to incorporate genotyping errors (Cheng and Chen et al., 2007; [143, 179, 180]). Nonetheless, family-based association designs will still plan an important role in GWA studies, given that a large number of family data have already been collected in linkage studies and are now available for association studies. The great challenges facing statistical geneticists in the coming years are to develop statistically powerful and computationally feasible methods to fully utilize such data, and to search for optimal study strategies to map complex disease genes in the post-genome era.

Acknowledgements We are grateful to the authors of reference [24] for the use of the Oxford ACE data, which is available on request from Dr. Martin Farrall (mfarrall@well.ox.ac.uk). This work was supported in part by the grants NIH GM074913 (KZ) and GM59507 (HZ) from the National Institute of General Medical Sciences.

References

1. Ogden CL, Flegal KM, Carroll MD, Johnson CL (2002) Prevalence and trends in overweight among US Children and adolescents, 1999–2000. JAMA – J Am Med Assoc 288:1728–1732
2. Flegal KM, Carroll MD, Ogden CL, Johnson CL (2002) Prevalence and trends in obesity among US adults, 1999–2000. JAMA-J Am Medical Assoc 288:1723–1727
3. Harris MI, Flegal KM, Cowie CC, Eberhardt MS, Goldstein DE, Little RR, Wiedmeyer HM, Byrd-Holt DD (1998) Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in US adults - The Third National Health and Nutrition Examination Survey, 1988–1994. Diabetes Care 21:518–524
4. Risch N (2000) Searching for genetic determinants in the new millennium. Nature 405: 847–856
5. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234–238
6. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073–1080
7. Strathdee CA, Gavish H, Shannon WR, Buchwald M (1992) Cloning of cDNAs for Faconi's anaemia by functional complementation. Nature 356:763–767
8. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W et al. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 253:66–71
9. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G (1995) Identification of the breast cancer susceptibility gene BRCA2. Nature 378:789–792
10. Lewontin RC (1964) The interaction of selection and linkage I. general considerations. Genetics 49:49–67
11. Boehnke, M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Human Genet 55:379–390
12. Kruglyak L, Lander ES (1995) High-resolution genetic mapping of complex traits. Am J Human Genet 56:1212–1223
13. Jorde LB (1995) Linkage disequilibrium as a gene mapping tool. Am J Human Genet 56:11–14
14. Ardlie K, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299–309
15. Carlson CS, Eberle MA, Kruglyyak L, Nickerson DA. 2004. Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
16. Clark AG (2003) Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Develop 13:296–302
17. Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7:385–394

18. Nordborg M, Tavaré S (2002) Linkage disequilibrium: what history has to tell us. Trend Genet 18:83–90
19. Gudmundsson G, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA et al. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39:631–637
20. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M et al. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 39:984–988
21. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z et al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39:645–649
22. McGinnis R, Shifman S, Darvasi A (2002) Power and efficiency of the TDT and case-control design for association scans. Behav Genet 32:135–144
23. Witte JS, Gauderman WJ, Thomas DC (1999) Asymptotic bias and efficiency in and case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epiemiol 149:693–705
24. Keavney B, McKenzie CA, Connell JM, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, Farrall M (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. Human Molecular Genet 7:1745–1751
25. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Human Genet 52:506–516
26. Ewens WJ, Spielman RS (2005) What is the significance of a significant TDT? Human Heredity 60:206–210
27. Bickeboller H, Clerget-Darpoux F (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Genet Epidemiol 12:865–870
28. Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. Human Heredity 48:67–81
29. Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. Ann Human Genet 59:323–336
30. Spielman RS, Ewens WJ (1996) The TDT and other family based tests for linkage disequilibrium and association. Am J Human Genet 59:983–989
31. Sham PC (1997) The transmission/disequilibrium tests for multiallelic loci. Am J Human Genet 61:774–778
32. Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449
33. Cleves MA, Olson JM, Jacobs KB (1997) Exact transmission-disequilibrium tests with multiallelic markers. Genet Epidemiol 14:337–347
34. Morris AP, Whittaker JC, Curnow RN (1997) A likelihood ratio test for detecting patterns of disease-marker association. Ann Human Genet 61:335–350
35. Whittaker JC, Thompson DJ (1999) Finite-sample properties of family-based association tests. Am J Human Genet 64:910–915
36. Clayton DG (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Human Genet 65:1170–1177
37. Cordell H J, Barratt BJ, Clayton DG (2004) Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol 26:167–185
38. Harley JB, Moser KL, Neas BR (1995) Logistic transmission modeling of simulated data. Genet Epidemiol 12:607–612
39. Rice JP, Neuman RJ, Hoshaw SL, Daw EW, Gu C (1995) TDT with covariates and genomic screens with mod scores: their behavior on simulated data. Genet Epidemiol 12:659–664
40. Waldman ID, Robinson BF, Rowe DC (1999) A logistic regression based extension of the TDT for continuous and categorical traits. Ann Human Genet 63:329–340

41. Sinsheimer JS, Blangero J, Lange K (2000) Gamete competition models. Am J Human Genet 66:1168–1172
42. Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Human Genet 62:969–978
43. Weinberg CR (1999b) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Human Genet 65:229–235
44. Baksh MF, Balding DJ, Vyse TJ, Whittaker JC (2005) A likelihood ratio approach to family-based association studies with covariates. Annal Human Genet 70:131–139
45. Koeleman BPC, Dudbridge F, Cordell HJ, Todd JA (2000) Adaptation of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the Conditional Extended Transmission/Disequilibrium Test. Ann Human Genet 64:207–213
46. Schaid DJ (1999) Likelihoods and TDT for the case-parents design. Genet Epidemiol 16: 250–260
47. Clayton D, Jones H (1999) Transmission/disequilibrium tests for extended marker haplotypes. Am J Human Genet 65:1161–1169
48. Lunetta KL, Faraone SV, Biederman J, Laird NM (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am J Human Genet 66:605–614
49. Rabinowitz D, Laird NM (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Human Heredity 504:227–233
50. Sun FZ, Flanders WD, Yang QH, Khoury MJ (1999) The transmission disequilibrium test (TDT) when only one parent is available: The 1-TDT. Am J Epidemiol 150:97–104
51. Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: The pedigree disequilibrium test. Am J Human Genet 67:146–154
52. Lake SL, Blacker D, Laird NM (2000) Family-based tests of association in the presence of linkage. Am J Human Genet 67:1515–1525
53. Zhang S, Sha Q, Chen H, Dong J, Jiang R (2003) Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. Am J Human Genet 73:566–579
54. Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Human Genet 61:439–448
55. Wicks J (2000) Exploiting excess sharing: A more powerful test of linkage for affected sib pairs than the transmission/disequilibrium test. Am J Human Genet 66:2005–2008
56. Guo CY, Lunetta KL, DeStefano AL, Ordovas JM, Cupples LA (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
57. Whittaker JC, Lewis CM (1998) The effect of family structure on linkage tests using allelic association. Am J Human Genet 63:889–897
58. Siegmund KD, Gauderman WJ (2001) Association tests in nuclear families. Human Heredity 52:66–76
59. Martin ER, Bass MP, Hauser ER, Kaplan NL (2003b) Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Human Genet 73:1016–1026
60. Cordell HJ, Clayton DG (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 Human Genet 70:124–141
61. Cordell HJ (2004) Properties of Case/Pseudocontrol Analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. Genet Epidemiol 26:186–205
62. Siegmund KD, Langholz B, Kraft P, Thomas DC (2000) Testing linkage disequilibrium in sibships. Am J Human Genet 67:244–248
63. White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50: 1–25
64. Zou GY (2006) Statistical methods for the analysis of genetic association studies. Ann Human Genet 70:262–276

65. Millstein J, Siegmund KD, Conti DV, Gauderman WJ (2005) Testing association and linkage using affected-sib-parent study designs. Genet Epidemiol 29:225–233
66. Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Human Genet 61:319–333
67. Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Human Genet 56:811–812
68. Spielman RS, Ewens WJ (1999) TDT clarification. Am J Human Genet 64:668–668
69. Knapp M (1999a) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. Am J Human Genet 64:861–870
70. Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Human Genet 62:450–458
71. Knapp M (1999b) Using exact p-values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib transmission/disequilibrium test. Am J Human Genet 65:1208–1210
72. Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. Am J Human Genet 63:1507–1516
73. Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Human Genet 62:950–961
74. Schaid DJ, Rowland C (1998) Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Human Genet 63:1492–1506
75. Laird NM, Blacker D, Wilcox M (1998) The sib transmission/disequilibrium test is a Mantel-Haenszel test. Am J Human Genet 63:1915–1916
76. Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Human Genet 63:1886–1897
77. Sun FZ, Flanders WD, Yang Q, Khoury MJ (1998) A new method for estimating the risk ratio in studies using case-parental control design. Am J Epidemiol 148:902–909
78. Wang D, Sun F (2000) Sample sizes for the transmission disequilibrium tests: TDT, S-TDT and 1-TDT. Commun Statistic – Theory Meth 29:1129–1142
79. Sun FZ, Yang QH, Zhao HY, Flanders WD (2000) Transmission/disequilibrium tests for quantitative traits. Ann Human Genet 64:555–565
80. Schaid DJ, Li H (1997) Genotype relative-risks and association tests for nuclear families with missing parental data. Genet Epidemiol 14:1113–1118
81. Martin RB, Alda M, MacLean CJ (1998) Parental genotype reconstruction: applications of haplotype relative risk to incomplete parental data. Genet Epidemiol 15:471–490
82. Weinberg CR (1999a) Allowing for missing parents in genetic studies of case-parent triads. Am J Human Genet 64:1186–1193
83. Whittemore AS, Tu IP (2000) Detection of disease genes by use of family data. I. Likelihood based theory. Am J Human Genet 66:1328–1340
84. Shih MC, Whittemore AS (2002) Tests for genetic association using family data. Genet Epidemiol 22:128–145
85. Jonasdottir G, Humphreys K, Palmgren J (2007) Testing association in the presence of linkage - a powerful score for binary traits. Genet Epidemiol 31:528–540
86. Croiseau P, Genin E, Cordell HJ (2007) Dealing with missing data in family-based association studies: a multiple imputation approach. Human Heredity 63:229–238
87. Allen AS, Rathouz PJ, Glen A, Satten GA (2003) Informative missingness in genetic association studies: case-parent designs. Am J Human Genet 72:671–680
88. Chen Y (2004) New approach to association testing in case-parent designs under informative parental missingness. Genet Epidemiol 27:131–140
89. Schaid DJ, Sommer SS (1993) Genotype risk ratio: methods for design and analysis of candidate-gene association studies. Am J Human Genet 53:127–130
90. Sebastiani P, Abad MM, Alpargu G, Ramoni MF (2004) Robust transmission/disequilibrium test for incomplete family genotypes. Genetics 168:2329–2337
91. Allison DB (1997) Transmission disequilibrium tests for quantitative traits. Am J Human Genet 60:676–690

92. Allison D B, Neale MC (2001) Joint tests of linkage & association for quantitative traits. Theoretical Population Biol 60:239–251
93. Allison DB, Heo M, Kaplan N, Martin ER (1999) Sibling based tests of linkage and association for quantitative traits. Am J Human Genet 64:1754–1764
94. George V, Tiwari HK, Zhu X, Elston RC (1999) A test of Transmission/Disequilibrium for quantitative traits in pedigree data using multiple regression. Am J Human Genet 65:236–245
95. Zhu X, Elston RC (2001) Transmission/disequilibrium test for quantitative traits. Genet Epidemiol 20:57–74
96. Zhu X, Elston RC (2000) Power comparison of regression methods to test quantitative traits for association and linkage. Genet Epidemiol 18:322–330
97. Zhu X, Elston RC, Cooper RS (2001) Testing quantitative traits for association and linkage in the presence or absence of parental data. Human Heredity 51:183–191
98. Yang Q, Rabinowitz D, Isasi C, Shea S (2000) Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits. Human Heredity 50:227–233
99. Liu Y, Tritchler D, Bull SB (2002) A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. Genet Epidemiol 22:26–40
100. Kistner EO, Weinberg CR (2004) Method for using complete and incomplete trios to identify genes related to a quantitative trait. Genet Epidemiol 27:33–42
101. Kistner EO, Weinberg CR (2005) A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents. Genet Epidemiol 29:155–165
102. Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Human Genet 64:259–267
103. Cardon LR (2000) A sib-pair regression model of linkage disequilibrium for quantitative traits. Human Heredity 50:350–358
104. Abecasis GR, Cardon LR, Cookson WO (2000a) A general test of association for quantitative traits in nuclear families. Am J Human Genet 66:279–292
105. Abecasis GR, Cookson WOC, Cardon LR (2000b) Pedigree tests of transmission disequilibrium. Euro J Human Genet 8:545–551
106. Chen W, Abecasis GR (2007) Family-based association tests for genome wide association scans. Am J Human Genet 81:913–926
107. Purcell S, Sham P, Daly MJ (2005) Parental phenotypes in family-based association analysis. Am J Human Genet 76:249–259
108. Diao G, Lin DY (2006) Improving the power of association tests for quantitative traits in family studies. Genet Epidemiol 30:301–313
109. Box GEP, Cox DR. 1964. An analysis of transformations. J Roy Stat Soc, Series B 26: 211–246
110. Xiong MM, Krushkal J, Boerwinkle E. 1998. TDT statistics for mapping quantitative trait loci. Ann Human Genet 62:431–452
111. Fan RZ, Floros J, Xiong MM (2002) Models and tests of linkage and association studies of quantitative trait locus for multi-allele marker loci. Human Heredity 53:130–145
112. Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Human Heredity 47:342–350
113. Monks SA, Kaplan NL (2000) Removing the sampling restrictions from family-based test of association for a quantitative trait locus. Am J Human Genet 66:576–592
114. Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Euro J Human Genet 9:291–300
115. Chapman JM, Copper JD, Todd JA, Clayton DG (2003) Detecting disease association due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Human Heredity 56:18–31
116. Roeder K, Bacanu S, Sonpar V, Zhang X, Devlin B (2005) Analysis of single-locus tests to detect gene/disease associations. Genet Epidemiol 28:207–219
117. Xiong M, Zhao J, Boerwinkle E (2002) Generalized T^2 test for genome association studies. Am J Human Genet 70:1257–1268

118. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Human Heredity 53:79–91
119. Zhao H, Zhang S, Merikangas KR, Trixler M, Widenauer DB, Sun FZ, Kidd KK. 2000. Transmission/disequilibrium tests using multiple tightly linked markers. Am J Human Genet 67:936–946
120. Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233
121. Liang KY, Hsu FC, Beaty TH, Barnes KC (2001) Multipoint linkage-disequilibrium-mapping approach based on the case-parent trio design. Am J Human Genet 68:937–950
122. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
123. Hsu FC, Liang KY, Beaty TH, Barnes KC (2002) Unified sampling approach for multipoint linkage disequilibrium mapping of qualitative and quantitative traits. Genet Epidemiol 22: 298–312
124. Fan R, Xiong M (2002) Combined high resolution linkage and association mapping of quantitative trait loci. Euro J Human Genet 11:125–137
125. Fan RZ, Jung JS (2003) High-resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on sibship data. Human Heredity 56:166–187
126. Fan RZ, Spinka C, Jin L, Jung JS (2005b) Pedigree linkage disequilibrium mapping of quantitative trait loci. Euro J Human Genet 13:216–231
127. Fan RZ, Knapp M, Wjst M, Zhao CX, Xiong MM (2005a) High resolution T2 association tests of complex diseases based on family data. Ann Human Genet 69:187–208
128. Xu X, Tian L, Wei LJ (2003) Combining dependent tests for linkage or association across multiple phenotypic traits. Biostatistics 4:223–229
129. Xu X, Rakovski C, Xu X, Larid N (2006) An efficient family-based association test using multiple markers. Genet Epidemiol 30:620–626
130. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM (2003a) Using the noninformative families in family-based association tests: a powerful new testing strategy. Am J Human Genet 73:801–811
131. Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C (2005) Genomic screening and replication using the same data set in family-based association testing. Nat Genet 37:683–691
132. Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM (2007) A New multimarker test for family-Based association studies. Genet Epidemiol 31:9–17
133. Van der Meulen MA, te Meerman GJ (1997) Haplotype sharing analysis in affected individuals from nuclear families withy at least one affected offspring. Genet Epidemiol 14:915–919
134. Bourgain C, Genin E, Quesneville H, Clerget-Darproux F (2000) Search multifactorial disease susceptibility genes in founder populations. Ann Human Genet 64:255–265
135. Bourgain C, Genin E, Holopainen P, Mustalahti K, Maki M, Partanen J, Clerget-Darproux F (2001) Use of closely related affected individuals for the genetic study of complex disease in founder populations. Am J Human Genet 68:154–159
136. Bourgain C, Genin E, Ober C, Clerget-Darproux F (2002) Missing data in haplotype analysis: a study on the MILC method. Ann Human Genet 66:99–108
137. Lange EM, Boehnke M (2004) The haplotype runs test: the parent-parent-affected offspring trio design. Genet Epidemiol 27:118–130
138. Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25:115–121
139. Seltman H, Roeder K, Devlin B (2001) Transmission/Disequilibrium test meets measured haplotype analysis: Family-based association analysis guided by evolution of haplotypes. Am J Human Genet 68:1250–1263
140. Knapp M, Becker T (2003) Family-based association analysis with tightly linked markers. Human Heredity 56:2–9

141. Lin S, Chakravarti A, Cutler DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nature Genet 36:1181–1188
142. Knapp M, Becker T (2004) Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). Am J Human Genet 74:589–591
143. Sha QY, Dong JP, Jiang RF, Chen HS, Zhang SL (2005) Haplotype sharing transmission/disequilibrium tests that allow for genotyping errors. Genet Epidemiol 28:341–351
144. Toivonen HTT, Onkamo P, Vasko K, Ollikainen V, Sevon P, Mannila H, Herr M, Kere J (2000) Data mining applied to linkage disequilibrium mapping. Am J Human Genet 67:133–145
145. Zhang S, Sha Q, Chen H, Dong J, Jiang R (2004) Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT) - Reply. Am J Human Genet 74:591–593
146. Epstein MP, Kwee LC (2008) Haplotype Association Analysis. In: Shili Lin, Hongyu Zhao (eds). Handbook on analyzing human genetic data: computational approaches and software. In press.
147. Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol 26:61–69
148. Allen AS, Satten GA (2007) Inference on haplotype/disease association using parent affected-child data: the projection conditional on parental haplotypes method. Genetic Epidemiol 31:211–223
149. Allen-Brady K, Wong J, Camp NJ (2006) PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinformatics 7:209
150. Dudbridge F, Koeleman BP, Todd JA, Clayton DG (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Human Genet 66:2009–2012
151. Li C, Boehnke M (2006) Haplotype association analysis for late onset diseases using nuclear family data. Genet Epidemiol 30:220–230
152. Lo S, Zheng T (2002) Backward haplotype transmission association (BHTA) algorithm - a fast multiple-marker screening method. Human Heredity 53:197–215
153. Yu K, Gu CC, Province M, Xiong CJ, Rao DC (2004) Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet Epidemiol 27:182–191
154. Yu K, Xu J, Rao DC, Province M (2005a) Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. Ann Human Genet 69:577–589
155. Yu K, Zhang SL, Borecki I, Kraja A, Xiong CJ, Myers R, Province M (2005b) A haplotype similarity based transmission/disequilibrium test under founder heterogeneity. Ann Human Genet 69:455–467
156. Zhang S, Zhang K, Li J, Zhao H (2002) On a family-based haplotype pattern mining method for linkage disequilibrium mapping. Proc Pacific Symp Biocomputing 7:100–111
157. Martin ER, Bass MP,Gilbert JR, Pericak-Vance MA, Hauser ER (2003a) Genotype-based association test for general pedigrees: The genotype-PDT. Genet Epidemiol 25:203–213
158. Martin ER, Bass MP, Kaplan NL (2001) Correcting for a potential bias in the pedigree disequilibrium test. Am J Human Genet 68:1065–1067
159. Liu X, Gordon D (2003) A general class of association tests for family-based data using weight functions. Genet Epidemiol 24:208–219
160. Cantor RM, Chen GK, Pajukanta P, Lange K (2005) Association testing in a linked region using large pedigrees. Am J Human Genet 76:538–542
161. Xiong M, Jin L (2000) Combined linkage and linkage disequilibrium mapping for genome screens. Genetic Epidemiol 19:211–234
162. Umbach DM, Weinberg CR (2000) The use of case-parent triads to study joint effects of genotype and exposure. Am J Human Genet 66:251–261
163. Hsu FC, Liang KY, Beaty TH (2003) Multipoint linkage disequilibrium mapping approach: incorporating evidence of linkage and linkage disequilibrium from unlinked region. Genet Epidemiol 25:1–13

164. Lake SL, Laird NM (2004) Tests of gene-environment interaction for case-parent triads with general environmental exposures. Ann Human Genet 68:55–64
165. Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH (2006) A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. Genet Epidemiol 30:111–123
166. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Human Genet 69:138–147
167. Lange C, Blacker D, Laird NM (2004a) Family-based association tests for survival and times-to-onset analysis. Stat Med 23:179–189
168. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM (2004b) PBAT: tools for family-based association studies Am J Human Genet 74:367–369
169. Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. Am J Human Genet 66:1616–1630
170. Purcell S, Cherny SS, Sham PC (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19:149–150
171. Hoffmann T, Lange C (2006) P2BAT: a massive parallel implementation of PBAT for genome-wide association studies in R. Bioinformatics 15:3103–3105
172. Lange C, Laird NM (2002a) Analytical sample size and power calculations for a general class of family-based association tests: dichotomous traits. Am J Human Genet 71:575–584
173. Lange C, Laird NM (2002b) On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. Genet Epidemiol 23:165–180
174. Van Steen K, Lange C (2005) PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. Human Genom 2:67–69
175. Gauderman WJ (2002a) Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 21:35–50
176. Gauderman WJ (2002b) Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 155:478–484
177. Gauderman WJ (2003) Candidate gene association studies for a quantitative trait, using parent-offspring trios. Genet Epidemiol 25:327–338
178. Brown BW (2004) Power calculations for the transmission/disequilibrium and affected sib pair tests using elementary probability methods. Genet Res 83:133–141
179. Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data Am J Human Genet 69:371–380
180. Gordon G, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. Euro J Human Genet 12:752–761
181. Zhang K, Sun F, Zhao H (2005) HAPLORE: A Program for Haplotype Reconstruction in General Pedigrees without Recombination. Bioinformatics 21:90–103
182. Falls JG, Pulford DJ, Wylie AA, Jirtle RL (1999) Genomic imprinting: implications for human disease. Am J Pathol 154:635–647
183. Morison IM, Paton CJ, Cleverley SD (2001) The imprinted gene and parent-of-origin effect database. Nucleic Acids Res 29:275–276
184. Whittaker JC, Gharani N, Hindmarsh P, McCarthy MI (2003) Estimation and testing of parent-of-origin effects for quantitative traits. Am J Human Genet 72:1035–1039
185. Hu YQ, Zhou JY, Sun F, Fung WK (2007) The Transmission Disequilibrium Test and imprinting effects test based on Case-Parent Pairs. Genet Epidemiol 31:273–287
186. Thomas D, Xie R, Gebregziabher M (2004) Two-stage sampling designs for gene association studies. Genet Epidemiol 27:401–414
187. Feng T, Zhang S, Sha Q (2007) Two-stage association tests for genome-wide association studies based on family data with arbitrary family structure. Euro J Human Genet 15:1169–1175

188. Van Steen K, Laird NM, Markel P, Molenberghs G (2007) Approaches to handling incomplete data in family-based association testing. Ann Human Genet 71:141–151
189. Whittaker JC, Morris AP (2001) Family-based tests of association and/or linkage. Annal Human Genet 65:407–419
190. Zhao H (2000) Family based association studies. Stat Meth Medical Res 9:563–587
191. Ho GY, Bailey-Wilson JE (2000) The transmission/disequilibrium test for linkage on the X chromosome. Am J Human Genet 66:1158–1160
192. Horvath S, Laird NM, Knapp M (2000) The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. Am J Human Genet 66:1161–1167
193. Ghosh S, Reich T (2004). The Sib TDT adjusted for age of disease onset. Ann Human Genet 68:249–256
194. Jiang HY, Harrington D, Raby BA, Bertram L, Blacker D, Weiss ST, Lange C (2006) Family-based association test for time-to-onset data with time-dependent differences between the hazard functions. Genet Epidemiol 30:124–132
195. Mokliatchouk O, Blacker D, Rabinowitz D (2001) Association tests for traits with variable age at onset. Human Heredity 51:46–53
196. Epstein MP, Veal CD, Trembath RC, Barker JNWN, Li C, Satten GA (2005) Genetic association analysis using data from triads and unrelated subjects. Am J Human Genet 76:592–608
197. Kazeem GR, Farrall M (2005) Integrating case-control and TDT studies. Ann Human Genet 69:329–335
198. Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG (2004) Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression Euro J Human Genet 12:964–970
199. Lange C, Silverman EK, Xu X, Weiss ST, Laird NM (2003c) A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics 4:195–206
200. Lange C, Lyon H, DeMeo D, Raby B, Silverman EK, Weiss ST (2003b) A new powerful non-parametric two-Stage approach for testing multiple phenotypes in family-based association studies. Human Heredity 56:10–17
201. Cervino ACL, Hill AVS (2000) Comparison of tests for association and linkage in incomplete families. Am J Human Genet 67:120–132
202. Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Human Genet 60:691–702
203. Li Z, Gastwirth JL, Gail MH (2005) Power and related statistical properties of conditional likelihood score tests for association studies in nuclear families with parental genotypes. Ann Human Genet 69:296–314
204. Nicodemus KK, Luna A, Shugart YY (2007) An evaluation of power and type I error of single-nucleotide polymorphism transmission/disequilibrium-based statistical methods under different family structures, missing parental data, and population stratification. Am J Human Genet 80:178–185
205. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Human Genet 65:220–228
206. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Human Genet 67:170–181
207. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208
208. Abecasis GR, Cherny SS, Cardon LR (2001) The impact of genotyping error on family-based analysis of quantitative traits. Euro J Human Genet 9:130–134
209. Mitchell AA, Cutler DJ, Chakravarti A (2003) Undetected genotyping errors cause apparent over transmission of common alleles in the transmission/disequilibrium test. Am J Human Genet 72:598–610
210. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 33:228–237

211. Curtis (1999) Combining the sibling disequilibrium test and Transmission/Disequilibrium test for multiallelic markers. Am J Human Genet 64:1785–1786
212. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
213. Lange C, DeMeo DL, Laird NM (2002) Power and design considerations for a general class of family based association tests: quantitative traits. Am J Human Genet 71:1330–1341
214. O'Connell JR (2000) Zero-recombinant haplotyping: applications to fine mapping using SNPs. Genet Epidemiol 19:S64–S70
215. Onkamo P, Ollikainen V, Sevon P, Toivonen HT, Mannila H, Kere J (2002) Association analysis for quantitative traits by data mining: QHPM. Ann Human Genet 66:419–429

Haplotype Association Analysis

Michael P. Epstein and Lydia C. Kwee

Abstract Haplotypes serve many useful roles in the design and implementation of genetic studies of complex traits. In this chapter, we focus on the use of haplotypes as variables of interest for detecting association between a genomic region and a complex trait. Such haplotype analyses are appealing because, in certain instances, they can be more powerful for association mapping compared to traditional methods based on the analysis of individual SNPs. At the same time, haplotype analyses are more complicated to implement than single-SNP analyses since the sample genetic data often consist of unphased genotypes (which often lead to haplotype ambiguity). However, statisticians have developed many innovative methods for haplotype analysis that accommodate such haplotype ambiguity using existing missing-data algorithms. In this section, we describe a variety of such statistical methods for haplotype mapping, which are applicable to genetic datasets collected under traditional population-based and family-based study designs. We further describe software packages that are publicly available for implementing these haplotype approaches. Finally, we illustrate many of these statistical methods and related software packages using unphased genotype data from the Finland-United States Investigation of NIDDM Genetics (FUSION) study.

1 Introduction

A haplotype commonly refers to a set of alleles at tightly linked marker loci that are transmitted as a unit from parent to child (see Fig. 1). As noted in other chapters, haplotypes are valuable for genetic analysis as they help summarize genetic variation and linkage-disequilibrum (LD) patterns throughout the human genome [12].0pt In addition, haplotypes assist in the selection of single-nucleotide

M. P. Epstein (✉)
Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322,
e-mail: mpepste@emory.edu

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_8, © Springer-Verlag Berlin Heidelberg 2010

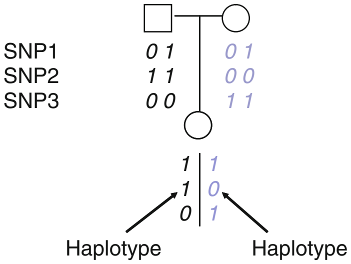


Fig. 1 Visual display of 3-SNP haplotypes in an offspring

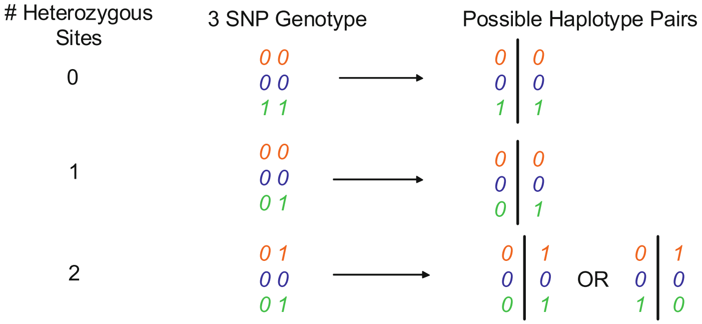


Fig. 2 Demonstration of potential haplotype ambiguity in 3-SNP (unphased) genotype data

polymorphisms (SNPs) that tag such genetic variation for association studies of complex traits [5, 62]. In this chapter, we consider another important role for haplotypes in genetic studies: their use as variables of interest for detecting association between a chromosomal region and a disease or trait of interest. As demonstrated by simulation studies [1, 40, 50], multi-marker association analyses based on haplotypes can be more powerful than single-marker association methods for mapping trait-influencing variants. Moreover, only haplotype-based methods inherently can model the trait-influencing effects of *cis*-acting variants, which have known to arise in disorders such as neural tube defects [25] and prostate cancer [64].

Although haplotype methods are valuable for genetic analysis, they are more complicated to implement compared to single-marker approaches due to the likely haplotype ambiguity within the observed genotype data. As shown in Fig. 1, a subject inherits two haplotypes (one of maternal descent and the other of paternal descent), and this haplotype pair automatically determines the person's multilocus genotype. However, the converse relationship that a multilocus genotype corresponds to a specific haplotype pair is false whenever the multilocus genotype is heterozygous at more than one locus (as demonstrated in Fig. 2). As the observed genetic data will often consist of only the multilocus genotype, the phase (allelic arrangement) of the underlying haplotypes of a subject may be unknown. To resolve the haplotype ambiguity, one could determine haplotypes directly using expensive molecular techniques [14, 17, 38]. However, a more cost-effective, yet still accurate, approach is to infer haplotypes from the observed genotype data by using one of the many existing missing-data algorithms [19, 41, 61, 76] . As we will show, we can use such missing-data algorithms in a variety of ways to construct valid tests of association between haplotypes and various phenotypes of interest.

In this chapter, we discuss methods and software for haplotype-based association analysis of both discrete and continuous phenotypes under a variety of study designs (including popular population-based and family-based designs for gene mapping). To help illustrate these methods, we will apply a subset of the approaches to unphased SNP genotype data from the Finland-United States Investigation of NIDDM Genetics (FUSION) study [68]. The remainder of this chapter proceeds as follows: we first provide a brief description of the FUSION study, followed by some general notation that we will use throughout this work. Next, we describe haplotype methods and software for analysis of unrelated subjects from cross-sectional, cohort, and case–control study designs. Finally, we describe similar haplotype methods and software for analysis of related samples from both case–parent triad and more general family-based study designs.

1.1 The FUSION Study

The FUSION study [68] is a long-term effort to identify susceptibility genes for type 2 diabetes and related quantitative traits. The study involves the phenotyping and genotyping of over 5,000 individuals living in Finland, utilizing a study design initially based on affected sib pairs for linkage analysis. Phenotypes include diabetes outcome, as well as diabetes-related quantitative traits such as fasting insulin, fasting glucose, body-mass index, high-density lipoprotein cholesterol, and blood pressure. The FUSION family-based sample consists of 737 familes ascertained based on an affected sib pair and has a total of 1,709 affected subjects [59]. In recent years, the FUSION study used this family-based sample to develop a case–control sample. For such case–control analysis, the FUSION study chose cases by selecting a single affected individual from each family and adding additional affected individuals from families excluded from linkage analysis for failing to have a genotyped affected sibling in the study. For controls, the FUSION study chose elderly subjects who had normal glucose tolerance at ages 65 and 70, and normoglycemic spouses of affected subjects. The FUSION case–control sample that we analyze here consists of 796 cases and 415 controls (225 elderly and 190 spouses).

Based on previous linkage and association analyses, the FUSION study has identified several regions linked to disease and quantitative traits [21, 59, 70]. Here, we focus haplotype analysis on five SNPs (distance between adjacent SNPs <300 kb) within a region along chromosome 22 that may harbor diabetes-susceptibility loci. Missing genotype rates for the five SNPs ranged between 2.4% and 5.6% with 17.1% of cases and 19.8% of controls missing genotype data at one or more SNP.

1.2 General Notation

We let Y denote the subject's phenotype of interest. Y can be a discrete outcome (e.g., 1 or 0 denoting the presence or absence of disease, respectively) or a continuous outcome (e.g., body-mass index). Also, we let E denote a subject's set of environmental covariates (e.g., age, gender) for the subject whom we wish to include in the genetic analysis of Y. For genetic data, we assume information from L biallelic SNPs within a chromosomal region of interest. Letting 0 and 1 denote the two alleles at each SNP, we can represent an L-SNP haplotype as a sequence of L numbers that take values of 0 or 1. For example, when $L = 2$, the possible 2-SNP haplotypes are 00, 01, 10, and 11. When $L = 3$, the possible 3-SNP haplotypes are 000, 001, 010, 011, 100, 101, 110, and 111. From these examples, it is straightforward to show that the total number of possible haplotypes for L SNPs is 2^L. We index these haplotypes by $k\,(k = 1, \ldots, 2^L)$ and let h_k denote the kth haplotype.

For a given study participant, we define $H = (h_k, h_{k'})$ as the subject's haplotype pair consisting of unordered haplotypes h_k and $h_{k'}$. Next, we define $G = h_k + h_{k'}$ as the subject's multi-SNP genotype, which we can represent as a sequence of L numbers that takes values of 0, 1, or 2. As mentioned earlier, while the haplotype-pair H determines the multi-SNP genotype G, the converse relationship is not necessarily true. Therefore, we let $S(G)$ denote the set of haplotype pairs $\{H = (h_k, h_{k'})\}$ consistent with G. We define $S(G)$ such that $(h_k, h_{k'}) \in S(G)$ implies that $(h_{k'}, h_k) \in S(G)$ when $h_k \neq h_{k'}$. It is important to note that G itself may include missing data (e.g., missing genotype at a specific SNP). In this situation, we can accommodate the missing data in G by including all haplotype pairs in $S(G)$ that are consistent with the known genotype information. We assume in this chapter that missing genotype data are missing at random, although we can relax this assumption in certain analyses [34].

2 Haplotype Analysis of Unrelated Samples

2.1 Cross-Sectional Studies

Under the cross-sectional study design, one collects phenotype, genotype, and covariate data from a random sample of n subjects from the population. We let Y_i, G_i, and E_i denote the phenotype, genotypes, and covariates, respectively, for the ith subject in the sample $(i = 1, \ldots, n)$. Further, we let H_i denote the haplotype pair of the subject.

2.1.1 Analyses Using Phased Haplotypes

Within the sample, the goal of the analysis is to assess the relationship between the haplotype H and phenotype Y, adjusting for the potential covariate effects in

E. If we knew the phase of the haplotypes for all subjects, we simply could conduct the association analysis using the popular generalized-linear-model (GLM) regression framework [36]. GLM analysis requires the construction of an appropriate likelihood that models the probability of the phenotype data Y conditional on the haplotype data H and the environmental data E within the sample. Assuming all subjects are unrelated (i.e., independent), we can write this likelihood as

$$L_{\text{OBS}} = \prod_{i=1}^{n} P\big[Y_i|H_i, E_i\big]. \tag{1}$$

The specific form of $P\big[Y_i|H_i, E_i\big]$ will depend on the distribution of Y. For continuous Y, $P\big[Y_i|H_i, E_i\big]$ often follows a probability-density function for a normal random variable with mean $\mu = E[Y|H, E]$ and variance σ^2. For binary Y, $P\big[Y_i|H_i, E_i\big]$ follows a probability-density function for a Bernoulli random variable with mean $\mu = E[Y|H, E] = P[Y = 1|H, E]$.

To assess the effects of H and E on Y, the GLM framework relates the mean $\mu = E[Y|H, E]$ described in the previous paragraph to a linear predictor of effects due to H and E. We can express this relationship using the following link function:

$$g(\mu) = \alpha + X_H \cdot \beta + X_E \cdot \gamma. \tag{2}$$

Here, X_H denotes a design vector that models the effects of a subject's haplotype pair H on μ and β denotes the related vector of regression coefficients. Likewise, X_E denotes a design vector for modeling the subject's environmental effects with respective coefficient vector γ. Finally, α denotes a scalar intercept parameter.

The choice of the link function $g(\cdot)$ depends on the distribution of the phenotype Y. For a continuous (and normally distributed) outcome, we typically apply the identity link $g(\mu) = \mu$ such that the resulting analysis is analogous to multiple linear regression. For a binary outcome, we typically assume the logistic link $g(\mu) = \log\big[\mu/\left(1 - \mu\right)\big]$, which leads to a logistic-regression analysis. In either scenario, we can then use the relationship in (2) to rewrite the likelihood in (1) as a function of the unknown parameters of interest (α, β, γ). We can then maximize the likelihood in (1) with respect to these parameters using standard maximum-likelihood procedures. After estimation, we can then construct test statistics with particular interest on assessing the effects of the haplotype-related parameters β on the phenotype Y, adjusting for the effects of environmental covariates. We can test null hypotheses regarding haplotype–phenotype associations by considering tests of the form $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ using appropriate likelihood-based test statistics (e.g., likelihood-ratio statistics, Wald statistics, or score statistics) that asymptotically follow a χ^2 distribution with degrees of freedom equivalent to the dimension of β.

Prior to haplotype analysis, we must specify the form of the haplotype design vector X_H in (2). In general, the form of X_H can be quite flexible. As an example, suppose we were interested in assessing the effects of a specific haplotype h^* relative to all remaining haplotypes. Then, X_H typically reduces to a scalar function whose form would depend on the assumed genetic mechanism for h^*. Define $I(A)$ as an indicator function that takes the value 1 or 0 depending on whether the event A is true or false, respectively. For a subject with $H = (h_k, h_{k'})$, we can model a recessive effect for h^* using $X_H = I(h_k = h_{k'} = h^*)$, a dominant effect for h^* using $X_H = I(h_k = h^*) + I(h_{k'} = h^*) - I(h_k = h_{k'} = h^*)$, and an additive effect for h^* using $X_H = I(h_k = h^*) + I(h_{k'} = h^*)$. We can also consider a co-dominant model for h^* by making X_H a two-dimensional vector with elements $[I(h_k = h^*) + I(h_{k'} = h^*) - I(h_k = h_{k'} = h^*), I(h_k = h_{k'} = h^*)]$.

While we considered the modeling of a specific target haplotype in the previous paragraph, we note that we can easily extend X_H to model the simultaneous effects of multiple haplotypes together and consider composite tests of haplotype–phenotype association. Such modeling simply requires the addition of the appropriate haplotype elements to the design vector. In theory, we can model the effects of all observed sample haplotypes within X_H, although we need one specific haplotype to serve as the baseline category.

2.1.2 Analyses Using Unphased Haplotypes

In the previous section, we made the unlikely assumption that we directly observed the phased haplotypes for all subjects. However, given data typically consist of unphased genotypes, we must consider a different observed-data likelihood for inference that allows for haplotype ambiguity. Here, we use a observed-data likelihood that we base on the joint probability of phenotype Y and genotype G conditional on environment E within the sample. We can write this likelihood as

$$L_{\text{OBS}} = \prod_{i=1}^{n} P\big[Y_i, G_i | E_i\big]. \tag{3}$$

We next express this likelihood in (3) as a function of the underlying haplotypes by writing $P\big[Y_i, G_i | E_i\big]$ as the sum of the haplotype pairs H_i consistent with G_i. Assuming that haplotypes are independent of environment within the sample, we can then write the likelihood as

$$L_{\text{OBS}} = \prod_{i=1}^{n} P\big[Y_i, G_i | E_i\big] = \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} P\big[Y_i | H_i, E_i\big] P\big[H_i\big]. \tag{4}$$

$P\big[Y_i | H_i, E_i\big]$ denotes the probability of the phenotype Y given haplotypes H and environment E, which we can model using the GLM procedure described in the previous section. $P[H_i]$ denotes the probability of the subject's haplotype pair H within the sample, which we can model in a variety of different ways. Letting $H_i = (h_k, h_{k'})$, we could model $P[H_i]$ under the popular assumption of Hardy–Weinberg Equilibrium (HWE) such that

$$P[H_i = (h_k, h_{k'})] = p_k p_{k'}, \tag{5}$$

where p_k denotes the frequency of the kth possible haplotype. While the HWE assumption typically holds in a cross-sectional study sample, substantial amounts of inbreeding or population stratification within the sample may cause Hardy–Weinberg departure (HWD) that leads to increased amounts of homozygosity or heterozygosity. In this situation, we can implement a model for $P[H_i]$ that allows for HWD by modifying the HWE model in (5) to include an additional parameter F (defined as a fixation index) that allows for excessive/reduced homozygosity within the sample. We write this HWD model for $P[H_i]$ as

$$P[H_i = (h_k, h_{k'})] = \begin{cases} p_k^2 + Fp_k(1 - p_k) & k = k' \\ (1 - F)p_k p_{k'}, & k \neq k'. \end{cases} \tag{6}$$

Comparison of the HWD model (6) with the HWE model (5) shows that an F value greater than 0 corresponds to excess homozygosity relative to HWE, while an F value less than 0 corresponds to excess heterozygosity.

On the basis of (2) and (5), it is straightforward to show that the observed-data likelihood L_{OBS} in (4) is a function of the earlier unknown parameters (α, β, γ) as well as the unknown haplotype frequencies $\boldsymbol{p} = \{p_k;\ k = 1, \ldots, 2^L\}$ and (if modeled) the fixation-index F. In the presence of haplotype ambiguity within the genotype data, we estimate these parameters by maximizing L_{OBS} in (4) indirectly using an Expectation–Maximization (EM) algorithm [11]. In this context, the EM algorithm obtains maximum-likelihood estimates of model parameters by iteratively maximizing the expectation of the natural log of the full-data likelihood $L_{\text{FULL}} = \prod_{i=1}^{n} P\big[Y_i, H_i | E_i\big]$, conditional on the observed data $\{Y_i, G_i, E_i; i = 1, \ldots, n\}$. For a more detailed description of the EM algorithm for cross-sectional studies, please see Appendix A of [33]. Once estimated, we can evaluate the variance–covariance matrix of the parameters using the formulas of [35] and [37], which account for the haplotype ambiguity within the observed genotype data.

After estimating the model parameters using the EM algorithm, we can use the fitted version of L_{OBS} in (4) to construct tests of haplotype effects on the phenotype Y $\big(H_0 : \beta = 0 \text{ vs. } H_A : \beta \neq 0\big)$. [55] developed score statistics for this purpose while [29] developed Wald statistics. In either situation, resulting statistics should asymptotically follow a χ^2 distribution with degrees of freedom equivalent to the dimension of β.

2.1.3 Stability Issues in Haplotype Analysis

Haplotype-association analysis using the likelihood framework in (4) requires the simultaneous estimation of the GLM model parameters (α, β, γ) as well as the haplotype-frequency parameters $\boldsymbol{p} = \{p_k;\ k = 1, \ldots, 2^L\}$. It is straightforward to show that the size of $\boldsymbol{p}$ increases exponentially with the number of SNPs within the haplotype analysis. As the number of haplotype-frequency parameters increases, there is an increased chance of numerical instability using the EM algorithm to maximize (4). To improve stability in this situation, one can instead conduct haplotype analysis under a “sliding window” design that sequentially examines smaller sets of SNPs within the region. For example, using a 4-SNP overlapping sliding-window, one would first conduct a haplotype analysis of SNPs 1–4, followed by SNPs 2– 5, followed by SNPs 3–6, and so on until the last SNP in the region is reached. We can then evaluate the empirical significance of the statistics using permutation procedures or efficient Monte-Carlo methods [24**?**].

2.1.4 Modeling Interaction Effects

In addition to examining main haplotype and environmental effects on a particular phenotype, interest may also focus on modeling and testing haplotype–environment interaction effects. Using the GLM framework, modeling such interaction effects is straightforward as it requires only the following alteration to the relationship shown in (2):

$$g(\mu) = \alpha + X_H \cdot \beta + X_E \cdot \gamma + X_{H \cdot E} \cdot \nu, \tag{7}$$

where $X_{H \cdot E}$ is a design vector that codes the haplotype–environment interaction (each element of the vector is generally the product of the respective elements of X_H and $X_E)$ with respective coefficient vector ν. Based on this relationship, the likelihood L_{OBS} in (4) is now a function of ν as well. We can then estimate this parameter using a variation of the EM algorithm described earlier and consider hypothesis tests of interaction effects using Wald statistics similar to those described in [29].

2.1.5 Haplotype Clustering

For rare haplotypes, the estimates of p_k and related effects in β for rare haplotypes often demonstrate large variability due to sampling variation and phase uncertainty, which potentially can lead to model instability [54] or invalid test statistics [18]. In addition, the modeling of rare haplotypes increases the number of model parameters in β, which increases the degrees of freedom of the resulting haplotype statistics that then leads to weakened power of global tests for detecting association with the phenotype.

To avoid these problems, one can pool rare haplotypes (defined by a frequency less than some threshold; typically 0.01–0.05) into a single haplotype category. While this resolves model instability, the resulting haplotype category is heterogeneous, which makes inference of its related effect on the trait difficult to interpret. A more appealing solution to this issue is to cluster rare haplotypes with their more common ancestral haplotypes using a model based on some evolutionary framework [16, 39, 58, 65]. Assuming haplotypes contained within each ancestral haplotype group have similar impact on the phenotype, one can use the haplotype clusters (rather than the individual haplotypes) to assess association with the phenotype.

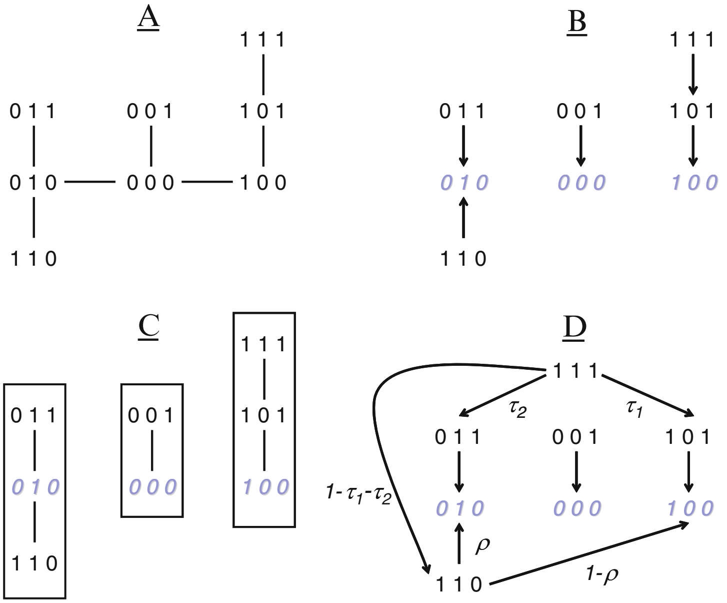


Fig. 3 (**a**) Cladogram of 3-SNP haplotypes. (**b**) Assignment of haplotypes to closest cluster base (shown in *grey shadow*). (**c**) Resulting three haplotype clusters from (**b**). (**d**) Haplotype clustering using a probabilistic algorithm

This haplotype clustering reduces the degrees of freedom in the resulting haplotype test and, therefore, should increase the power to detect trait-influencing variants.

We illustrate the process of haplotype clustering using an example based on [65] and detailed in Fig. 3. Within a particular genetic region, suppose we focus on three SNPs such that there are eight possible haplotypes consisting of $\{000, 001, 010, 100, 110, 101, 011, 111\}$. Suppose these haplotypes evolved as shown in the cladogram of Fig. 3a. Further, assume haplotypes 010, 000, and 100 are common enough such that they account for the majority of the haplotypes within the sample. We then let these three haplotypes form distinct cluster bases (shown in blue within Fig. 3b) and assign the remaining rarer haplotypes to their closest cluster base. This results in the formation of three haplotype clusters (Fig. 3c), which we can use as a surrogate for the eight distinct haplotypes within our haplotype analysis. As the number of parameters required for the haplotype clusters ((2), assuming one cluster serves as baseline) is smaller than the number required using distinct haplotypes (7), the former test statistic should be more powerful than the latter statistic for haplotype analysis.

Figure 3a–3c assume that the haplotype genealogy is unambiguous, which is unlikely in practice. Therefore, one instead uses a probabilistic approach for haplotype clustering, such as the one described in [65]. After determining the set of haplotype cluster bases $\mathcal{H}_{(0)}$, the approach identifies the set of haplotypes $\mathcal{H}_{(1)}$ that differ from $\mathcal{H}_{(0)}$ by one SNP allele (i.e., one mutation), the set of haplotypes $\mathcal{H}_{(2)}$

that differ from $\mathcal{H}_{(0)}$ by two SNP alleles (i.e., two mutations), and so forth until all haplotypes are assigned to a set $\mathcal{H}_{(j)}$ $(j = 0, \ldots, J)$. Starting with $\mathcal{H}_{(J)}$, we then assign each haplotype in the set to a particular (ancestral) haplotype in $\mathcal{H}_{(J-1)}$ with a specific probability that depends primarily on the estimated frequencies of the haplotypes in $\mathcal{H}_{(J-1)}$(offspring haplotypes are more likely derived from a common ancestral haplotype than a rarer one). We then assign each haplotype in $\mathcal{H}_{(J-1)}$ to a haplotype in $\mathcal{H}_{(J-2)}$ in similar fashion and repeat the process over and over until each haplotype in $\mathcal{H}_{(1)}$ is assigned to one of the haplotype cluster bases in $\mathcal{H}_{(0)}$. Figure 3d shows such a probabilistic model for cluster assignment, with haplotype 111 being assigned to haplotype 101 with probability τ_1, to haplotype 011 with probability τ_2, and to haplotype 110 with probability $1 - \tau_1 - \tau_2$. Haplotype 110, in turn, is assigned to haplotype cluster 010 with probability ρ and haplotype cluster 100 with probability $1 - \rho$. To determine cluster allocation for haplotypes that differ by >1 SNP, one simply takes the product of the relevant single-step allocation probabilities. For example, it is straightforward to show that haplotype 111 will be assigned to haplotype cluster 010 with probability $\tau_2 + (1 - \tau_1 - \tau_2) \cdot \rho$, haplotype cluster 100 with probability $\tau_1 + (1 - \tau_1 - \tau_2) \cdot (1 - \rho)$, and haplotype cluster 000 with probability 0.

Using this probabilistic algorithm, [67] proposed an association method using haplotype clusters based on a modified version of the GLM framework in (2). The approach replaces the haplotype-design vector X_H in (2) with a modified vector X_{CH} that models the clustered haplotypes (typically coded under an additive model). In particular, one can write $X_{\mathrm{CH}} = X_H B$, where B is an allocation matrix that probabilistically assigns each subject's pair of distinct haplotypes to the appropriate haplotype clusters (note that the elements of B are a function of the underlying haplotype frequencies). The resulting GLM model then takes the following form:

$$g(\mu) = \alpha + X_{CH} \cdot \beta_C + X_E \cdot \gamma, \tag{8}$$

where β_C denotes a vector of regression coefficients that model the haplotype–cluster effects. Relating this GLM model to the likelihood in (4), Tzeng et al. applied an EM algorithm for parameter estimation and then constructed score statistics for testing the null hypotheses of the form $H_0 : \beta_C = 0$ vs. $H_A : \beta_C \neq 0$. Such score statistics asymptotically follow a χ^2 distribution with degrees of freedom equivalent to the dimension of β_C. Extensions of this clustering framework to allow for interaction effects is currently under study (Tzeng, personal communication).

2.1.6 Software Packages

Many software packages exist for conducting a haplotype-based association analysis of subjects collected under a cross-sectional study. Many such packages run on common operating systems (OS), including Windows (2000 and XP), Macintosh OSX, Solaris, and Linux. Some packages may consist of a suite of routines that can be installed within a general analysis system like S-Plus or R, whereas other packages function as a self-contained executable. In the former category, [55] and [29] developed a suite of S-Plus/R routines for haplotype analysis called haplo.stats that, once installed, implements a haplotype-based association analysis within the GLM framework. The package actually consists of two separate procedures: haplo.score and haplo.glm. On the one hand, haplo.score constructs GLM-based score statistics [55] for testing global and individual haplotype effects on the outcome of interest (using either asymptotic or permutation-based p-values), adjusting for the effects of covariates. On the other hand, haplo.glm uses GLM regression to model, estimate, and test main haplotype and environmental effects, as well as haplotype–environment interaction effects. In addition, the software deals with rare haplotypes by pooling all such haplotypes (defined as those whose frequencies are below a user-specified threshold) into a single haplotype category. haplo.glm does not provide an option for permutation-based inference. However, one can obtain permutation-based p-values for the haplotype and environmental effects simply by creating a for loop within S-Plus or R that repeatedly permutes the outcome data (using the sample command).

Self-contained software executables also exist for haplotype analysis in cross-sectional studies. One such package is PLINK, which is a terminal-based application that runs on Windows, Macintosh OSX, and Linux platforms. PLINK shares many of the same features as haplo.stats; both software packages can model, estimate, and test the effects of haplotype and environment on the outcome of interest. PLINK also has an internal feature that allows for permutation-based testing of these various effects. Further, PLINK allows for haplotype analysis using sliding windows of various size. Many of these features are also implemented in the software package whap (Purcell et al. 2007b) , which was the predecessor to PLINK but is no longer supported.

HAPSTAT is another self-contained executable for haplotype analysis in cross-sectional studies. The software runs on Windows OS and utilizes a graphic-user-interface (GUI), rather than a terminal interface, to facilitate analysis using a "point-and-click" strategy. The application provides a flexible framework as it can model, estimate, and test haplotype effects (under additive, dominant, or recessive mechanisms), environmental effects, and haplotype–environmental interactions effects on traits of interest. Finally, the software has the appealing feature of allowing for HWD among haplotypes using the model shown in (6).

The above software packages model the effects of discrete haplotypes only. For analysis using haplotype clusters, one can use a set of R routines called Hapclustering [67] that initially identifies the haplotype cluster bases (using a modified version of the Shannon Information Criterion described in [65]) and then derives the allocation matrix B using estimated haplotype frequencies (calculated using an EM algorithm contained within the program). For both binary and continuous outcomes, the software provides score statistics and corresponding p-values for testing global hypotheses of haplotype effect on the trait. The routines also allow for the modeling of covariates, although they provide neither parameter estimates nor test statistics of the covariate effects. Also, the routines do not accommodate haplotype–environment interaction effects.

2.1.7 Software Application to FUSION Data

We applied a couple of the software packages described earlier to the data from the FUSION study consisting of five tightly linked SNPs found along chromosome 22. Here, we examined associations between the SNP-based haplotypes and a continuous outcome consisting of body-mass index (BMI). We note that the FUSION dataset is based on a case–control design rather than a cross-sectional design, which likely affects the distribution of BMI within the sample. In attempts to make the sample as close to a cross-sectional sample as possible, we analyze only the 415 control subjects from the study.

We first performed haplotype analyses on BMI in the FUSION sample using the haplo.score procedure [55] in the haplo.stats package. Using an EM algorithm, the procedure estimated 14 haplotypes with nonzero frequency within the FUSION sample. However, of these 14 haplotypes, only five had an estimated frequency greater than 0.05. We show the haplotypes and their respective frequencies below:

Haplotype	Frequency
00110	0.002
10011	0.357
10110	0.032
11111	0.002
01111	0.002
01011	0.129
01100	0.251
10000	0.014
11100	0.011
11011	0.139
10100	0.052
00011	0.004
01101	0.001
00100	0.004

Next, haplo.score constructed a global score statistic with 13 degrees of freedom for testing association between the haplotypes and BMI in the FUSION sample. The value of the score statistic was 17.07 ($p = 0.196$) indicating no significant association between the haplotypes within the region of interest and BMI. Inference using permutation procedures yielded similar results ($p = 0.189$ using 10,000 permutations of the data) .

Finally, haplo.score constructed separate score statistics for testing the effects of each of the 14 haplotypes on BMI. Using either asymptotic or permutation-based inference, results again suggest little evidence of association between the individual haplotypes and BMI assuming a Bonferroni-corrected p-value of 0.004.

Haplotype	Frequency	Asymptotic p-value	Permutation p-value
00110	0.002	0.173	0.132
10011	0.357	0.192	0.196
10110	0.032	0.296	0.292
11111	0.002	0.356	0.316
01111	0.002	0.836	0.826
01011	0.129	0.989	0.989
01100	0.251	0.920	0.921
10000	0.014	0.808	0.803
11100	0.011	0.486	0.462
11011	0.139	0.343	0.349
10100	0.052	0.274	0.262
00011	0.004	0.211	0.191
01101	0.001	0.171	0.130
00100	0.004	0.006	0.015

Since only five of the 14 haplotypes have an estimated sample frequency greater than 0.05, the global score statistic described earlier expends many degrees of freedom to model rare haplotypes in the sample, which could decrease power. Therefore, to reduce the dimension of the global haplotype association test with BMI, we next applied the haplotype-clustering procedure of [67] implemented in Hap-clustering to the FUSION data. Using the default settings within the program, the software clustered the 14 observed haplotypes shown above into six distinct cluster bases with respective frequencies

Haplotype cluster	Frequency
01011	0.132
01100	0.265
10011	0.360
10100	0.068
10110	0.033
11011	0.141

Note that these six haplotype clusters are based on the six most common haplotypes found in the FUSION sample.

As a result of the clustering, the global association test between the haplotypes and BMI now has only five degrees of freedom compared to the 13 degrees of freedom of the original global test. Nevertheless, the clustered global test yields a p-value of 0.439, which still indicates no significant evidence of association between the 5-SNP haplotypes and BMI in the FUSION sample. Subsequent examination of individual haplotype-cluster effects on BMI (shown below) further suggest that none of the haplotype clusters are associated with BMI

Haplotype cluster	p-value
01011	0.989
01100	0.556
10011	0.193
10100	0.249
10110	0.152
11011	0.432

2.2 Cohort Studies

Under a cohort study design, we assume that one collects a random sample of n subjects who are at risk for a particular outcome (e.g., disease) and then follows such subjects over time to determine the age-of-onset (AOO) for the outcome. Subjects who drop out of the study prior to completion or who fail to manifest the outcome by the end of the study will have censored outcomes, in that AOO is only known to be later than the censored time point. For the ith subject in the sample $(i = 1, \ldots, n)$, we let T_i denote AOO and let C_i denote the censoring time. The AOO phenotypic data is then $Y_i = \min(T_i, C_i)$ and $\Delta_i = I\left(T_i \leq C_i\right)$, with the latter datum being an indicator function that identifies whether the AOO observation is censored (equal to 0) or not (equal to 1). As described previously, we let G_i denote the subject's multi-SNP genotype and let E_i denote the subject's environmental covariates (which can be time-dependent).

Lin [31] used the popular proportional-hazards model [10] to assess associations between AOO and the underlying haplotypes H. This approach models the hazard function of AOO as a function of the haplotypes H and the environmental covariates E. We can write this hazard function as

$$\lambda\left(T_i = t | H_i, X_i\right) = \lambda_0\left(t\right) \exp\Big(X_H \cdot \beta + X_E(t) \cdot \gamma\Big), \tag{9}$$

where X_H denotes the haplotype design vector with log hazard-ratios β, $X_E(t)$ denotes the design vector for the (possibly time-varying) covariates, and $\lambda_0(t)$ denotes the arbitrary baseline hazard function. Note that the hazard function could also incorporate haplotype–environment interactions, if desired.

Based on this hazard function, one can write the likelihood of the observed data $\big\{Y_i, \Delta_i, G_i, E_i;\ i = 1, \ldots, n\big\}$ as [31, 33]

$$L_{\mathrm{OBS}} = \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} P[Y_i, \Delta_i | H_i, E_i] \cdot P[H_i],$$

$$= \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} \lambda\left(Y_i | H_i, X_i\right)^{\Delta_i} \exp\Big\{-\lambda(Y_i|H_i, E_i)\Big\} \cdot P[H_i]. \qquad (10)$$

Here, $\lambda(y|H_i, E_i) = \int_0^y \exp\Big(X_H \cdot \beta + X_E(t) \cdot \gamma\Big)\, \mathrm{d}\lambda_0(t)$, where $\lambda_0(t) = \int_0^t \lambda_0\,(s)\; \mathrm{d}s$. We can model the haplotype-pair frequencies $P[H_i]$ as shown previously in (5) and (6).

Using an EM algorithm described in [31], we can maximize L_{OBS} in (10) and obtain maximum-likelihood estimates (MLEs) of β and γ. We can estimate the variances of the MLEs using the observed information matrix or by a profile-likelihood approach described in [32]. For inference, we can easily construct likelihood-ratio statistics based on (10) or Wald statistics for testing $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$.

As an alternative to the cohort design, one can also implement a case–cohort design that studies subjects who have manifested the outcome of interest (e.g., disease) plus a random subcohort of subjects from the entire cohort. For rare outcomes, the efficiency of statistical inference using this case–cohort design is similar to that of the cohort design. However, by reducing the sample size and thereby reducing genotyping cost, the case–cohort design can be far more economical than the standard cohort design for genetic analysis. The likelihood for haplotype analysis in a case–cohort design is similar to L_{OBS} in (10), but includes an additional term to model the AOO data of the non-genotyped subjects within the cohort. For more details of the approach, please see [33] and [75].

2.2.1 Software Packages

The only software package currently available for haplotype analysis in both cohort and case–cohort designs is the previously described HAPSTAT package. The application allows for estimation and testing of haplotype effects (under additive, dominant, or recessive mechanisms), environmental effects, and haplotype–environmental interactions effects on censored outcomes.

2.3 Case–Control Studies

Under the popular case–control design, we assume that one collects genetic and environmental data from a sample of n subjects consisting of c controls and d cases $(c + d = n)$. For the ith subject $(i = 1, \ldots, n)$, we let Y_i denote disease outcome $(1 = \text{affected}, 0 = \text{unaffected})$, G_i denote the multi-SNP genotype, and E_i denote environmental covariates.

A simple approach for testing an overall association between haplotype H and disease Y is to apply an omnibus test that compares the estimated haplotype frequencies within the case and control samples [20, 79, 80]. A popular omnibus haplotype test first applies the EM algorithm to genotype data to estimate the haplotype frequencies within the control sample (denoted by $\boldsymbol{p}^{(0)}$), the case sample (denoted by $\boldsymbol{p}^{(1)}$), and the pooled sample of cases and controls (denoted by $\boldsymbol{p}^{(0+1)}$). Next, one constructs the observed-data likelihood of the genotypes within the controls only, the cases only, and the pooled sample. For controls, this likelihood takes the form $L_{\text{control}} = \prod_{i=1}^{c} P[G_i|Y_i = 0] = \prod_{i=1}^{c} \sum_{(h_k,h_{k'})\in S(G_i)} p_k^{(0)} \cdot p_{k'}^{(0)}$, where $p_k^{(0)}$ denotes the frequency of the kth haplotype in the controls. We define the likelihoods L_{case} and L_{pooled} in similar fashion using $\boldsymbol{p}^{(1)}$ and $\boldsymbol{p}^{(0+1)}$, respectively. Using these three likelihoods, we can then construct the omnibus test statistic as

$$S = 2\ \log_{\text{e}}\Big(\big(L_{\text{control}}\, L_{\text{case}}\big)/L_{\text{pooled}}\Big)$$

Under the null hypothesis of no association between any of the haplotypes and disease, the omnibus test S should asymptotically follow a χ^2 distribution with $K-\ 1$ degrees of freedom, where K denotes the number of observed haplotypes in the sample. In the likely occurrence of rare haplotypes, the asymptotic distribution of S may not hold and so analysts often rely on permutation-based approaches to assess significance of the omnibus statistic.

The omnibus test S suffers from two serious limitations. First, S cannot easily model the effects of any influential covariates E. Second, S does not provide inference on the effects of specific haplotypes or haplotype features. Such inference may be valuable, especially for identifying specific chromosomal segments that harbor disease-influencing variants [26]. Therefore, many analysts instead rely on more flexible approaches for case-control haplotype analysis. One common approach is to apply the logistic-regression framework previously described for cross-sectional studies to the case–control data. In this context, we consider the prospective likelihood

$$L_{\text{OBS}} = \prod_{i=1}^{n} P\big[Y_i, G_i|E_i\big] = \prod_{i=1}^{n} \sum_{H_i\in S(G_i)} P\big[Y_i|H_i, E_i\big]P\big[H_i\big], \tag{11}$$

where $P\big[H\big]$ denotes the frequency of haplotype pair H in the pooled sample of cases and controls. Typically, one models $P\big[H\big]$ under HWE using (5) or HWD using (6). Such a model for $P\big[H\big]$ignores the fact that the sample contains an oversampling of cases such that disease-susceptibilty haplotypes will be overrepresented in the data. While this may lead to bias in the estimation of haplotype effects, [63] noted that this situation arises only in practical situations when there is substantial haplotype ambiguity within the genotype data.

$P\big[Y|H, E\big]$ in (11) denotes the probability of the disease outcome conditional on haplotypes H and environment E, which can be modeled under a Bernoulli distribution with success probability $\mu = E[Y|H, E] = P[Y = 1|H, E]$. Letting $\theta_{H,E} = \frac{P[Y=1|H,E]}{P[Y=0|H,E]} = \frac{\mu}{1-\mu}$ denote the odds of disease given haplotype pair H and environmental covariates E, we can use the logistic link in GLM to write

$$\log_{\text{e}}\big(\theta_{H,E}\big) = \alpha + X_H \cdot \beta + X_E \cdot \gamma \tag{12}$$

Note that we can also amend (12) to model haplotype-environment interaction effects by adding the additional predictor vector $X_{H \cdot E}$ with corresponding regression coefficient ν.

By applying an EM algorithm to the likelihood in (11), [55] and [74] constructed statistics for testing either global or specific haplotype effects on disease, adjusting for environmental covariates. Such statistics asymptotically follow a χ^2 distribution with degrees of freedom equal to the dimension of β (for global tests). For testing global haplotype effects, [67] proposed to improve the power of the score statistic of [55] by using haplotype clusters. Within (12), Tzeng et al. replaced X_H with $X_{\mathrm{CH}} = X_H B$ (B is the allocation matrix described previously that probabilistically assigns discrete haplotypes to specific haplotype clusters) and replaced the disease-risk parameters β for discrete haplotypes with disease-risk parameters β_C for the haplotype clusters. In addition to testing, other investigators considered the estimation of haplotype effects on disease. Lake et al. [29] estimated main haplotype effects and haplotype–environment interactions using the prospective likelihood in (11), while [81] estimated these effects using a prospective estimating-equation framework that yields similar inference under a rare-disease assumption.

We refer to the likelihood in (11) as prospective because it models the probability of disease conditional on the haplotypes and environment. However, we note that this model does not reflect the manner by which one collects data in a case–control study. Rather, a case–control association study retrospectively samples genetic and environmental data conditional on disease status. Therefore, given the nature of case–control sampling, a retrospective likelihood that considers the probability of haplotype and environmental conditional on the disease outcome is likely more appropriate for case–control association analysis. However, most analysts still typically rely on the prospective likelihood in (11) for inference. This choice is due mainly to the theoretical work of [26], which demonstrated that general analysis of case–control data using the easy-to-implement prospective likelihood yields identical inference compared to analysis using the retrospective likelihood. However, this theoretical result only holds under the assumption of a saturated distribution for the model predictors. For case–control haplotype analysis, the critical assumption of [26] is violated because one cannot estimate a saturated haplotype-pair distribution from unphased genotype data. When the assumption is violated in general, [6] noted that retrospective analysis of case–control data can be much more powerful than prospective analysis. For haplotype analysis, [51] confirmed this finding using simulation studies that showed that retrospective approaches for case–control haplotype analysis generally are more powerful than prospective approaches for detecting main haplotype effects on disease. Kwee et al. [28] noted similar power improvements of retrospective methods over prospective methods for detecting haplotype–environment interaction effects.

A variety of retrospective likelihood and profile-likelihood approaches exist for investigating haplotype and haplotype–environment interaction effects on disease [18, 28, 32, 33, 51, 60]. Here, we consider the approach of [28] and construct the observed-data retrospective likelihood for haplotype analysis by considering the probability of the genotype data G and the environmental data E conditional on the disease data Y. We can then write the retrospective likelihood as a function of the underlying haplotype data H as

$$L_{\mathrm{OBS}} = \prod_{i=1}^{n} P[G_i, E_i|Y_i] = \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} P[H_i|E_i, Y_i] \cdot P[E_i|Y_i]. \tag{13}$$

To model $P[H_i|E_i, Y_i]$ in (13), we make the reasonable assumptions that the disease is rare (i.e., prevalence less than 0.10 in the target population) and that haplotype and environment are independent in the population. For control subjects, these two assumptions imply that $P[H_i|E_i, Y_i = 0] = P[H_i|Y_i = 0]$. We can model $P[H_i|Y_i = 0]$ under HWE using (5) or under HWD using (6).

For case subjects, we can write $P[H_i|E_i, Y_i = 1]$ as [52]

$$P[H_i|E_i, Y_i = 1] = \frac{\theta_{H_i,E_i} \cdot P[H_i|Y_i = 0]}{\sum_{H^*} \theta_{H^*,E_i} \cdot P[H^*|Y_i = 0]}. \tag{14}$$

Here, $\theta_{H,E}$ denotes the odds of disease given H and E, which we can model using the logistic model shown in (12). Note that the intercept parameter α and γ in (12) cancels from numerator and denominator of $P[H_i|E_i, Y_i = 1]$ in (14).

To complete the construction of the retrospective likelihood, we must develop an appropriate model for $P[E_i|Y_i]$ in (13). The derivation of such a model is challenging, particularly when E is either continuous or categorical with many possible outcomes. However, if we assume a saturated distribution for E in the population, we can use the work of [26] to rewrite $P[E_i|Y_i]$ as being simply proportional to $P[Y_i|E_i]$, which is much easier to specify. In particular, one can write this probability as

$$P[E_i|Y_i = y] \propto P[Y_i = y|E_i] = \frac{(\sum_{H^*} \theta_{H^*,E_i} P[H^*|Y_i = 0])^y}{1 + \sum_{H^*} \theta_{H^*,E_i} P[H^*|Y_i = 0]}. \tag{15}$$

Using (14) and (15), we can construct the retrospective observed-data likelihood in (13) and estimate (α,β, γ, ν) and the haplotype frequencies using a variation of the EM algorithm (see [28] for details).

2.3.1 Related Study Designs

Genetic association studies of complex disease sometimes employ specific variations of the case–control design for analysis. For example, some studies employ a finely stratified design that matches cases and controls on particular covariates (e.g., age, gender) to remove their confounding effects from the analyses. For such a matched case–control design, [27] developed a conditional logistic-regression framework for detecting associations between haplotypes and disease, but this procedure ignores the case–control sampling design in haplotype-frequency estimation and, as well, ignores the uncertainty of such frequency estimates within the association analysis. Zhang et al. [77, 78] proposed improvements to the procedure of Kraft et al., which effectively resolved these issues. However, to avoid dealing with the ascertainment bias in haplotype-frequency estimation, the authors performed estimation using genotype data from controls only. By not using all the genotype data available, this procedure is then likely inefficient relative to a retrospective procedure that uses genotype data from both controls and cases while simultaneously accommodating the sampling design within the analysis. Kwee et al. [28] recently developed such a retrospective approach for haplotype analysis of such matched data. Power and efficiency improvements of this novel approach relative to the approach of Zhang et al. still need to be investigated.

Researchers can also employ a case-only study design [42, 73] to investigate the effects of haplotype–environment interactions on disease. Under specific assumptions of a rare disease, independence between haplotype and environment in the population, and multiplicative effects of haplotypes on disease risk, we can show that a case-only analysis is as efficient for detecting interaction effects as a full retrospective case–control analysis, but at substantially reduced cost as no controls are required to be genotyped. Kwee et al. [28] developed a case-only approach for estimating and testing haplotype–environment effects by showing that a likelihood based on $P[G|E, Y = 1] = \sum_{H \in \mathrm{S}(G)} P[H|E, Y = 1]$ contains all essential information for testing the effects of interaction parameters. Using simulated data, the authors demonstrated that the case-only approach had near-identical power and efficiency for interaction effects relative to a full retrospective analysis of case–control data under multiplicative models of haplotype effect.

Lin and Zeng [32] noted that a more efficient approach for case-control haplotype analysis may be possible if one has knowledge of the total number of cases and controls within the population of interest [56]. Such information may be available from databases, such as hospital records or cancer registries. In this scenario, one can perform haplotype analysis by multiplying the observed-data prospective likelihood in (11) by a likelihood that models the outcome data of the cases and controls that were not selected and genotyped for analysis. Please see [33] and [32] for more details about analysis under this modified study design.

2.3.2 Haplotype Similarity Analyses

The majority of haplotype-based methods for association mapping of disease examines differences in haplotype frequencies among cases and controls within the chromosomal region of interest. While this is a valid manner by which to map susceptibility variants, it is not the only viable strategy. A second and increasingly popular approach for haplotype mapping of disease is based on the idea that, for a disease mutation of recent origin, a set of haplotypes from a sample of case subjects should share longer stretches of identical sequence around the disease locus compared to a set of haplotypes from a control sample [69]. This difference in the length of haplotype sharing between the two samples results from the shorter coalescence time of the recent mutation in the case sample relative to the normal allele in the control sample. As a result, we can use haplotype-based methods to identify chromosomal regions, where the average haplotype similarity among the pairwise combinations of case haplotypes is significantly greater than the average similarity among the pairwise combinations of control haplotypes.

Using the concepts described in [9] and [4], [66] developed an approach for disease mapping based on haplotype similarity. The approach requires the definition of a measure $M_{h_k h_{k'}}$, which quantifies the similarity between haplotypes h_k and $h_{k'}$. A variety of possible measures exists, including a matching measure that equals 1 if all SNP alleles of the two haplotypes are shared identical by state (IBS) and 0 otherwise. Also, one can implement a counting measure that equals the total number of SNP alleles of the two haplotypes shared IBS, as well as a length measure that equals the maximum number of adjacent SNP alleles of the two haplotypes shared IBS.

To illustrate these various similarity measures, suppose we consider two haplotypes that are each comprised of six SNPs. Let $h_k = 001011$ and $h_{k'} = 001111$. If we assume a matching measure, then $M_{h_k h_{k'}} = 0$, as the two haplotypes are not IBS at each SNP. If we assume a counting measure, then $M_{h_k h_{k'}} = 5$, as the two haplotypes share five SNP alleles IBS. Finally, if we assume a length measure, then $M_{h_k h_{k'}} = 3$ because the longest stretch of adjacent marker alleles shared IBS by the two haplotypes consist of SNPs 1–3 within the haplotype.

Given a specific similarity measure M, [66] constructed a nonparametric statistic that tested for excessive haplotype similarity in the case sample relative to the control sample. Let S_1 and S_0 denote the mean value of M in the case and control sample, respectively. We can write S_1 and S_0 as the sum of M over all pairwise haplotype combinations weighted by the frequency of the haplotype pair in the respective sample (with such haplotype frequencies estimated using a missing-data algorithm like the EM algorithm). Assuming that the frequencies of any two compared haplotypes within a specific sample are independent of one another (which is analogous to the HWE assumption), we can express S_1 and S_0 as

$$S_1 = \sum_{h_k} \sum_{h_{k'}} M_{h_k h_{k'}} \pi_k \pi_{k'} \qquad S_0 = \sum_{h_k} \sum_{h_{k'}} M_{h_k h_{k'}} \rho_k \rho_{k'}, \tag{16}$$

where π_k and ρ_k denote the estimated frequency of haplotype h_k in the case and control sample, respectively.

Once we evaluate S_1 and S_0, we construct the test of [66] as the normalized difference between the two values:

$$T = \frac{S_1 - S_0}{\sqrt{Var(S_1) + Var\,(S_0)}}. \tag{17}$$

We calculate Var(S_1) and Var(S_0) using the expressions in Appendix B of [66]. Under the null hypothesis of no difference in haplotype similarity between the two samples, T should follow a standard normal distribution. As a disease mutation should induce excessive similarity in cases relative to controls, one only typically examines the hypothesis $H_0 : S_0 = S_1$ vs. $H_A : S_0 < S_1$ for fine mapping and therefore constructs only one-sided tests for inference. For length and match similarity measures, the authors note that the variance of T fails to accommodate the extra variability due to unknown haplotype phase in genotype data (this extra variability is not an issue when using the counting measure for reasons discussed in Tzeng et al). Therefore, in this setting, the authors recommend establishing empirical significance of results using bootstrap sampling that randomly draws case and control haplotypes based on estimated haplotype frequencies from the pooled sample.

An appealing feature of the haplotype-similarity approach of Tzeng et al. is that the resulting test statistic has reduced degrees of freedom relative to statistics based on comparing haplotype frequencies among cases and controls (like the omnibus test). Using evolutionary simulations, [66] demonstrated that their haplotype-similarity statistic was often more powerful than a typical omnibus test, particularly in the likely situation where the disease mutation occurred on a common haplotype. However, this similarity approach does have some drawbacks for analysis: it can neither accommodate nor test the effects of environmental covariates and interactions.

2.3.3 Software Packages

For case–control haplotype analysis, we can use many of the software packages described previously for cross-sectional haplotype analysis of binary data. For omnibus tests of haplotype association with disease, one can easily apply PLINK or haplo.stats for inference using asymptotic-based or permutation-based procedures. In addition to omnibus tests, both PLINK and haplo.stats can also conduct general haplotype inference using the prospective likelihood in (11) and the logistic model in (12). Each software package can test and estimate the effects of specific haplotypes, environmental factors, and haplotype–environment interactions on disease. Using the same likelihood and framework, one can also perform haplotype-cluster analysis of case–control data using the R routines in Hap-clustering. As mentioned previously, this software package tests only global hypotheses of haplotype effect on disease risk. The routines also allow for the modeling (but not testing or estimation) of covariates.

As a retrospective likelihood will likely have improved power for detecting haplotype and haplotype–environment interaction effects on disease relative to a prospective likelihood, several software packages exist for implementing analyses based on the former type of likelihood. The previously described software package HAPSTAT uses a variation of a retrospective likelihood described in [33] and [32] for case–control haplotype inference that allows for covariates and permits testing of haplotype–environment interactions. Another useful software package is CHAPLIN, which utilizes the retrospective likelihood shown in [28] for inference. CHAPLIN is a self-contained executable that currently runs on various Windows, Linux, and Unix OS and will soon run on Macintosh OSX. CHAPLIN can run either via a graphic-user interface or a terminal-user interface (with the latter interface facilitating permutation-based analyses). Like HAPSTAT, CHAPLIN can model haplotype frequencies under HWD using a fixation index as shown in (6). The program can currently can accommodate only main haplotype effects, but will soon be extended to model and test both environmental effects and haplotype–environment interactions. In addition, the package will soon allow for haplotype analysis in case-only and matched case–control studies.

To implement a haplotype-similarity approach for analysis, one can use a set of R routines called QSHS (Quadratic Statistics of Haplotype Similarity) developed by [66]. The software constructs the statistic shown in (16) and (17) under length, counting, and matching measures of haplotype similarity and establishes the asymptotic significance of the results. The software also considers both full-dimensional and reduced-dimensional analyses (using a user-defined threshold for haplotype frequencies to include in analysis). One thing that QSHS does not do is estimate the necessary haplotype frequencies in the case and control samples necessary for haplotype-similarity analysis. A user must estimate these quantities separately (using independent software packages like PHASE), save them to a file, and then input them directly into the program. Another thing that QSHS does not automatically do is establish empirical significance of results based on the matching and length measures of haplotype similarity, as suggested in [66]. However, we can conduct a bootstrap analysis for this purpose in R by constructing a for loop that, within each of the many iterations,

1. Applies the sample command to randomly assign haplotypes based on the haplotype frequencies in the pooled sample to the case and control samples.
2. Estimates the (phase-known) haplotype frequencies in the bootstrap sample using gene-counting techniques.
3. Inputs these haplotype estimates into QSHS for subsequent analysis.
4. Saves the bootstrap values of the resulting QSHS test statistics into a vector or file.

2.3.4 Software Application to FUSION Data

We conducted a case–control haplotype analysis of the five SNPs from the FUSION dataset using a variety of the software packages available for this purpose. First, we constructed an omnibus goodness-of-fit statistic between the FUSION sample haplotypes and disease using the PLINK package. Using this software, we identified 17 haplotypes with nonzero frequency in the sample. PLINK then produced an omnibus statistic value of $S = 32.628$, which yields an asymptotic p-value of 0.008 (based on a χ^2 distribution with 16 degrees of freedom). For robust inference, we next established the empirical significance of the omnibus statistic using permutation procedures and obtained a robust p-value of 0.007, which closely matches the asymptotic result. Collectively, these results suggest a significant overall association between the 5-SNP haplotypes and disease within the FUSION sample.

We next used PLINK to construct individual association tests of each specific haplotype found in the FUSION sample against all other haplotypes. Such analyses assume the specific haplotype of interest has a log-additive effect on disease risk. We present the results of these individual haplotype tests from PLINK in the following table. We show p-values smaller than a Bonferroni-corrected significance level of 0.003 in bold.

Haplotype	Frequency	Asymptotic p-value
10011	0.311	**0.00042**
01100	0.296	**0.00034**
11011	0.133	0.362
01011	0.132	0.647
10100	0.057	0.602
10110	0.032	0.795
10000	0.012	0.683
11100	0.009	0.892
00011	0.006	0.588
00100	0.003	0.918
00100	0.003	0.0647
10010	0.001	0.266
00110	0.001	0.642
11110	0.001	0.241
01111	0.001	0.0662
11111	0.001	0.0576
01101	<0.001	0.145

Examination of the PLINK results reveal that only haplotypes 01100 and 10011 within the region are significantly associated with disease. We then conducted refined analyses on these two haplotypes using CHAPLIN. In particular, we considered models that regressed the joint effects of these two haplotypes on disease under different genetic models of haplotype effect. We then chose the best model to be the one that yielded the smallest value of the Akaike Information Criterion (AIC), which CHAPLIN provides. After examining a variety of models, we found the model that yielded the smallest AIC was the one that assumed a multiplicative effect of both haplotype 01100 and 10011. Based on this model, we show estimates and tests of the two haplotypes below.

Haplotype	$\widehat{\beta}$ (SE)	Asymptotic p-value (Wald statistic)
01100	0.239 (0.114)	0.037
10011	−0.216 (0.110)	0.049

These results suggest that haplotype 01100 is associated with an increased risk in disease while haplotype 10011 is a protective haplotype. Curiously, it is interesting to note that these two haplotypes share no SNP alleles in common. CHAPLIN also constructed a joint test of the effects of the two haplotypes and obtained a p-value of 0.0003 using either a likelihood-ratio test or a robust-score test.

Finally, we analyzed the FUSION data using the haplotype-similarity approach implemented in QSHS. We considered matching, length, and counting measures of haplotype similarity within the analysis. We show results of the analyses below.

Similarity measure	T statistic	Asymptotic p-value
Matching	−0.350	0.637
Length	−3.316	1.0
Counting	−4.491	1.0

Somewhat surprisingly, these results suggest no evidence of association between the genetic region of interest and disease, which conflicts with the earlier significant results from PLINK and CHAPLIN. However, examination of the T statistics under length and counting measures suggest that there is increased haplotype similarity in controls compared to cases (note that the haplotype-similarity tests are one-sided tests and only look for excessive haplotype similarity in cases compared to controls). While this result is unusual, it could be explained by the presence of a recent protective allele of recent origin that originated in control subjects.

3 Haplotype Analysis of Family-based Samples

In addition to using unrelated samples for association mapping of complex traits, investigators may also use family-based study designs for such analyses. Implementation of a family-based design might be attractive for economic reasons (e.g., using families previously collected for a linkage study). More importantly, family-based association studies are attractive as resulting tests of association between SNPs and phenotype are typically robust to the confounding effects of population stratification. Association tests in unrelated samples are susceptible to such confounding (resulting in inflated type-I error and bias) without some additional corrections (see [7, 13, 44]).

The most common family-based study design collects units defined as a case–parent triad, which consist of an affected proband and the proband's parents. However, studies can also consider nuclear pedigrees for analysis, such as those used in linkage studies. Such designs could collect samples consisting of affected proband(s) and a various number of the proband's relatives (including parents as well as affected and unaffected siblings). One can also collect and analyze more distant relatives like grandparents, although we will not discuss such family designs here (see [71] for more details). For case–parent triads and nuclear pedigrees, statistics for testing association generally are based on a framework that evaluates the distribution of offspring genotypes within a family conditional on the parental genotypes as well as the phenotypic data within the family. The use of such a framework is popular for evaluating genotype risks because parental allele frequencies and familial phenotypes are sufficient statistics for nuisance parameters corresponding to parameters related to population stratification and the trait distribution, respectively [23]. By conditioning on these sufficient statistics, the framework produces test statistics that are robust to population stratification and misspecification of the phenotype distribution.

Under a null hypothesis of no linkage and no association with the phenotype, the above framework suggests that the distribution of SNP alleles in offspring conditional on the parental genotypes follow Mendelian segregation patterns independent of the phenotypic data within the family [30]. Given this result, we would ideally construct a family-based test for haplotype analysis by considering an approach that compares the observed segregation patterns of haplotypes within offspring to the expected segregation patterns under the null hypothesis of no linkage and no association. In essence, this ideally would require us to develop a test of the offspring haplotype data that is conditional on the parental haplotype data and the phenotypic trait data of the family. However, in practice, such tests must deal with the likely haplotype ambiguity within the familial genotype data. While missing haplotype data can be inferred probabilistically from the sample of observed genotypes (using a missing-data algorithm like the EM algorithm), such inference is sensitive to misspecification of the haplotype distribution. Such misspecification can arise in the presence of population stratification and hence may lead to invalid results (which defeats the primary rationale behind using a family-based design for gene mapping). Valid test statistics for haplotype analysis must be able to circumvent this issue. In addition, such test statistics must also be able to handle missing parental genotype data; a common occurrence in family-based association studies.

While a variety of statistical methods exist for family-based association mapping using haplotypes [8, 15, 79, 80], few approaches are both generally robust to population stratification and able to accommodate missing genotype data. We describe two such approaches in the remainder of this section. First, we describe the haplotype approach of [23], which is applicable for association mapping within nuclear pedigrees. Second, we describe a robust approach recently developed by [2] for haplotype and haplotype–environment interaction analysis within case–parent triads. After these descriptions, we then discuss the software packages that implement these two approaches.

3.1 Haplotype Approach of Horvath et al. [23]

We assume a collection of n nuclear pedigrees. For the ith pedigree ($i = 1, \dots n$) that consists of two parents and n_i offspring, we let $Y_{P,i} = \left(Y_{P,1,i},\, Y_{P,2,i}\right)$ denote a trait vector for parents and let $Y_{O,i} = \left(Y_{O,1,i}, Y_{O,2,i}, \dots, Y_{O,n_i,i}\right)$ denote a trait vector for the offspring. Similarly, we define $G_{P,i}\ \left(H_{P,i}\right)$ and $G_{O,i}\ \left(H_{O,i}\right)$ as vectors of genotypes (haplotype pairs) for the parents and offspring, respectively, in the

ith pedigree. Although not explicitly shown, we note that we can generalize these phenotypic and genotypic vectors to accommodate missing parental data.

The family-based haplotype association test of [23] is an extension of the popular collection of family-based association tests (FBATs) originally developed in [49] and [22]. In general, the FBAT approach constructs a test statistic under the null hypothesis of no linkage and no association that is based on the distribution of $G_{O,i}$ conditional on sufficient statistics for nuisance parameters that could potentially invalidate inference (e.g., population stratification, misspecification of distribution for continuous phenotypes). For haplotype analysis, [23] used sufficient statistics based on the genotype-based FBAT, which (under the likely scenario of missing parental data) consists of all observed genotype and trait data within the family. Using this finding, the authors then developed a conditioning algorithm that, for the ith pedigree, determines the set $\mathcal{G}_{O,i}$ of possible offspring genotype vectors that have the same value of the minimal-sufficient statistic as the observed genotype vector. Then, for each genotype vector $G^*_{O,i} \in \mathcal{G}_{O,i}$, one determines the vector's probability conditional on $\mathcal{G}_{O,i}$ and the minimal-sufficient statistic. Under the null hypothesis, this conditional probability is a simple function of Mendelian proportions.

For the ith pedigree, [23] next defined a function S_i of the offspring genotypes $G_{O,i}$ and phenotypes $Y_{O,i}$ to be used in the evaluation of a test statistic for haplotype–disease association. In their paper, the authors initially chose the function to be $S_i = \sum_{j=1}^{n_i} Y_{O,j,i} \cdot X_{G_{0,j,i}}$, where $X_{G_{0,j,i}}$ is a vector that denotes the specific coding of allelic effects for the genotype of the jth offspring. Since the interest on modeling and testing haplotype effects, the authors then rewrote S_i to be a function of the underlying haplotype data as follows:

$$S_i = \sum_{j=1}^{n_i} Y_{O,j,i} \cdot \left(\sum_{H_{O,j,i} \in G_{O,j,i}} X_{H_{0,j,i}} \cdot \frac{P[H_{O,j,i}]}{\sum_{H^*_{O,j,i} \in G_{O,j,i}} P[H^*_{O,j,i}]} \right), \tag{18}$$

where $X_{H_{0,j,i}}$ denotes the design vector of the haplotype effects for the jth offspring. We can model this design vector using similar approaches as described earlier for haplotype analysis using GLM models. Also, $P[H_{O,j,i}]$ denotes the frequency of the haplotype pair of the jth offspring. To estimate $P[H_{O,j,i}]$, the authors used haplotype frequencies estimated via an EM algorithm to the unphased genotype data across pedigrees. While the presence of population stratification in the sample could yield biased estimates of such haplotype frequencies, the authors noted that inference should still be valid under the null hypothesis of no linkage and no association as one estimates $P[H_{O,j,i}]$ conditional on the minimal-sufficient statistic consisting of the observed parental and offspring genotypes. The authors confirmed the validity of this assertion using simulations under a population-stratification model.

Under the null hypothesis, [23] then calculated the expected value $E[S_i]$ and variance Var $[S_i]$ of S_i using the conditional probability of offspring genotype vectors $G^*_{O,i} \in \mathcal{G}_{O,i}$ conditional on the minimal-sufficient statistic (described earlier). The authors then calculated the haplotype FBAT as a Mantel–Haenszel statistic of the form $T = U^{'}V^{-1}U$, where $U = \sum_i\{S_i - E[S_i]\}$ and $V = \sum_i Var(S_i)$, which asymptotically follows a χ^2 distribution with degrees of freedom equal to the rank of V.

Occasionally, interest may focus on a genetic region previously identified via a linkage study. In this situation, it may be more appropriate to use the haplotype FBAT to test a null hypothesis of linkage but no association. Under this particular null hypothesis, the linkage will induce additional correlation among offspring genotypes within the same pedigree. This additional correlation is not accommodated by the variance estimate shown above in V. To accommodate this phenomenon, [23] suggested the use of a robust variance estimator $V_{\mathrm{R}} = \sum_i\{S_i - E[S_i]\}\{S_i - E[S_i]\}$ and subsequently construct the haplotype FBAT statistic as $T_{\mathrm{R}} = U^{'}V_{\mathrm{R}}^{-1}U$.

3.2 Haplotype Approach of Allen and Satten [2]

For haplotype mapping of disease in case–parent triads, [2] developed an alternative method that appeared to have improved power relative to the FBAT approach of [23]. In addition, unlike FBAT, this proposed method permitted robust estimation of haplotype effects on disease. The authors based this method on the framework of [47, 48] and [72], who developed a valid family-based association test by ensuring a family's contribution to an asymptotically normal test statistic had a mean value of zero conditional on any pattern of parental genotype data. As a result, such an approach is robust to misspecification of the parental–genotype distribution and hence robust to population stratification. Allen et al. [3] extended the general approach of Rabinowitz to the area of haplotype analysis by identifying an efficient score function whose family-specific contributions to a test statistic under the null hypothesis had a mean value of zero conditional on any pattern of parental haplotype data. The approach of [2] described in this section is a further modification of [3] to accommodate missing parental genotype data and also to allow for the testing and estimation of interaction effects between haplotypes and environment.

Given the focus on case–parent triads, some of the notations used in the previous section for haplotype FBAT can be simplified for this section. For the ith triad $(i = 1, \ldots, n)$, $Y_{O,i}$, $G_{O,i}$, and $H_{O,i}$ are now scalar variables that denote the phenotype, genotype, and haplotype pair, respectively, of the lone offspring in the triad. On the basis of the triad sample design, it obviously follows that $Y_{O,i} = 1$ for each of the n triads. Further, to consider haplotype–environment interactions, we now additionally define a vector of environmental variables $E_{O,i}$ for the offspring of the ith triad. Such environmental variables can be either categorical or continuous in nature.

To examine haplotype and haplotype–environment interaction effects on disease, [2] used a log-linear model for the relative risk (RR) of disease in a triad offspring. Suppressing the triad index i, we can write the general RR model as

$$\log\left(\frac{P[Y_O = 1|H_O, E_O]}{P[Y_O = 1|H^*, E_O]}\right) = \alpha + X_{H_0} \cdot \beta + X_{E_0} \cdot \gamma + X_{H_0} \cdot X_{E_0} \cdot \nu, \quad (19)$$

where H^* denotes a baseline haplotype pair category. Within (19), X_{H_0} and X_{E_0} denote design vectors for haplotype and environmental effects, respectively, while (β, γ, ν) denote parameter vectors for haplotype, environment, and haplotype–environmental interaction effects.

Allen and Satten [2] performed inference on these parameters by constructing a likelihood of the genotype data in a triad conditional on the environmental exposure in the offspring. Again suppressing the triad index, we can write the likelihood contribution for a triad as

$$P[G_O, G_P|E_O, Y_O = 1]$$

$$= \sum_{H_O \in G_O} \sum_{H_P \in G_P} P[H_O|H_P, E_O, Y_O = 1] \cdot P[H_P|E_O, Y_O = 1]. \quad (20)$$

Using the RR model in (19), one can write $P[H_O|H_P, E_O, Y_O = 1]$ in (20) as

$$P[H_O|H_P, E_O, Y_O = 1] = \frac{\exp\Big(X_{H_0} \cdot \beta + X_{H_0} \cdot X_{E_0} \cdot \nu\Big) \cdot P[H_O|H_P, E_O]}{\sum_{H'_O} \exp\Big(X_{H'_0} \cdot \beta + X_{H'_0} \cdot X_{E_0} \cdot \nu\Big) \cdot P[H'_O|H_P, E_O]}, \quad (21)$$

where $P[H_O|H_P, E_O]$ denotes the probability of the offspring's haplotype pair conditional on the parental haplotypes and the offspring's environmental variable. Assuming no recombination within the haplotype region and further assuming that the environmental variables do not influence segregation, one can show that $P[H_O|H_P, E_O] = P[H_O|H_P]$ is a simple function of Mendelian transmission probabilities.

Also, we should note that this probability in (21) is not a function of main environmental effects γ (which cancel from both numerator and denominator of the probability). In general, estimates of γ are nonidentifiable in case–parent triad analyses [53, 57] and thus will not be considered further in this section. For simplicity, we define the set of estimable parameters as $\theta = (\beta, \nu)$.

$P[H_P|E_O, Y_O = 1]$ in (21) denotes the frequency of the parental haplotype pairs conditional on the offspring environmental variables in the triad. Specification of this probability is unappealing as it depends on a set of nuisance parameters (related to population stratification) that is likely difficult to model in practice. Worse, misspecification of this probability (i.e., not properly modeling the stratification) can yield biased inference [3]. Because of this, [2] developed estimating equations for θ that were unbiased for any possible distribution of the parental haplotypes and, hence, robust to stratification.

Allen and Satten [2] developed estimating equations based on the finding that

$$P[G_O, G_P|H_P, E_O, Y_O = 1] = \sum_{H_O \in S(G_O)} P[H_O|H_P, E_O, Y_O{=}1] \cdot I\big(H_P{\in}S(G_P)\big) \tag{22}$$

is independent of $P[H_P|E_O, Y_O = 1]$. Therefore, an estimating function $U_\theta(G_O, G_P)$ with the property

$$E\big[U_\theta(G_O, G_P)|H_P, E_O, Y_O = 1\big] = 0 \text{ for all } H_P \tag{23}$$

leads to $E\big[U_\theta(G_O, G_P)|E_O, Y_O = 1\big] = 0$. Such an estimating equation is then robust to misspecification of $P[H_P|E_O, Y_O = 1]$ and, hence, robust to population stratification. The goal then is to find an estimating function with the properties shown in (23).

To find an appropriate function $U_\theta(G_O, G_P)$, [2] first considered a matrix Ψ of conditional probabilities $P\big[G_O, G_P\,|H_P, E_O, Y_O = 1\big]$ with the rows indexing all possible combinations of offspring and parental genotype outcomes (including outcomes that denote missing data) and the columns indexing all possible outcomes of the parental haplotypes. Therefore, the (j, k)th element of Ψ corresponds to $P\big[(G_O, G_P) = j\,|H_P = k, E_O, Y_O = 1\big]$ (which can be evaluated using (21) and (22)).

Next, Allen and Satten [2] considered the vector $\mathcal{U}_\theta$ with jth element corresponding to $U_\theta\big((G_{O,i}, G_{P,i}) = j\big)$. The authors then noted that the required condition in (23) holds when $\mathcal{U}_\theta$ is orthogonal to the columns of Ψ. Therefore, the authors applied a projection approach to construct the estimating equation

$$\widetilde{\mathcal{U}}_\theta = \mathcal{U}_\theta - \Psi(\Psi'\Psi)^{-1}\Psi'\mathcal{U}_\theta,$$

which is the portion of $\mathcal{U}_\theta$ that is orthogonal to the columns of Ψ. Therefore, $\widetilde{\mathcal{U}}_\theta$ satisfies the necessary properties in (23) and can be used for inference that is robust to misspecification of parental haplotype frequencies that can arise due to stratification. Given the nature of their approach, the authors named the method the projection-conditional-on-parental-haplotypes (PCPH) approach for haplotype inference in case–parent triads.

Based on the above derivation, its important to note that the PCPH approach will lead to valid inference and unbiased estimators of θ regardless of the choice of $U_\theta(G_O, G_P)$ used in $\mathcal{U}_\theta$. However, the choice of the estimating equation can affect the efficiency and power of PCPH under alternative models. Allen and Satten [2] noted that the optimal choice for $U_\theta(G_O, G_P)$ is the score vector $S_\theta(G_O, G_P)$ for $P[G_O, G_P|E_O, Y_O = 1]$ in (20) assuming correct specification of $P[H_P|E_O, Y_O = 1]$. Therefore, the authors chose the estimating-equation vector $\mathcal{U}_\theta$ to be $\mathcal{S}_\theta$ with jth element corresponding to $S_\theta\big((G_O, G_P) = j\big)$ and assuming a working model for $P[H_P|E_O, Y_O = 1]$ that uses parental haplotype frequencies (estimated via an EM algorithm from parental genotype data) under the assumptions of HWE and random mating. Based on this model, it is straightforward to show that $\widetilde{\mathcal{S}}_\theta = \mathcal{S}_\theta - \Psi(\Psi'\Psi)^{-1}\Psi' S_\theta$ is the portion of $\mathcal{S}_\theta$ orthogonal to the columns of Ψ.

Using the PCPH approach, one can estimate the haplotype and haplotype–environment interaction effects in θ by solving $\widetilde{\mathcal{S}}_{\theta} = 0$ using an iterative algorithm described in [2]. Further, one can also use the PCPH approach to construct robust score statistics for testing haplotype and haplotype–environment interaction effects on disease. Define $\widetilde{\mathcal{S}}_{\theta,i}$ as the estimating-equation contribution for the ith triad ($i = 1, \ldots, n$). One can then construct the robust score statistic as $T = n\overline{\mathcal{S}}'_{\theta_0} V_{\theta_0} \overline{\mathcal{S}}_{\theta_0}$, where

$$\overline{\mathcal{S}}_{\theta_0} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{S}}_{\theta_0,i}; \;\; V_{\theta_0} = \frac{1}{n}\sum_{i=1}^{n} \left(\widetilde{\mathcal{S}}_{\theta_0,i} - \overline{\mathcal{S}}_{\theta_0}\right)\left(\widetilde{\mathcal{S}}_{\theta_0,i} - \overline{\mathcal{S}}_{\theta_0}\right)'$$

and θ_0 denotes the vector of parameter estimates evaluated under the null hypothesis. The score statistic T should asymptotically follow a χ^2 distribution with degrees of freedom equivalent to the rank of V_{θ_0}.

As the PCPH approach derives from the RR model in (19) and further allows for estimation of haplotype effects, the procedure permits more flexible modeling and testing of haplotype effects relative to haplotype FBAT. In particular, PCPH permits the testing of composite null hypotheses, where one is interested in assessing the effects of a specific haplotype while allowing the effects of the other haplotypes to be unconstrained and freely estimated. This is as opposed to haplotype FBAT which, in testing the effects of a specific haplotype, assumes the remaining haplotypes have no effect on disease risk. Such an assumption may be unappealing, particularly if multiple susceptibility haplotypes exist within the region of interest [2].

3.3 Software Packages

The haplotype approach developed by [23] for analysis of qualitative and quantitative outcomes is implemented in the popular FBAT software package. FBAT is a self-contained executable that runs via a terminal interface on a variety of OS, including Windows, Macintosh OSX, Linux, and Solaris. The software has many appealing features for general haplotype analyses in families. FBAT estimates haplotype frequencies using an internal EM algorithm and deals with the issue of rare haplotypes in association testing by removing those haplotypes with an estimated frequency less than a user-defined threshold (default setting of 0.05). The software also permits testing the null hypothesis of linkage, but no association (for families with more than 1 offspring) using the robust variance V_{R} have been noted earlier. In addition, the software can perform permutations to establish significance, such that one can obtain exact p-values for both global and single-haplotype test statistics. For single-haplotype analyses, note that resulting tests assume that the remaining haplotypes have no effect on disease risk.

The PCPH approach of [2] for haplotype analysis of qualitative outcomes in case–parent triads is implemented in the PCPH package, which is a set of self-contained executables that runs on a terminal interface on the Windows OS. Using estimated haplotype frequencies (computed using one of program's executables), the software constructs both individual and global score tests of haplotype effect of disease, as well as similar tests for detecting haplotype–environment interactions. Further, the software provides estimates of haplotype and haplotype–environment interaction effects with corresponding variances. For tests of individual haplotypes, PCPH allows the effects of the other haplotypes to be unconstrained in analysis. Unlike FBAT, PCPH does not provide permutation-based inference for calculation of exact p-values.

4 Summary

Haplotype analyses provides a complementary (and potentially more powerful) procedure for association mapping relative to traditional approaches that consider each SNP in a separate analysis. In this chapter, we have presented statistical methods for haplotype mapping of complex traits under many traditional population-based and family-based study designs and further described many public software packages that implement these approaches. While the majority of the methods presented here were developed to investigate the main effects of genetic factors on trait outcome, we showed that many of these methods can be extended to examine interaction effects between gene and environment. Unless otherwise noted, analyses using these haplotype methods are generally computationally efficient and can run on a single processor on a standard desktop computer in a matter of minutes.

We described the use of haplotype-based statistical methods for association mapping primarily in the context of the analysis of a small genomic region that one might investigate as part of a candidate-gene study. However, the recent arrival of improved high-throughput genotyping technology has facilitated the use of genomewide-association scans (see Chapter 11) for identifying loci that influence a complex trait of interest. Huang et al. (2007) investigated the application of haplotype-based methods to genomewide data by employing a sliding-window design and adjusting for the multiple testing of windows using an efficient Monte-Carlo procedure. PLINK implements a similar sliding-window procedure for haplotype analysis in genomewide scans, although the package appears to use computationally-intensive permutations to adjust for multiple testing. Nevertheless, there remains some open issues regarding the most powerful manner by which to employ haplotypes for association mapping across the human genome. While the use of sliding-windows of haplotypes across the genome is simple and easy to implement, it ignores the underlying LD structure in the human genome and therefore may result in the analysis of windows where there is little correlation among SNPs (and hence little LD information contained within the haplotypes). To rectify this problem, studies may want to define windows based on underlying haplotype-block structure, which one can assess by applying block algorithms implemented in software like Haploview (Barrett et al. 2005) to appropriate reference samples from the

International HapMap Project (2005). Further investigation of this strategy may be warranted.

Overall Software Recommendations: While our advice regarding software for haplotype analysis depends on the study design, our recommendations also substantially depend on the computational background of the analyst conducting the study. We feel that naive users of haplotype software with little computational expertise would likely prefer a software package with features that provide an intuitive platform for haplotype analysis, such as point-and-click capabilities and a friendly graphic-user interface. On the other hand, we feel that analysts with substantial computational and statistical background would likely prefer software that is flexible, easily assimilated into other software and scripts (e.g. software written using R code), can be run in batch, and is portable across different operating systems. With that in mind, we stratify our recommendations for haplotype software based not only on the study design, but also on the computational expertise of the user. Our recommendations follow below:

I. Cross-Sectional Study: HAPSTAT (naive user)

PLINK or haplo.stats (advanced user)

II. Cohort Study: HAPSTAT (naive and advanced users)
III. Case-Control Study: CHAPLIN or HAPSTAT (naive user)

PLINK or haplo.stats (advanced user)

IV. Family-based Study: PCPH (naive user)

FBAT (advanced user)

Electronic-Database Information

CHAPLIN (http://www.genetics.emory.edu/labs/epstein/software)
FBAT (http://www.biostat.harvard.edu/∼fbat/default.html)
Hap-Clustering (http://www4.stat.ncsu.edu/∼jytzeng/Softwares/Hap-Clustering/R/)
haplo.stats (http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm)
HAPSTAT (http://www.bios.unc.edu/∼lin/hapstat/)
PCPH (http://www.duke.edu/∼asallen/Software.html)
PLINK (http://pngu.mgh.harvard.edu/∼purcell/plink)
QSHS (http://www4.stat.ncsu.edu/∼jytzeng/Softwares/QSHS/)

Acknowledgements We thank the FUSION study investigators for allowing us to present results from the analysis of FUSION data. We also thank Dr. Glen Satten for his comments on a previous version of this chapter. This work was supported by National Institutes of Health grants HG003618 and GM074909.

References

1. 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
2. Allen AS, Satten GA (2007) Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method. Genet Epidemiol 31:211–223
3. Allen AS, Satten GA, Tsiatis AA (2005) Locally efficient robust estimation of haplotype-disease association in family-based studies. Biometrika 92:559–571
4. Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F (2000) Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 64:255–265
5. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120
6. Carroll RJ, Wang S, Wang CY (1995) Prospective analysis of logistic case–control studies. J Am Stat Assoc 90:157–169
7. Chen H-S, Zhu X, Zhao H, Zhang S (2003) Qualitative semi-parametric test to detect genetic association in case–control design under structured population. Ann Human Genetics 67: 250–264
8. Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177
9. Clayton D, Jones H (1999) Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65:1161–1169
10. Cox DR (1972) Regression models and life-tables (with discussion). J R Stat Soc B 34:187–220
11. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. J R Stat Soc 39:1–38
12. Devlin B and Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322
13. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
14. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001) Experimentally derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28:361–364
15. Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25:115–121
16. Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35–43
17. Eitan Y, Kashi Y (2002) Direct micro-haplotyping by multiple double PCR amplifications of specific alleles (MD-PASA). Nucleic Acids Res 30:e62
18. Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329
19. Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927
20. Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork N (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE variation and Alzheimer's disease. Genome Res 11:143–151
21. Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, Mohlke KL, Silander K, Kohtamäki K, Chines P, Balow J, Birznieks G, Chang J, Eldridge W, Erdos MR, Karanjawala ZE, Knapp JI, Kudelko K, Martin C, Morales-Mena A, Musick A, Musick T, Pfahl C, Porter R, Rayman JB, Rha D, Segal L, Shapiro S, Sharaf R, Shurtleff B, So A, Tannenbaum J, Te C, Tover J, Unni A, Welch C, Whiten R, Witt A, Blaschak-Harvan J, Douglas JA, Duren WL, Epstein MP, Fingerlin TE, Kaleta HS, Lange EM, Li C, McEachin RC, Stringham HM, Trager E, White PP, Eriksson J, Toivanen L, Vidgren G, Nylund SJ, Tuomilehto-Wolf E, Ross EH, Demirchyan E, Hagopian WA, Buchanan TA, Tuomilehto J,

Bergman RN, Collins FS, Boehnke M (2000) The Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus Genetics (FUSION) Study. I. An autosomal genome scan for genes that predispose to Type 2 diabetes. Am J Hum Genet 67:1174–1185

22. Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype-phenotype associations. Eur J Hum Genet 9:301–306
23. Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol 26:61–69
24. Huang BE, Amos CI, Lin DY (2007) Detecting haplotype effects in genomewide association studies. Genet Epidemiol 31:803–812
25. Joosten PH, Toepoel M, Mariman EC, Van Zoelen EJ (2001) Promoter haplotype combinations of the platelet-derived growth factor alpha-receptor gene predispose to human neural tube defects. Nat Genet 27:215–217
26. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
27. Kraft P, Cox DG, Paynter RA, Hunter D, De Vivo I (2005) Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet Epidemiol 28:261–272
28. Kwee LC, Epstein MP, Manatunga AK, Duncan R, Allen AS, Satten GA (2007) Simple methods for assessing haplotype–environment interactions in case-only and case–control studies. Genet Epidemiol 31:75–90
29. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ (2003) Estimation and tests of haplotype–environment interaction when linkage phase is ambiguous. Hum Hered 55:56–65
30. Lewinger JP, Bull SB (2006) Validity, efficiency, and robustness of a family-based test of association. Genet Epidemiol 30:62–76
31. Lin DY (2004) Haplotype-based association analysis in cohort studies of unrelated individuals. Genet Epidemiol 26:255–264
32. Lin DY, Zeng D (2006) Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc 101:89–104
33. Lin DY, Zeng D, Millikan R (2005) Maximum-likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol 29:299–312
34. Liu N, Beerman I, Lifton R, Zhao H (2006) Haplotype analysis in the presence of informatively missing genotype data. Genet Epidemiol 30:290–300
35. Louis T (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc B 44:226–233
36. McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, London
37. McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York
38. Michalatos-Beloin S, Tishkoff S, Bentley K, Kidd K, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 24:4841–4843
39. Molitor J, Majoram P, Thomas DC (2003) Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73:1368–1384
40. Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233
41. Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169
42. Piegorsch WW, Weinberg CR, Taylor JA (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. Stat Med 13:153–162.
43. Prentice RL, Pyke R (1979) Logistic disease incidence models and case–control studies. Biometrika 66:403–412
44. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000). Association mapping in structured populations. Am J Hum Genet 67:170–181

45. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007a) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81: 559–575
46. Purcell S, Daly MJ, Cham PC (2007b) WHAP: haplotype-based association analysis. Bioinformatics, 23:255–256
47. Rabinowitz D (2002) Adjusting for population heterogeneity and misspecified haplotype frequencies when testing non-parametric null hypotheses in statistical genetics. J Am Stat Assoc 97:742–751
48. Rabinowitz D (2003) Adjusting for population heterogeneity: a framework for characterizing statistical information and developing efficient test statistics. Genet Epidemiol 24:284–290.
49. 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 504:227–233
50. Rosenberg PS, Che A, Chen BE (2006) Multiple hypothesis testing strategies for genetic case-control association studies. Stat Med 25:3134–3149
51. Satten GA, Epstein MP (2004) Comparison of prospective and retrospective methods for haplotype inference in case–control studies. Genet Epidemiol 27:192–201
52. Satten GA, Kupper LL (1993) Inferences about exposure-disease associations using probability-of-exposure information. J Am Stat Assoc 88:200–208
53. Schaid DJ (1995) Relative-risk regression models using cases and their parents. Genet Epidemiol 12:813–818
54. Schaid DJ (2004) Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348–364
55. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434
56. Scott AJ, Wild CJ (1997) Fitting regression models to case–control data by maximum likelihood. Biometrika 84:57–71
57. Self SG, Longton G, Kopecky KJ, Liang KY (1991) On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–61
58. Seltman H, Roeder K, Devlin B (2003) Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58
59. Silander K, Scott LJ, Valle TT, Mohlke KL, Stringham HM, Wiles KR, Duren WL, Doheny KF, Pugh EW, Chines P, Narisu N, White PP, Fingerlin TE, Jackson AU, Li C, Ghosh S, Magnuson VL, Colby K, Erdos MR, Hill JE, Hollstein P, Humphreys KM, Kasad RA, Lambert J, Lazaridis KN, Lin G, Morales-Mena A, Patzkowski K, Pfahl C, Porter R, Rha D, Segal L, Suh YD, Tovar J, Unni A, Welch C, Douglas JD, Epstein MP, Hauser ER, Hagopian W, Buchanan TA, Watanabe RM, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2004) A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. Diabetes 53:821–829
60. Spinka C, Carroll RJ, Chatterjee N (2005) Analysis of case–control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol 29:108–127
61. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
62. Stram DO (2005) Tag SNP selection for association studies. Genet Epidemiol 27:365–374
63. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike ML (2003a) Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 55:27–36
64. Tavtigian S, Simard J, Teng D, Abtin V, Baumgard M, Beck A, Camp J, et al. (2001) A candidate prostate cancer susceptibility gene at chromosome 17p. Nat Genet 27:172–180
65. Tzeng JY (2005) Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol 28:220–231

66. Tzeng J-Y, Devlin B, Wasserman L, Roeder K (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72:891–902
67. Tzeng JY, Wang CH, Kao JT, Hsiao CK (2006) Regression-based association analysis with clustered haplotypes through use of genotypes. Am J Hum Genet 78:231–242
68. Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M (1998) Mapping genes for NIDDM: design of the Finland-United States Investigation of NIDDM Genetics (FUSION) study. Diabetes Care 21:949–958
69. Van der Meulen MA, te Meerman GJ (1997) Association and haplotype sharing due to identity by descent with an application to genetic mapping. In: Edwards JH, Pawlowitzki IH, Thompson E (eds) Genetic mapping of disease genes. Academic Press, London, pp. 115–135
70. Watanabe RM, Ghosh S, Langefeld CD, Valle T, Hauser ER, Magnuson VL, Mohlke KL, Silander K, Ally DS, Chines P, Blaschak-Harvan J, Douglas JA, Duren WL, Epstein MP, Fingerlin TE, Kaleta HS, Lange EM, Li C, McEachin RC, Stringham HM, Trager E, White PP, Balow J, Birznieks G, Chang J, Eldridge W, Erdos MR, Karanjawala ZE, Knapp JI, Kudelko K, Martin C, Morales-Mena A, Musick A, Musick T, Pfahl C, Porter R, Rayman JB, Rha D, Segal L, Shapiro S, Sharaf R, Shurtleff B, So A, Tannenbaum J, Te C, Tovar J, Unni A, Welch C, Whiten R, Witt A, Kohtamaki K, Ehnholm C, Eriksson J, Toivanen L, Vidgren G, Nylund SJ, Tuomilehto-Wolf E, Ross EH, Demirchyan E, Hagopian WA, Buchanan TA, Tuomilehto J, Bergman RN, Collins FS, Boehnke M (2000) The Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus Genetics (FUSION) Study. II. An autosomal genome scan for diabetes-related quantitative-trait loci. Am J Hum Genet 67:1186–1200
71. Weinberg CR (2003) Studying parents and grandparents to assess genetic contributions to early-onset disease. Am J Hum Genet 72:438–447
72. Whittemore AS (2004) Estimating genetic association parameters from family data. Biometrika 91:219–225
73. Yang Q, Khoury MJ, Flanders WD (1997) Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 146:713–720
74. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91
75. Zeng D, Lin DY, Avery CL, North KE, Bray MS (2006) Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. Biostatistics 7:486–502
76. Zhang S, Pakstis A, Kidd K, Zhao H (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data, Am J Hum Genet 69:906-912
77. Zhang H, Zheng G, Li Z (2006) Statistical analysis for haplotype-based matched case-control studies. Biometrics 62:1124–1131
78. Zhang H, Zhang H, Li Z, Zheng G (in press)Statistical methods for haplotype-based matched case-control association studies. Genet Epidemiol
79. Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. 2000. Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 67:936–946
80. Zhao JH, Curtis D, Sham PC (2000) Model-free analysis and permutation tests for allelic associations. Hum Hered 50:133–139
81. Zhao LP, Li SS, Khalid N (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231–1250

Multiple Comparisons/Testing Issues

Qingrun Zhang and Jurg Ott

Abstract The statistical testing of multiple genetic markers in genetic linkage and association studies is discussed and shown to lead to a multiple-testing problem. Various solutions are discussed and demonstrated on published data. The false discovery rate (FDR) and several approaches of estimating it, are mentioned. Randomization (permutation) testing is highly recommended.

1 Introduction

At one of the early Genetic Analysis Workshops, a participant asked whether trying different inheritance models and phenotype definitions in linkage analysis would lead to an increase in the rate of false positive results. There seemed to be uncertainty about this point and the reason for it soon became clear: Shortly before the workshop, a statistical investigation was published showing that "when misspecifying the genetic parameter values, neither linkage nor heterogeneity can be falsely concluded" [1]. Whereas, this result is undoubtedly true it was widely misinterpreted to mean that no matter how many parameter values and disease models were tried, the rate of false positive results would not be increased. For example, lod scores for schizophrenia linkage of 6.5 corresponding to odds for linkage exceeding 3,000,000 were reported [2]. Let's consider in detail what is going on here.

A statistical test lets researchers come to a conclusion regarding two hypotheses, in this case, whether there is linkage (H_1 = alternative hypothesis) or not (H_0 = null hypothesis) between two loci, for example, a hypothesized disease locus and a marker locus. The result of such a test is a p-value, the empirical significance level, which indicates the probability that a result as extreme or more extreme than the one observed may occur by chance alone, that is, when in fact there is no linkage (H_0 is true). Small values of p (<0.01, say) convince researchers that the result is unlikely

J. Ott (✉)
Beijing Institute of Genomics, Chinese Academy of Sciences, No. 7 Bei Tu Cheng West Road, Beijing 100029, China,
e-mail: ottjurg@yahoo.com

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_9, © Springer-Verlag Berlin Heidelberg 2010

to be due to chance and they accept it as real, in this case, the actual existence of linkage.

Generally, each of a large number of markers is tested for genetic linkage or association to a disease phenotype. As genetic association studies are becoming more and more important, they will be the focus of our discussion here rather than linkage studies. In case–control association studies, 100,000s of SNP markers are subjected to a chi-square test based on a contingency table, with the two rows corresponding to case and control individuals and the columns referring to either the three genotypes or the two alleles at a given SNP. Each test result is considered "significant" (a true finding is inferred) if p falls below a threshold α. With a large number, m, of tests being carried out, rather than focusing on one test at a time, one may be interested in the more general question, *how many* of these tests could be due to chance alone? Or, similarly, what is the probability, α, that one or more of these tests show an extreme result by chance? The α probability is known as the experiment-wise (or global or overall) significance level, whereas p is the pointwise significance level. In many situations, the former is much larger than the latter, $\alpha >> p$, so that steps must be taken to keep α below a prescribed level such as 0.01. This situation is referred to as the multiple testing problem, and steps taken to adjust α to a proper level are multiple testing corrections.

Later, we outline various methods that have been proposed to overcome the multiple testing problem. Most of these methods have been developed in the statistical framework of significance testing, which is also how the multiple testing problem has been introduced and will be handled here.

2 Bonferroni Correction

Consider the situation that multiple association tests are independent. Each of m tests is carried out at the significance level, p. That is, with each test one runs the risk p of declaring the result significant in the absence of any true association. Still assuming no real association, we want to know the probability, P($\geq$ 1 test significant), that at least one of these tests is significant. This is equal to $1 -$ P(no test significant) or $1 -$ P(all tests nonsignificant). For any individual test, the probability of it not being significant is given by $1 - p$ and, since they are all independent, we find that P(all tests nonsignificant) $= (1 - p)^{\mathrm{m}}$. Therefore, $\alpha \equiv \mathrm{P}\,(\geq 1$ test significant$) = 1 - (1-p)^{\mathrm{m}}$, and this is approximately equal to *mp* if p is small. For example, if any one of 20 tests is declared significant when it results in $p \leq 0.01$, then the probability of one or more tests being (falsely!) declared significant is 0.182 or approximately $20 \times 0.01 = 0.20$. Thus, for the overall significance level to be small, for example, $\alpha = 0.05$, each individual test must be carried out at the more stringent significance level of $p = \alpha/m$, which is referred to as the [3] correction. It is known to be conservative, that is, the correction is too stringent so that power is lost. This is particularly true when tests are *de*pendent, which is easy to see for tests that are perfectly correlated – each gives the same result, so they may be represented by a single test and no correction is necessary. Some researchers seem to misunderstand this situation. In response to a review of a manuscript, an author claimed that no multiple testing correction was required because his tests were all independent. It is with independent tests that corrections for multiple testing are most needed!

An interesting concept is due to [4]. For m dependent tests, he proposed a procedure to compute an equivalent number, m_{eff}, of independent tests. Bonferroni corrections could then be carried out with m_{eff} instead of m, which is beneficial because $m_{\mathrm{eff}} < m$. Cheverud's procedure involves the computation of eigenvalues of the $m \times m$ correlation matrix of genotype codes (for example, $\mathrm{AA} = 0$, $\mathrm{AB} = 1$, $\mathrm{BB} = 2$) and may not be practical for large numbers m of markers. Permutation testing (see section on single test statistic) allows the derivation of m_{eff} but also computes significance levels directly thus circumventing the need for invoking the Bonferroni correction.

3 False Discovery Rate

Recall that the significance level is the probability of obtaining a positive (significant) result when in fact there is no true effect such as linkage or association (null hypothesis is true). This is also why the significance level is called the rate of false positive results, or the false positive rate. Conditioning on whether the null hypothesis is true or false is not very intuitive and is often misunderstood by nonstatisticians. A more plausible concept focuses on all positive (significant) test results and asks, among these, what is the proportion that is real (true positives) and what proportion is false positives? The latter proportion is called the false discovery rate (FDR), and methods have been developed to keep it to a low level such as 0.05 [5–7]. The QVALUE program [6] provides an easy way to estimate the FDR associated with a given test result and its p-value.

Various approaches have been implemented in statistical packages to identify "significant" test results, that is, those tests associated with an FDR no larger than a given small value. The simplest and best known method is probably the Benjamini-Hochberg (BH) method [8]. It works as follows: Order the p-values associated with each test by size from the smallest (most significant) to the largest, $p_{[1]} < p_{[2]} < \ldots p_{[\mathrm{i}]} \ldots < p_{[\mathrm{m}]}$. For each test, also compute $q_{\mathrm{i}} = i \times (0.05/m)$. Then, one starts with the largest p-value, $p_{[\mathrm{m}]}$, and compares it and successively smaller ones with the corresponding q values. Let k be the first test found in this manner for which $p_{[\mathrm{k}]} < q_{\mathrm{k}}$. Then all tests with p-values ranked $i \leq k$ are associated with an FDR of no more than 0.05 and are declared significant. This and other FDR approaches are adaptive in the sense that the decision criterion changes as one gradually eliminates one after the other hypothesis as not being significant. Thus, the largest p-value is compared with $m \times (0.05/m)$; if it is not significant then the next smallest p-value is compared with $(m-1) \times (0.05/m)$, and so on. Once a test

Table 1 The Benjamini-Hochberg (BH) procedure applied to the five smallest significance levels in a study on AMD (Klein et al. 2005) compared with Bonferroni-corrected p-values, p_{Bon}

Rank,i	$p_{[\text{i}]}$	$i \times (0.05/103,611)$	$\mathbf{p}_{\text{Bon}} = p_{[\text{i}]} \times 103,611$
1	4.02×10^{-8}	48.3×10^{-8}	0.0043
2	7.58×10^{-8}	96.5×10^{-8}	0.0080
3	1.36×10^{-6}	1.45×10^{-6}	0.1409
4	2.60×10^{-5}	0.193×10^{-5}	(2.69)
5	3.01×10^{-5}	0.241×10^{-5}	(3.12)
–	–	–	–
103,611		0.05	

has been eliminated as nonsignificant, there are fewer tests remaining that need to be examined, which is one way of seeing the rationale for the BH procedure [8].

The BH approach is known as a step-up procedure as it starts with the largest p-values at the bottom of the list and gradually steps up to the smallest one. Alternative (step-down) approaches go in the other direction. For example, starting with the smallest significance level, $p_{[1]}$, the BL procedure compares each $p_{[\text{i}]}$ with $h_{\text{i}} = \min\left[0.05, 0.05 \times m/\left(m + 1 - i\right)^2\right]$ until it finds $p_{[\text{k}]} > h_{\text{k}}$. Then, the smallest $k-1$ p-values are declared significant [8]. For large values of tests, h does not change much initially while the q criterion in the BH procedure does not change much for large p-values but does so for small p-values. BH may thus be more appropriate for large numbers of tests and is also generally better known as the BL approach.

As an example for the BH procedure, consider a well-known study of age-related macular degeneration (AMD) [9]. In their allele-specific association tests, these authors found SNPs with the five smallest significance levels shown in Table 1. According to the Bonferroni criterion, with $m = 103,611$, only the two smallest p-values are significant at the overall 0.05 level (Table 1). The BH procedure, however, also declares the third-smallest p-value significant (indeed, that SNP turned out to be important). Assuming that none of the p-values larger than those in the table have an $\text{FDR}<0.05$, one would then declare these three markers significant.

4 Randomization Testing

For small sample sizes and two groups of observations, Fisher [10] introduced the idea of a randomization test (often also called permutation test). For example, consider n_1 case and n_2 control individuals, and assume that a statistical test has been carried out to see whether there is a significant difference in genotype frequencies between cases and controls, where S_{o} is the observed test statistic and large values of S_{o} are considered indicative of a difference. Under the null hypothesis of no difference, a given genotype is equally likely to have occurred in a case or a control individual. Therefore, all possible permutations of "case" and "control" labels are equally likely. For such a randomization (permutation) sample of randomly assigned labels, one may compute the test statistic S and determine the proportion of S values obtained in all permutation samples that are at least as large as S_{o}. This proportion is the empirical significance level, p, associated with S_{o}. Many authors include S_{o} in the formation of p but this is not universally done.

Just as different conventional statistical tests yield different results when applied to the same data, may permutation tests furnish different significance levels depending on the specific test statistic used. Also, some test statistics may be more appropriate and more powerful than others to detect a given hypothesis. In the example provided in Chap. 6.1 of [11], for the same data two different test statistics resulted in p-values of 0.0263 and 0.0037, a sevenfold difference in significance levels!

With sample sizes of n_1 and n_2, the total number of permutations is given by $N_{\mathrm{p}} = (n_1 + n_2)!/(n_1!n_2!)$, which may be a very large number. For example, for $n_1 = n_2 = 15$, N_{p} already exceeds 115 million. Therefore, one generally works with a random sample of size N obtained from all possible permutations and carries out a Monte Carlo randomization test. Efficient algorithms for generating sequences of random permutations are available [12]. In human case–control studies, computer-based randomization tests have been proposed [13]. Empirical significance levels so obtained are estimates whose precision increases with N. Ideally, in addition to the estimated significance level one should also supply the associated confidence interval for the true but unknown significance level, which may be carried out with the BINOM program [1], which furnishes exact confidence intervals based on the binomial distribution. For example, assume that on the basis of $N = 10{,}000$ randomization samples, a significance level of $p = 0.04$ was estimated. Then the 2-sided 95% confidence interval is (0.036, 0.044). In practice, with a sufficient number of permutations, estimated significance levels are rather accurate so that usually no confidence intervals are reported.

The beauty of randomization samples is that they permit to construct the sampling distribution under the null hypothesis for an arbitrary test statistic, which often is impossible with classical statistical approaches. However, the test statistic must be computed for each randomization sample. If this is time-consuming then a randomization test can be very computationally intensive. In addition, because the total number N_{p} of permutations tends to be extremely high, the number N of permutation samples should be comparatively high in order for the sample space of permutations to be adequately represented in the randomization samples. Thus, values of $N \approx 10{,}000$ or higher should preferably be used.

A recent publication proposes permutation testing by importance sampling [14], that is, instead of sampling from the whole parameter space of permutations only those are sampled leading to an interesting result. Whereas, this approach reduces sampling effort dramatically it is based on some assumptions that are not required with general sampling. Furthermore, genome-wide significance levels in human case–control association studies don't tend to be extremely small (the focus of these authors' publication) so that moderate values of N should be sufficient to accurately estimate significance levels.

As mentioned earlier, significance levels obtained from randomization samples permit the determination of a number m_{eff} of effectively independent markers. For example, in our study on age-related macular degeneration (AMD), each of a total of $m = 103,611$ SNPs were tested for association with AMD [9]. In the comparison of allele frequencies between case and control individuals, the best SNP showed a Bonferroni-corrected significance level of $p_{\text{Bon}} = 0.0043$ (Table 1). On the other hand, permutation testing with our SUMSTAT program based on 30,000 randomization samples [2] resulted in $p_{\text{rand}} = 0.0035$. Now, the Bonferroni correction was obtained by multiplying the uncorrected significance level p by m, that is, $p_{\text{Bon}} = p \times m$, and we want to ask, what number m_{eff} of independent tests would have to be postulated so that a Bonferroni correction applied to p would lead to p_{rand}? That is, we require $p \times m_{\text{eff}} = p_{\text{rand}}$ or $m_{\text{eff}} = p_{\text{rand}}/p$. With $p = p_{\text{Bon}}/m$ from above, we can now write

$$m_{\text{eff}}/m = p_{\text{rand}}/p_{\text{Bon}},$$

which, in our case, furnishes $0.0035/0.0043 = 81\%$. Carrying out these calculations with 3,000,000 permutations each in a few other SNPs, we found values of m_{eff}/m between 73 and 94%. Evidently, these calculations are not very reliable but they give us an idea of how much correlation exists among test results for the 100 K SNP chip. These results are not really useful in practice because we have already obtained the proper significance levels from permutation samples, but knowing the effective number of SNPs may help in power calculations that assume independent tests, for example, with the CaTS program for power calculations in two-stage association studies [15]. Like in most such computations, independence of marker tests is assumed, where working with m instead of m_{eff} would require a much too stringent significance level resulting in an artificially low power. For 300 and 500 K SNP chips, we don't yet have many results but ratios of m_{eff}/m of around 50% or less may well be expected.

5 Single Experiment-Wise Test Statistic

So far, we have discussed several means of correcting results from multiple tests, where each test furnished a result, for example, a p-value. Now, we turn to an altogether different principle. Rather than correcting the p-values for large numbers of tests, an alternative approach is to define a single test statistic for all markers and derive its associated p-value, which by definition is an experiment-wise significance level as only one test is carried out [16, 17]. Because the null distribution of this statistic may be unknown, its associated significance level is preferably estimated from randomization samples. For example, for a number of markers, the largest of the transmission-disequilibrium test statistics, $\text{TDT}_{\max}$, over all the markers has been proposed as an experiment-wise test statistic [18]. Because of nonindependence among markers, the p-value associated with $\mathrm{TDT}_{\max}$ is best obtained from computer-based randomization samples.

For the genome-wide AMD study mentioned earlier [9], one may also consider the largest marker-specific test statistic as the single test statistic representative for the whole experiment. In 30,000 randomization samples, 105 of these samples showed a maximum chi-square at least as large as the one in the observed data (Table 1). Thus, the overall significance level turned out to be equal to $105/30,000 = 0.0035$.

A special genome-wide test statistic has been proposed as follows [17]. To evaluate the joint association of a number m of SNPs (irrespective of their genomic positions), one forms the sum, S_{m}, of the m largest test statistics and evaluates the empirical significance level, p_{m}, associated with S_{m} via permutation sampling. This is done for each of the values, $m = 1, 2, \ldots, m_{\max}$, where $m_{\max}$ is a suitable upper limit such as 20 (the same set of N_{p} permutation samples is used for each m). The smallest of the $m_{\max}$ values of p_{m} is then chosen as the experiment-wise test statistic, and *its* associated significance level is obtained from the randomization samples. Technical details of this procedure are given in [2]; see also [19]. Another somewhat related statistic, the scan statistic, is the largest sum of test statistics for m consecutive SNPs over the genome [20]; see also [3].

Now, let us return to the question of which test statistic is most appropriate for a given situation. As we will see, this is not always easy to answer. Consider a genetic case–control association study with a large number of SNP markers, each with two alleles, A and B, which for convenience may be coded as 0 and 1, respectively. For each marker, an allele test may be carried out by computing χ^2 for a 2×2 table, whose rows correspond to case and control individuals, with columns representing the two alleles at the SNP. One may then designate the largest observed χ^2 value as the single, genome-wide test statistic and evaluate its associated significance level by permutation testing. The χ^2 statistic is equivalent to the absolute difference, $|f_{\mathrm{case}} - f_{\mathrm{control}}|$, divided by the square root of its variance, where f is the frequency of B alleles in case or control individuals. Because the variance is small for small f values, differences are weighted more for rare alleles than for common alleles. For example, the difference between 0.07 and 0.05 is weighted twice as much as the difference between 0.36 and 0.34. However, we may want to consider these two differences the same. If so, we would choose as our SNP specific test statistic $|f_{\mathrm{case}} - f_{\mathrm{control}}|$, not weighted by any variance. Which of these two statistics is more powerful or more appropriate has not been thoroughly investigated but their statistical properties are likely to be different. Various test statistics have recently been proposed and shown to have different power in different situations [21–23].

6 Example Dataset: Parkinson Disease

Few case–control datasets are currently available to statistically minded researchers as test beds for their analysis methods. We chose the Parkinson disease (PD) dataset with 270 case and 271 control individuals [24], each genotyped for over 400,000

SNPs [4]. To minimize the amount of computing necessary we picked the chromosome 11 data, which contain 19,539 SNPs. An increasing number of analysis programs are available to data analysts, but we prefer to keep as close to the data as possible and write our own analysis programs.

All works were carried out on a desktop PC running Windows XP with 3.00 GB of RAM and 3.59 GHz clock speed. Programs were written in Free Pascal [5], and all chi-square statistics were calculated as likelihood ratio rather than Pearson chi-squares. A small number of SNPs had a large number of missing observations. We wanted to retain only those SNPs with no more than 25% missing observations. This eliminated three SNPs: rs4756052 (170 missing), rs6591003 (392 missing), and rs7110392 (169 missing), so we were left with 19,536 SNPs. In addition, it appears useful to eliminate those SNPs whose genotype frequencies are severely out of Hardy–Weinberg equilibrium (HWE) because this is indicative of genotyping errors. We used a Bonferroni-corrected significance threshold of $0.05/400,000 = 1.25 \times 10^{-7}$, corresponding to a Hardy–Weinberg disequilibrium (HWD) χ^2 of 27.9. This step eliminated an additional seven SNPs so that we were left with $m = 19,529$ SNPs. Figure 1 demonstrates that indeed exactly seven SNPs exhibit unusually high HWD values.

The following calculations refer to multiple testing on chromosome 11. With a whole-genome analysis, of course, one would need to correct for many more tests than done for this example. Testing for allelic associations, we found a largest chi-square (1 df) of 17.53 corresponding to an uncorrected empirical significance level of 0.000,028,28. Bonferroni correction leads to $p_{\text{Bon}} = 0.55228$ while randomization testing with 20,000 permutation samples resulted in $p_{\text{rand}} = 0.34615$. Thus, for the given marker density on chromosome 11, the effective proportion of independent SNPs is $p_{\text{rand}}/p_{\text{Bon}} = 63\%$. As expected, this is somewhat smaller than the values we found in our AMD study. In the genotype test, the largest chi-square (2 df) turned out to be 33.36 with a (chromosome-wide) Bonferroni-corrected significance level of 0.001,113. In 20,000 randomization samples, the significance level

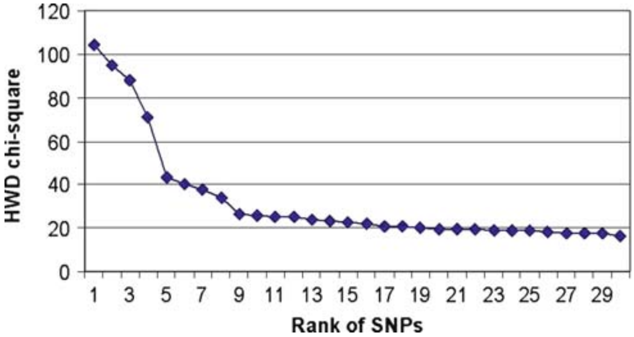


Fig. 1 Hardy–Weinberg disequilibrium chi-square values (y-axis) versus ranks (x-axis) of 30 SNPs with highest values. Outliers (eight largest chi-squares) show a deviation from the smooth curve of smaller chi-square values

was essentially the same as the one resulting from Bonferroni correction, for which we don't have a good explanation.

7 Discussion

Multiple testing corrections are not without controversy. Some authors argue that no corrections are needed but this does not seem quite right. However, valid questions remain. In particular, should corrections be applied to all experiments ever done in a given problem area? Clearly not. Should corrections be applied to the experiments described in a single publication? This is what the current consensus seems to be, but then one might argue that the best "strategy" would be to publish as few experiments in any single paper and distribute results over many papers. Although this argument sounds contrived, it does point out a dilemma that is difficult to overcome. I don't see a universally good answer except that the problem is self-regulating – editors will prevent people from publishing every minor advance in their research. Additional aspects of multiple testing have been briefly discussed by Balding [25]. As pointed out by that author, the problem is not only the number of tests carried out by one researcher but there are many other researchers who work on the same problem and thus contribute to the multiplicity of tests.

As is well known, effects of linkage may be detected over wide genomic regions, certainly over several centi-Morgans away from a disease locus, that is, several MB of sequence. On the other hand, loci tend to be in linkage disequilibrium (show genetic association) only when they are at most about 0.1 MB apart. Lander and Kruglyak [26] showed that for linkage analysis, a saturation point is eventually reached when markers become more and more dense. That is, the correction for multiple testing is not increased by much when a researcher increases the number of microsatellite markers from 600 to 800 in a genome-wide study, and a critical lod score limit of 3.3 ensures an experiment-wise significance level of 0.05 in standard linkage analysis for any number of markers. For genome-wide association testing, it has recently been proposed that in the limit of very dense maps, the multiple testing burden is equivalent to 1 mio. independent tests in Europeans and twice this number in Africans [27].

Many "significant" association studies have been published, but subsequently could not be replicated. Several possible reasons have been quoted for these failures [28]. While insufficient power is certainly one of them, a major reason must be that until recently little attention has been paid to rigorous corrections for multiple testing. It is understandable that researchers resist these efforts. After all, they want to get their papers published and editors are reluctant to publish work that is not significant. This has led to some play with words that might not be easy to recognize as such by nonstatisticians. Clearly, what we need is experiment-wise significance. Some authors have started using the term "point-wise significant" when one of many experiments or tests exhibits a significance level $p < 0.05$. While this is formally correct, editors could easily be misled as they see “significant” and don’t realize that “point-wise significant” is another way of saying nonsignificant.

We would like to close with a quote: “Significant results in abstracts are common but should generally be disbelieved” [29].

Web Resources

1. http://www.genemapping.cn/util.htm
2. http://www.genemapping.cn/sumstat.htm
3. http://www.genemapping.cn/scanstat.html
4. http://ccr.coriell.org/ninds/
5. http://www.FreePascal.org/
6. http://genomics.princeton.edu/storeylab/qvalue/

References

1. Clerget-Darpoux F, Babron MC, Bonaiti-Pellie C (1987) Power and robustness of the linkage homogeneity test in genetic analysis of common disorders. J Psychiatr Res 21:625–630
2. Sherrington R, Brynjolfsson J, Petursson H, Potter M, Dudleston K, Barraclough B, Wasmuth J, Dobbs M, Gurling H (1988) Localization of a susceptibility locus for schizophrenia on chromosome 5. Nature 336:164–167
3. Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3–62
4. Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52–58
5. Devlin B, Roeder K, Wasserman L (2003) Analysis of multilocus models of association. Genet Epidemiol 25:36–47
6. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440–9445
7. Efron B (2004) Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. Journal of the American Statistical Association 99:96–104
8. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125:279–284
9. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389. Epub 2005 Mar 2010
10. Fisher RA (1935) The design of experiments. Oliver & Boyd, Edinburgh
11. Manly BFJ (2007) Randomization, bootstrap, and Monte Carlo methods in biology. Chapman & Hall/CRC, Boca Raton, FL
12. Nijenhuis A, Wilf HS (1978) Combinatorial algorithms for computers and calculators. Academic, New York
13. Sham PC, Curtis D (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 59:97–105
14. Kimmel G, Shamir R (2006) A fast method for computing high-significance disease association in large population-based studies. Am J Hum Genet 79:481–492

15. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213. Epub 2006 Jan 2015
16. Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48:67–81
17. Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 11:2115–2119
18. McIntyre LM, Martin ER, Simonsen KL, Kaplan NL (2000) Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. Genet Epidemiol 19:18–29
19. Becker T, Cichon S, Jonson E, Knapp M (2005) Multiple testing in the context of haplotype analysis revisited: application to case-control data. Ann Hum Genet 69:747–756
20. Hoh J, Ott J (2000) Scan statistics to scan markers for susceptibility genes. Proc Natl Acad Sci U S A 97:9615–9617
21. Zheng G, Freidlin B, Gastwirth JL (2006) Comparison of robust tests for genetic association using case-control studies. IMS Lecture Notes-Monograph Series 49:253–265
22. Matthews AG, Haynes C, Liu C, Ott J (2008) Collapsing SNP genotypes in case-control genome-wide association studies increases the type I error rate and power. Statistical Applications in Genetics and Molecular Biology 7:Art. 23
23. Zhang Q, Wang S, Ott J (2008) Combining identity by descent and association in genetic case-control studies. BMC Genet 9:42
24. Fung HC, Scholz S, Matarin M, Simon-Sanchez J, Hernandez D, Britton A, Gibbs JR, Langefeld C, Stiegert ML, Schymick J, Okun MS, Mandel RJ, Fernandez HH, Foote KD, Rodriguez RL, Peckham E, De Vrieze FW, Gwinn-Hardy K, Hardy JA, Singleton A (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 5:911–916
25. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781–791
26. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247
27. Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32:381–385
28. Ott J (2004) Association of genetic loci: Replication or not, that is the question. Neurology 63:955–958
29. Gotzsche PC (2006) Believability of relative risks and odds ratios in abstracts: cross sectional study. Bmj 333:231–234

Estimating the Absolute Risk of Disease Associated with Identified Mutations

Mitchell H. Gail and Nilanjan Chatterjee

Abstract For a given mutation status, we define the absolute risk as the chance that disease develops in a defined age interval, given that the person is well at the beginning of the interval. Absolute risk is reduced by competing risks of mortality, that may cause the person to die before the disease of interest develops. We distinguish absolute risk from the pure cumulative risk of disease that is often estimated in the genetic epidemiologic literature, and we concentrate on estimating marginal risks for members selected at random from the population, rather than family specific risks. We review cohort, population-based case–control, case–control family, and kin-cohort designs for estimating absolute and pure cumulative risks associated with a measurable genetic mutation.

1 Introduction

Once a disease-causing mutation has been identified and can be measured, it is useful to characterize the risk associated with the mutation. Although relative risk is a useful quantity for etiologic studies and can be estimated from case–control designs, absolute risk is more useful for clinical applications. For example, the concept of absolute risk is helpful in answering such questions as: “Given that I am carrying a mutation in the BRCA1 gene and am 20 years old, what is the chance that I will develop breast cancer before age 50?” or “How many BRCA1 carriers need to be enrolled in this prevention trial to have good statistical power to detect a fifty percent reduction in breast cancer risk from prophylactic treatment with tamoxifen?” or “If there were a treatment that reduced the risk of breast cancer by fifty percent, both for carriers and noncarriers of BRCA mutations, how many breast cancers could be prevented in a population of Ashkenazi women that had a prevalence of BRCA

M. H. Gail (✉)
Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville MD 20852, USA
e-mail: gailm@mail.nih.gov

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_10, © Springer-Verlag Berlin Heidelberg 2010

mutation carriers of 2.5%?" These examples illustrate some of the many uses of absolute risk. Other applications, such as weighing the risks and benefits of taking tamoxifen to prevent breast cancer have been discussed elsewhere [15].

For chronic diseases, such as cancer, we are interested in the distribution of age at disease onset, and we define absolute risk to be the probability that a person who is disease-free at age a will be diagnosed with the disease in a subsequent interval $[a, a+\tau]$ of duration τ, given the person's genetic status and perhaps some other covariates. This probability is influenced by competing causes of death, which can reduce the chance that the individual will be diagnosed with the disease of interest. For simplicity, we ignore other risk factors apart from age and mutation status, and let $g = 1$ denote a mutation carrier and $g = 0$ the homozygous wild type. If $h_g(t)$ is the age-specific incidence rate of the disease of interest in those with carrier status g, and if $h_2(t)$ is the age-specific mortality rate of dying from causes of death except the disease of interest, then the absolute risk (sometimes called "crude" risk in the competing risks literature) is given by

$$P(a, a+\tau|g) = \int_a^{a+\tau} h_g(t) \exp\left[-\int_a^t \{h_g(u) + h_2(u)\}\,\mathrm{d}u\right] dt. \tag{1}$$

The need to take competing risks into account is especially important when projecting risks over long time intervals, such as risks from age 30 to 70. Yet it is common in genetic epidemiology to ignore competing risks and present estimates of the cumulative "pure" probabilities

$$1 - \exp\left\{-\int_a^{a+\tau} h_g(u)\,\mathrm{d}u\right\}. \tag{2}$$

Such "pure" cumulative probabilities represent the risk that would be observed hypothetically if all other causes of death could be eliminated without altering the cause-specific hazard $h_g(t)$. Pure cumulative probabilities can be estimated as 1 minus the ratio of the Kaplan–Meier estimate of survival from age 0 to $a+\tau$, divided by the Kaplan–Meier estimate of survival from age 0 to a. Hereafter, we call (2) the "pure cumulative risk ."

To estimate the absolute risk, however, more specialized software is needed [1, 3, 14]. Once good estimates of $h_g(t)$ have been obtained, the absolute risks can be computed from (1) by substituting appropriate national estimates of mortality rates for $h_2(t)$. Although it is preferable to use absolute risk in consulting, we shall discuss pure cumulative risk in the following, in keeping with much of the genetic epidemiologic literature. Frequently, the pure cumulative risk from age 0 to t for carriers is termed the disease "penetrance" to age t. These concepts can be generalized for disease models in which $g = 2, 1, 0$ denote homozygous mutant, heterozygous, and homozygous wild types, each with its own hazard of disease. In what follows, however, we shall emphasize dominant models and use the previous notation for carrier status.

Some investigators have questioned the usefulness of estimating penetrance because many factors, apart from competing risks, can modify the absolute risk caused by a mutation. These may be host factors or environmental factors. For example, a person who is homozygous for a recessive mutation that causes phenylketonuria can be protected from mental retardation by controlling the diet. Thus, the risk from this mutation depends on the environment. Estimates of breast cancer risk from mutations in the BRCA1 and BRCA2 genes are often much higher in subjects who are enrolled because they are members of families with many affected members than in subjects from the general population [12, 20, 31]. This fact suggests that other familial factors, either behavioral or genetic, can modify the risk from these mutations. Another issue that complicates the concept of penetrance is the very definition of mutation. A large gene like BRCA1 is subject to mutations at many different loci, and there are seldom sufficient data to characterize the risks from mutations at each locus. Thus, risk is often ascribed to any mutation in a gene that alters its function substantially, and the penetrance represents an average of the penetrances of the various possible mutations. Despite these potential limitations on the usefulness of the concept of penetrance, we shall discuss some of the strengths and weakness of designs that are commonly used to estimate it.

Some of the most efficient designs to estimate penetrance are based on sampling families. This gives rise to an important definitional issue that we previously alluded to. Suppose the hazard for a member of a randomly selected family is $h_g(t)b$, where b is a positive random variable, sometimes called a "frailty," that characterizes the residual familial effect apart from the gene under study. Conditional on b, one could calculate an individual's absolute risk from (1) or pure cumulative risk from (2) by replacing $h_g(t)$ by $h_g(t)b$. However, b will not be observed. If G is the distribution of b in the general population and if b is independent of g, then the pure probability that a randomly selected member of the population would survive to age t is

$$\int \exp\left\{-b\int_0^t h_g(u)\,\mathrm{d}u\right\}\mathrm{d}G(b). \tag{3}$$

The corresponding hazard is

$$h_g^{\dagger}(t) = \frac{h_g(t)\int \exp\left\{-b\int_0^t h_g(u)\,\mathrm{d}u\right\} b\,\mathrm{d}G(b)}{\int \exp\left\{-b\int_0^t h_g(u)\,\mathrm{d}u\right\}\mathrm{d}G(b)} \tag{4}$$

What we are really estimating when studying random samples of individuals from the population is the "marginal" hazard $h_g^{\dagger}(t)$ and corresponding survival function (3). Of course, if residual familial effects are small, so that the variance of b is small, $h_g(t)$ will nearly equal $h_g^{\dagger}(t)$. Unless stated otherwise, we will let $h_g(t)$ stand for the marginal hazard $h_g^{\dagger}(t)$ in what follows.

A cohort study of a random sample of individuals naturally yields an estimate of the marginal genotype-specific hazards in the population (Sect. 2). A population-based case–control study of unrelated individuals, coupled with information on the overall age-specific disease rates in the population, also gives rise to such estimates, as we discuss in Sect. 3. When, however, sampling is based on membership in a family, one must take both the ascertainment scheme and the familial correlations into account to obtain estimates of the marginal genotype-specific hazard rates, pure marginal cumulative risks, and marginal absolute risks in the general population (Sects. 4 and 5).

Our discussion assumes that the mutation has been identified and can be measured. Methods like those in segregation analysis have been used to estimate the penetrance of a putative mutation, even before the mutation has been identified. For example, Claus et al. [9] estimated the penetrance of a putative dominant breast cancer gene from data on the dates of onset of breast cancer in mothers and sisters of women with breast cancer (cases) and of controls in a population-based case–control study. The likelihoods they used were similar to those found in segregation analysis but allowed for estimation of survival distributions corresponding to carriers and noncarriers. These calculations are the basis of widely used tables used for estimating pure cumulative breast cancer risk based on a woman's age and the history of breast cancer in her relatives [8]. We focus instead on designs in which at least some members of the study population are genotyped. Here, we use the term "mutation" to denote any genetic variant that is associated with increased disease risk. Sometimes "mutation" is reserved for highly penetrant variants that tend to be rare, such as mutations in BRCA1. The concepts in this paper also apply to more common genetic variants, such as a particular single nucleotide polymorphism (SNP) that may be associated with increased disease risk, even if the SNP itself is only a marker and does not cause the functional genetic defect. It may still be useful to characterize the risk associated with such a genetic marker for purposes of risk projection. Typically, it is easier to estimate the absolute risk for a marker with a prevalence of 5% or greater than for a rare mutation, because it can be difficult to assemble a study population with a sufficient number of rare mutations.

2 Population-Based Cohort Studies

If a random or representative sample of individuals is available for prospective follow-up, then standard survival methods can be used to estimate absolute risk and pure cumulative risk from (1) and (2), with $h_g(t)$ interpreted as a marginal hazard. A potential complication arises if the hazard from competing causes of death also depend on g, denoted by $h_2(t; g)$. In principle, if the cohort is large enough, both $h_g(t)$ and $h_2(t; g)$ can be estimated using standard methods [28], and the absolute risk can be calculated without any special assumptions on the "independence" of the competing risks. In this context, it would usually be misleading to use national mortality rates for competing causes in calculating (1), because the national rates would represent an average of the rates for carriers and noncarriers, and, for rare mutations, would correspond mainly to noncarriers. Suppose, for example, that carriers of BRCA mutations also have higher mortality from causes of death other than breast cancer. Then, national mortality rates would be smaller than the true competing hazard rates for carriers, and application of (1) would overestimate the absolute risk of breast cancer in carriers.

Estimates of $h_g(t)$ and of the pure cumulative risk depend on the assumption that censoring by death from other causes is independent of the development of breast cancer, conditional on genotype g. Thus, provided each member of the cohort is genotyped and conditional independence holds, unbiased estimates of $h_g(t)$ and of the pure cumulative risk can still be obtained, even if $h_2(t; g)$ depends on g.

In addition to its simplicity in concept and analysis, the cohort design offers the possibility of obtaining baseline materials and information on covariates that affect risk before the disease process has influenced these measurements. Efficient sampling of the cohort through case–cohort or nested case–control designs captures most of the information on baseline covariates by analyzing only the cases that develop disease and a comparable number of controls.

The main disadvantages to the cohort design are the large sample sizes and long follow-up times that are typically needed to estimate genetic relative risks or absolute risks with required precision. Long follow-up times and large cohorts are required because $h_g(t)$ is usually quite small, especially at younger ages. In addition, if interest centers on a rare genetic factor, such as a mutation in a BRCA gene, huge numbers of women in the general population would need to be screened to obtain an adequate number of randomly sampled carriers. A further complication is that medical practice may change and affect the risk of disease during a long study. For example, breast cancer incidence rates in a cohort of women may be altered by preventive measures such as the use of tamoxifen, or by changes in screening practices.

One approach to overcome these difficulties is to combine information from several cohorts from general populations. Such cohort consortia could potentially provide information needed to estimate penetrance precisely, even for a rather uncommon genetic variant. A second approach is to recruit women from "high-risk clinics." Women recruited from such clinics often have a strong family history of disease, however. Although estimates of absolute risk or pure cumulative risk obtained from such women may be appropriate for a high risk population, they may overestimate the penetrance for women in the general population. A third approach to shortening such studies is to use a retrospective cohort design. In an ideal retrospective cohort study, one can define the cohort at an earlier time, for example, from a roster of all persons working in a given plant on a specified previous date. Further, one can obtain complete and accurate disease ascertainment on all cohort members, irrespective of exposure status, and one can obtain accurate exposure data on all cohort members or on a properly sampled nested case–control or case–cohort subsample. Attempts to apply this paradigm to genetic studies pose difficulties that can lead to serious bias. Although one may find women with breast cancer and their relatives from lists of current clinic patients, it may be difficult to reconstruct a list of all the women in a defined cohort at risk at an earlier time, let alone assure complete follow-up and exhaustive or appropriately sampled genotyping from this cohort.

3 Case–Control Designs

Population-based case–control studies compare the genotypes of randomly sampled cases with those of controls to estimate relative risks, and, by coupling this information with data on the age-specific composite risk of disease in the population, genotype-specific pure cumulative risks (and absolute risks) can be estimated. To illustrate, suppose we have a random sample of cases $(Y = 1)$ aged 50–54 and corresponding controls $(Y = 0)$. From these data we can estimate $P(g|Y)$, and, if we also know the probability of getting disease from age 50 to 54 from population data, then, from Bayes' Theorem, we can compute the genotype-specific disease probabilities

$$P(Y = 1|g) = \frac{P(Y = 1)P(g|Y = 1)}{\sum_{y=0}^{1} P(Y = y)P(g|Y = y)}. \tag{5}$$

The fact that population-based case–control data can yield estimates of exposure-specific disease risk has been known since the path-breaking paper of Cornfield [11]. Examples of the use of this idea for piecewise constant hazard models are found in Gail et al. [14]. Langholz [22] and Benichou and Gail [4] provide analytical methods for absolute risk for the survival setting in which cases and controls are selected from nested case–control designs, and Self and Prentice [29] provide such techniques for the case–cohort design.

For rare mutations, genetic epidemiologists often compute a relative risk from the case–control data and multiply this relative risk times the population age-specific hazard to obtain the hazard in for a mutation carrier, $h_{g=1}(t)$. This strategy works if the mutation is rare because the population hazard then corresponds to the baseline hazard $h_{g=0}(t)$. For more common variants, however, (5) or its survival analysis equivalents must be used.

The population-based case–control design has several potential advantages. Because it is retrospective, there is no need to wait for disease to develop, as in a cohort study. If one is interested in the "natural history" associated with a mutation, such retrospective data from an era when preventive strategies were seldom used, may be more useful than a prospective cohort study. If the mutation is common, the required sample size for a case–control design is much smaller than for a corresponding cohort study.

A potential disadvantage of the population-based case–control design is the possibility of bias in the information on some risk factors, because cases may recall antecedent events in a different manner from controls. Provided samples from controls and cases are handled comparably, such differential errors do not typically affect the germ line genotype itself, however, because the genotype is stable. A second potential problem is difficulty obtaining representative samples of cases and controls, because a substantial proportion of those invited to participate and give a DNA sample may refuse. If the tendency for a case to participate is increased if that case has a strong family history, but the chance that a control will participate is not influenced by family history, differential nonresponse bias may result.

Some of these difficulties may be overcome by using a hospital-based, rather than a population-based case–control design. In the hospital based design, cases diagnosed at a hospital are compared to patients from that hospital with other diseases that are not thought to be affected by the gene under study. An advantage of this design is that hospitalized patients are more likely to participate in the study and give blood or other material for a DNA sample. A disadvantage is that the control diseases may be associated with the gene under study, distorting the association. Relative risks obtained from such studies can be multiplied by population-based estimates of $h_{g=0}(t)$ to yield an estimate of $h_{g=1}(t)$.

Even though case–control designs are generally thought to require comparatively small sample sizes, this is not the case for studies of rare mutations. For example, to estimate the lifetime cumulative pure risk of a rare mutation such as a mutation in BRCA1 with a precision ± 0.05, Gail et al. [16] calculated that over 17,030 genotypes would be required, based on an optimal control to case ratio with 15,506 controls and 1,524 cases. Many more controls than cases were needed for an efficient design because the mutation was so much more common in cases than in controls for this hypothetical mutation with lifetime penetrance 0.92 and mutant allele frequency 0.0033.

Family-based case–control studies, such as discordant sib-pair designs in which one sib is a case and another nondiseased sib a control, can be used to estimate relative risks, $rr = h_{g=1}(t)/h_{g=0}(t)$, but not the hazards $h_g(t)$ themselves. In a typical discordant sib-pair analysis, conditional logistic regression will be used, and any random familial effect b, will cancel from the conditional likelihood. The resulting relative hazard is therefore a family-specific relative hazard rather than a ratio of marginal hazards. If the mutation is rare, one can multiply this family-specific relative hazard by the age-specific population rates to approximate the marginal hazard for a carrier, and thereby estimate pure cumulative marginal risks and absolute risks. This approximation may not work well if the marginal and family-specific genetic relative risks are very different. To show that the differences between marginal relative risks and family-specific relative risks can be substantial, we considered the case of exponential survival and family-specific genetic relative risk 3.0. Assuming 30% of families had $b = 0.5$, 50% had $b = 1$, and 20% had $b = 1.75$, we calculated from (4) that the marginal genetic relative hazard was as low as 2.22 for some parts of the age range, which is 26% lower than the family-specific relative hazard. Hence, family-based estimates of relative risk could lead to overestimates of pure marginal cumulative risk. In practice, the familial frailties may have less variation, and the differences between family-specific relative hazards and marginal relative hazards may be smaller.

4 Case–Control Family Study Design

The case–control family study design involves the recruitment of population-based families through an index sample of case–control subjects. Consider a study design where N_0 cases and N_1 controls have been randomly sampled from the healthy and diseased subjects, respectively, in an underlying population. Let Y_{i0} and G_{i0}, $i = 1, \ldots, N_0 + N_1$, denote the disease status and the mutation status, respectively, for the $N_0 + N_1$ case–control subjects (probands). Here, if the ith subject is a case, $Y_{i0} = 1$ and if the ith subject is a control $Y_{i0} = 0$. Assume r_i relatives are recruited for the ith proband. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ir_i})$ and $\boldsymbol{G}_i = (G_{i1}, \ldots, G_{ir_i})$ denote the vector of disease outcome and mutation status, respectively, for the r_i relatives of proband i. The likelihood for case–control family data is given by

$$L_{FBCC} = \prod_{i=1}^{N_0+N_1} \Pr(\boldsymbol{Y}_i^R, \boldsymbol{G}_i, G_{i0} | Y_{i0}),$$

where one conditions on the event Y_{i0} to reflect the fact that families are sampled conditional on the disease status of the probands. Whittemore [33] considered inference based on L_{FBCC} under "reproducible" multivariate models for the distribution of disease-risk in families. Reproducible risk models imply that the marginal distribution of the disease status for any subset of subjects in a family depends only on their own covariates, but not on those for the other members of the same family. In particular, the marginal distribution of disease outcome for an individual subject given his/her own covariate cannot depend on the covariate status of the relatives of the subject. Many standard multivariate models for disease risk, including "marginal"[24] and "random effect" models, are reproducible. Whittemore points out that, "reproducibility" although desirable and mathematically convenient, requires an assumption of a certain type of independence between the covariates and the sources of familial correlation of disease.

Assuming reproducibility, Whittemore factorized L_{FBCC} as

$$\begin{aligned} L_{FBCC} &= \prod_{i=1}^{N_0+N_1} L_{1i} \times L_{2i} \times L_{3i}, \\ &= \prod_{i=1}^{N_0+N_1} \Pr(G_{i0}|Y_{i0}) \times \Pr(\boldsymbol{Y}_i|\boldsymbol{G}_i, G_{i0}, Y_{i0}) \times \Pr(\boldsymbol{G}_i|G_i). \end{aligned} \tag{6}$$

Further, under marginally specified multivariate distributions that assume a logistic model for the disease outcome of an individual subject, Whittemore used calculations analogous to Prentice and Pyke [27] to show that the estimate of the regression parameters that maximizes L_{FBCC} can be obtained from a "prospective" likelihood of the form

$$\begin{aligned} L^*_{FBCC} &= \prod_{i=1}^{N_0+N_1} L_{1i} \times L_{2i}, \\ &= \prod_{i=1}^{N_0+N_1} \Pr(Y_{i0}|G_{i0}) \times \Pr(\boldsymbol{Y}_i|\boldsymbol{G}_i, G_{i0}, Y_{i0}), \end{aligned}$$

where the intercept parameter of the marginal logistic model involved in the computation of L_{1i} needs to be distinguished from that involved in L_{2i}. Zhao et al. [34] considered an estimating equation based approach for inference on marginal models from family-based case–control studies, avoiding specification of the full multivariate distribution for families.

Hsu et al. [21] considered inference for case–control family data incorporating age-at-onset information. In this setting, the phenotype information for a subject can be summarized as $Y = (T, \Delta)$, where Δ denotes the indicator of whether the subject had the disease ($\Delta = 1$) or not ($\Delta = 0$) and $T = \min(T^*, C)$ denotes the minimum of age-at-onset (T^*) of the disease or censoring (C). In the following, we will assume both T and C are continuous with absolutely continuous distribution functions. If no covariates were present, the hazard function for the age-at-onset of the disease for a relative, given the outcome information of the proband, could be written as

$$h(t|t_0, \delta_0) = h(t|t_0, \delta_0 = 0)\psi(t, t_0)^{\delta_0}, \tag{7}$$

where

$$\psi(t, t_0) = \frac{S_{tt_0}(t, t_0)S(t, t_0)}{S_t(t, t_0)S_{t_0}(t, t_0)},$$

where $S(t, t_0) = \text{pr}(T^* > t, T_0^* > t_0)$ denotes the joint "survivor" function for the relative and the proband, $S_t(t, t_0) = \partial S(t, t_0)/\partial t$, $S_{t_0}(t, t_0) = \partial S(t, t_0)/\partial t_0$ and $S_{tt_0}(t, t_0) = \partial S(t, t_0)/\partial (tt_0)$. Clayton [10] introduced the "cross-ratio" function $\psi(t, t_0)$ as a measure of dependence between correlated failure times. Motivated by (7), Hsu et al. [21] proposed modeling disease incidence data of the relatives given their covariate information, such as mutation status (g), and the phenotype status of the index proband using a stratified Cox proportional hazard model of the form

$$h(t|g, t_0, \delta_0) = h_{0t_0} \exp\left[\beta g + \delta_0 \log\left\{\psi(t, t_0; \theta)\right\}\right], \tag{8}$$

where t_0, the proband's age, is considered to be the stratifying variable and the $\psi(t, t_0; \theta)$ is a parametric model for the cross-ratio function $\psi(t, t_0)$. The simplest model for the cross-ratio function [10] is to assume $\psi(t, t_0) = \theta$, i.e., it is constant for all values of (t, t_0). Oakes [26] described various other parametric forms for the cross-ratio function which could be generated by alternative copula distributions for bivariate failure times. Hsu et al. described an elegant and computationally simple approach for estimating the parameters β and θ of the model (8) based on a "pseudo-partial likelihood" of the data that essentially involves comparing the risk-score $\exp\left[\beta g + \delta_0 \log\left\{\psi(t, t_0; \theta)\right\}\right]$ between pairs of relatives from independent families within "matched" sets defined by the proband's age. They showed that the resulting estimator is consistent and developed a "sandwitch" approach for estimating the asymptotic variance of the estimator that accounts for the correlation in the partial-likelihood scores among the different relatives from the same family.

The approach of Hsu et al., however, has some limitations. First, interpretation of the parameter β in model (8) is defined conditional on the proband's disease outcome. It is not clear how β would define the genetic-risk of a randomly selected person from the population. Second, it is also not clear how to incorporate information on the covariates of the proband in model (8). Third, the pseudo-partial-likelihood approach may be inefficient because comparisons between cases and controls are restricted to be within highly stratified matched sets.

Shih and Chatterjee [30] considered an alternative approach. They proposed inference under a marginal Cox proportional hazard model of the form

$$h_1(t) = h_0(t)\exp(\beta), \tag{9}$$

where $h_g(t)$ is defined to be the hazard function of the disease for a randomly selected person from the population with mutation status g. Let

$$F_g(t) = 1 - \exp\left\{-\int_0^t h_g(s)\,\mathrm{d}s\right\}$$

be the corresponding pure cumulative-risk function and $S_g(t) = 1 - F_g(t)$ be the survivor function. Shih and Chatterjee proposed specifying the multivariate age-at-onset distribution for families using copula models [19, 26] with the marginal distributions for individual members being specified by the Cox model (9). Let T_0^*,T_1^*, …, T_r^* denote the random variables associated with the ages at onset of the disease for a family with $r+1$ subjects where the index "0" corresponds to the proband. The copula model corresponds to a specification of the multivariate survivor function

$$S_{g_0,g_1,\ldots,g_r}(t_0,t_1,\ldots,t_r) = \Pr(T_0^* > t_0, T_1^* > t_1, \ldots, T_r^* > t_r | g_0, g_1, \ldots, g_r)$$

of the form

$$S_{g_0,g_1,\ldots,g_r}(t_0,t_1,\ldots,t_r) = C_\theta\left\{S_{g_0}(t_0), S_{g_1}(t_1), \ldots, S_{g_r}(t_r)\right\}, \tag{10}$$

where $C_\theta(u_1,\ldots,u_m), \theta \in \Theta$ is any parametric class of multivariate distribution functions, defined on the product space of $[0,1]^{r+1}$, that has uniform marginals. The parameter θ can be interpreted as a measure of "residual familial aggregation" that characterizes familial correlation of the disease that cannot be explained by the gene under study. Clayton's model [10] of constant cross-ratio function corresponds to the copula function

$$C_\theta(u_0,u_1,\ldots,u_r) = \left[\sum_{m=0}^{r} {u_m}^{1-\theta} - r + 1\right]^{1/(1-\theta)}. \tag{11}$$

The value of $\theta = 1$ corresponds to independence and $\theta > 1$ corresponds to positive dependence. Although in a restricted range, values $\theta < 1$ can be allowed to model negative correlation, in this article we only allow for positive dependence ($\theta \geq 1$) for modeling familial aggregation. Oakes [26] described several classes of copula models induced by frailties.

Copula models are “reproducible.” If $i_{r_1}, \ldots, i_{r_k}$ denote the indices for a subset of the subjects for $r + 1$ relatives in a family, then

$$\Pr(T^*_{i_1} > t_{i_1}, \ldots, T^*_{i_k} > t_{i_k} | g_0, g_1, \ldots, g_r) = \Pr(T^*_{i_1} > t_{i_1} \ldots, T^*_{i_k} > t_{i_k} | g_{i_1}, \ldots, g_{i_k})$$

and

$$S_{g_{i_1}, \ldots, g_{i_k}}(t_{i_1}, \ldots, t_{i_k}) = C_\theta \left\{ S_{g_{i_1}}(t_{i_1}), \ldots, S_{g_{i_k}}(t_{i_k}) \right\},$$

where $C_\theta(u_{i_1}, \ldots, u_{i_r})$ is the same class of parametric function as $C_\theta(u_0, u_1, \ldots, u_r)$ except that it is defined on the space $[0,1]^k$ instead of $[0,1]^{r+1}$, with the interpretation of θ remaining unchanged. For the special case $k = 1$,

$$\Pr(T^*_{i_1} > t_{i_1} | g_0, g_1, \ldots, g_r) = C_\theta \left\{ S_{g_{i_1}}(t_{i_1}) \right\} = S_{g_{i_1}}(t_{i_1}),$$

where the last equality follows because of the uniform marginals of copula distributions.

Shih and Chatterjee showed that under copula models one can write the hazard of the age-at-onset of disease for a relative given his/her mutation status (g) and the index proband’s outcome $y_0 = (t_0, \delta_0)$ and mutation status (g_0), as

$$\lambda(t | g, t_0, \delta_0, g_0) = \lambda_0(t) \exp\left[\beta g + \delta_0 \log\left\{\psi(t, t_0; \theta)\right\}\right] \phi_\theta \left\{ S_g(t), S_{g_0}(t_0) \right\}, \tag{12}$$

where

$$\phi_\theta(u, v) = u \frac{\partial C_\theta(u, v)}{\partial u} / C_\theta(u, v).$$

There are several important differences between the specification of the conditional hazard function given in (12) from that in (8). First, (12) incorporates the genotype information of the proband as well as that for the relatives. Second, in (12), the age information from the proband is incorporated through the semiparametrically specified cross-ratio functions $\psi(t, t_0)$ and its derivatives. In contrast, in (8), age of the proband is treated as a stratifying variable. Third, the genetic-risk parameter in (12), unlike that in (8) has the desired marginal interpretation.

Shih and Chatterjee proposed estimation of β and θ using a likelihood decomposition similar to (6). They, however, considered several modifications. First they proposed replacing L_1, the “retrospective” likelihood for the case–control probands, by L_1^C, a conditional likelihood similar to that proposed by Li et al. [23], so that the resulting method can handle age-matched case–control sampling. Second, they suggested replacing L_2, the full likelihood of the relatives, by a composite likelihood L_2^* that treats the different relative-proband pairs from a family as independent units. This composite-likelihood approach is asymptotically unbiased for estimation of the parameters of the marginal model, although it may lose some information due to ignoring the full correlation structure in the families. It is also computationally simpler and less sensitive to model miss-specification compared to the full-likelihood approach. Both likelihoods L_2 and L_2^* involve the unknown baseline nonparametric baseline hazard function $\lambda_0(t)$. Shih and Chatterjee exploited the proportional hazard structure of the model (12) to propose an iterative estimation scheme. Each iteration yields estimates of β and θ by maximizing $L = L_1^c \times L_2$ or $L = L_1^c \times L_2^*$ with fixed $\lambda_0(t)$, and then estimating $\lambda_0(t)$ by a closed form Nelson–Aalen type estimator for fixed β and θ. Simulation studies suggested that this approach can produce estimates of the regression (β) and correlation parameters (θ) that are much more precise than those obtained by the partial-pseudo-likelihood method considered by Hsu et al. [21]. Moreover, it produces an estimate of the baseline hazard function $\lambda_0(t)$ that is essential for estimation of pure cumulative-risk and absolute risk.

In summary, the case–control family design has some advantages over a standard case–control design for estimation of genetic risk. Additional information from the relatives can substantially improve the efficiency of estimates of relative-risk parameters. Moreover, disease incidence data of the relatives also allows estimation of cumulative- and absolute risks associated with a genetic variant via internal estimation of the baseline hazard function. The application of this design, however, has been limited, as recruitment of relatives in a case–control study may difficult for a number of practical reasons, and it may not be feasible to obtain covariate information and genotypes on relatives. In Sect. 5, we will review an alternative more practical design that does not require recruitment of the relatives themselves, but relies instead on obtaining the relatives' disease history information from an interview of the proband.

5 Kin–Cohort Design

The kin–cohort design is based on probands who agree to be genotyped and to provide a history of the ages at onset of the disease of interest in their first-degree relatives. These relatives constitute a retrospective cohort, whose genotype distributions can be inferred from the probands' genotypes and Mendelian principles. Hence, pure cumulative risk and absolute risk can be estimated. Struewing et al. [31] used this design to estimate the risks of breast cancer and other cancers from BRCA1 and BRCA2 mutations in a population of Ashkenzi men and women in the Washington DC area. Letting $F_1(t)$ and $F_0(t)$ denote the pure cumulative risk in carriers and noncarriers respectively, they noted for this rare mutation that the cumulative risk in first-degree relatives of carrier probands was $0.5\ F_1(t) + 0.5\, F_0(t)$, because about half the first-degree relatives of carrier probands are expected to be carriers. Likewise, the cumulative risk in first-degree relatives of noncarrier probands is very nearly $F_0(t)$. Hence, $F_1(t)$ and $F_0(t)$ can be estimated from empirical estimates of the pure cumulative risks in first-degree relatives of carrier and noncarrier probands. Using this technique, Struewing et al. estimated the pure cumulative risk to age 70 as 56%, which is considerably less than the estimate of 84% obtained from a consortium of high risk families [12], but closer to the population-based estimate of 40% found in New South Wales, Australia [20].

Before describing likelihood-based approaches to estimation, we discuss some of the strengths and weaknesses of the kin–cohort design. A kin–cohort study can be completed quickly, because probands are sampled cross-sectionally and the disease histories of relatives are collected retrospectively from the proband. Morever, it is possible to estimate the risks from several disease outcomes from a single kin–cohort study simply by asking the probands to provide each relative's history for several disease outcomes. The kin–cohort design usually requires slightly smaller samples than corresponding cohort or case–control studies to estimate absolute risk [16].

Kin–cohort studies are subject to some potential biases. If subjects tend to volunteer to be probands more readily if they have affected relatives, estimates of penetrance will be upwardly biased [16, 17, 31, 32]. If probands mistakenly report disease that is not present in relatives, penetrance can be seriously overestimated, whereas if probands fail to report disease, penetrance will be underestimated [16–18]. For rare mutations, very large sample sizes may be needed to assure that Wald-type confidence intervals have proper coverage [17, 18].

We now consider how the ascertainment of probands affects the likelihood analysis and how biases can result, unless residual familial correlation is taken into account. For subject i, let $Y_i = (T_i, \delta_i)$ be the vector whose first component is the age at end of follow-up, T_i, and whose second component is an indicator, δ_i, of whether or not the disease was diagnosed at T_i. Let Y_0 denote the proband's disease history (or phenotype) and let $\mathbf{Y} = (Y_1, Y_2, \ldots Y_r)$ be the vector of disease histories for the relatives of the proband. First we assume that probands are sampled at random from the population; later we consider "case-enriched ascertainment" in which case probands ($\delta = 1$) are sampled at a higher rate than control probands ($\delta = 0$). With randomly sampled probands, the likelihood is the product over probands of

$$P(g_0)P(Y_0|g_0)P(\mathbf{Y}|g_0, Y_0). \tag{13}$$

The quantity $P(g_0)$ can be estimated directly from the genotypes of probands, with or without the assumption of Hardy–Weinberg equilibrium, and $P(Y_0|g_0)$ is obtained from standard survival methods. Under the strong assumption that outcomes within a family are conditionally independent given the corresponding genotypes g_0 and $\mathbf{g} = (g_1, g_2, \ldots g_r)'$,

$$P(\mathbf{Y}|Y_0, g_0) = \Sigma_{\mathbf{g}} \prod_{i=1}^{r} P(Y_i|g_i)P(\mathbf{g}|g_0). \tag{14}$$

The conditional distribution $P(\mathbf{g}|g_0)$ of the vector of genotypes of the relatives, $\mathbf{g}$, given the proband's genotype can be computed using standard Mendelian methods. If there are residual familial effects that influence phenotype in addition to the genotypes under study, however, then (14) is incorrect. The correlations among components of $\mathbf{Y}$ and between components of $\mathbf{Y}$ and Y_0 need to be taken into account. For example, if there is a random familial effect b, such that conditional on genotypes and b the familial phenotypes are independent, then

$$P(\mathbf{Y}|Y_0, g_0) = \Sigma_{\mathbf{g}} \left\{ \int \prod_{i=1}^{r} P(Y_i|g_i, b)dG(b) \right\} P(\mathbf{g}|g_0). \tag{15}$$

Ignoring residual familial correlation leads to overestimates of $F_1(t)$, underestimates of $F_0(t)$, and overestimates of $P(g = 1)$ [5, 16–18].

Chatterjee and Wacholder [5] developed a simple and elegant approach to circumvent this problem when probands are sampled at random. In this case, each pair (Y_i, Y_0) can be regarded as sampled at random from the population, and, under model (15), $P(Y_i|g_0) = \sum_{g_i} P(Y_i|g_i)P(g_i|g_0)$, as can be seen by integrating (15) over b. Using this idea, Chatterjee and Wacholder pretended that the r doublets were independent to produce a composite likelihood

$$P(g_0)P(Y_0|g_0) \prod_{i=1}^{r} \sum_{g_i} P(Y_i|g_i)P(g_i|g_0). \tag{16}$$

Even though the doublets are not independent, consistent estimates of the penetrance can be obtained from (16), and the estimated variances can be corrected for correlations among the doublets by using "sandwich" estimates [5].

The approach of Chatterjee and Wacholder depends on two key assumptions. First, the probands are sampled at random. Second, as in the random effects model (15), $P(Y_i|g_i, g_0) = P(Y_i|g_i)$. This is a special case of the "reproducibility assumption" that is discussed by Whittemore [33] and by Gail and Chatterjee [13]. It can be violated, for example, if the allele under study is in linkage disequilibrium with another nearby disease-producing allele at a separate locus. Then, knowing g_0 provides additional information about Y_i to that provided by g_i. Another example is in Gail and Chatterjee [13].

For rare mutations it is more efficient to over-sample case probands. For such case-enriched ascertainment, one is sampling conditional on the proband's phenotype, and the appropriate conditional likelihood that takes ascertainment into account is

$$\begin{aligned} P(g_0|Y_0)P(\mathbf{Y}|g_0, Y_0) &= P(g_0|Y_0) \sum_{\mathbf{g}} P(\mathbf{Y}|\mathbf{g}, g_0, Y_0)P(\mathbf{g}|g_0, Y_0), \\ &= P(g_0|Y_0) \sum_{\mathbf{g}} P(\mathbf{Y}|\mathbf{g}, Y_0)P(\mathbf{g}|g_0), \qquad (17) \end{aligned}$$

where the last equality follows from the assumption that $\mathbf{g}$ is conditionally independent of Y_0 given g_0, as is reasonable for Mendelian transmission, and from the reproducibility assumption [13, 33] that Y is conditionally independent of g_0 given $\mathbf{g}$ and Y_0.

The term $P(g_0|Y_0)$ in (17) can be computed from Bayes' Theorem in terms of the marginal hazards $h_g(t)$ without the need to take into account residual familial correlations among family members' survival information, Y_0 and the vector $\mathbf{Y}$, given genotypes. For computing $P(\mathbf{Y}|\mathbf{g}, Y_0)$ in (17), one needs to make some model assumption about the nature of residual familial aggregation. If one assumes no residual familial aggregation, then one can write

$$P(\mathbf{Y}|\mathbf{g}, Y_0) = P(\mathbf{Y}|\mathbf{g}) = \prod_{i=1}^{r} P(Y_i|g_i),$$

which can be computed in terms of marginal hazards [16,25]. Ignoring residual correlations, however, can lead to overestimates of penetrance of the disease-producing mutation [5, 16–18]. Begg [2] points out the possibility of such bias in the extreme case where all probands are diseased.

Chatterjee et al. [7] considered use of copula models to account for residual familial aggregation in "kin–cohort" analysis of relatives' data from a case-enriched sample of probands. They observed that the likelihood (17) has similar structure as that for case–control family design (see 6) except that (17) needs to account for the missing genotypes for the relatives. Thus, they developed an expectation-maximization type algorithm to extend the maximum-likelihood and maximum-composite-likelihood techniques described in Shih and Chatterjee [30] to estimate relative-risk parameters, cumulative risks and residual familial aggregation. They considered a variation of the method that allows analysis of studies with case-only probands.

Chatterjee et al. conducted extensive simulation studies to arrive at several important conclusions. They observed that accounting for residual familial aggregation, even with a mis-specified model, eliminates or reduces the bias in estimates of cumulative risk parameters from kin–cohort data that may be incurred with case-enriched sampling of the probands. However, the studies showed that if only case probands were used, the analysis was sensitive to mis-specification of the model for residual correlation. They further observed that the disease incidence data for the relatives adds substantial information for estimation of the genetic relative-risk parameters, even though the relatives are not genotyped.

Another source of bias that may affect the kin–cohort design arises when the gene under study influences competing risks of mortality or survival following cancer onset [13, 16]. If the hazard from competing causes of death is increased in carriers, but not in noncarriers, the estimate of $F_1(t)$ will be too small [13]. This dependence of competing risks on genotype will reduce the number of case-probands who are carriers in the study, because many of them would have died of other causes between the time they developed the disease of interest and the time the cross-sectional survey for probands was conducted. A more severe downward bias in estimates of $F_1(t)$ results if the hazard of death following onset of the disease of interest is greater in carriers than in noncarriers [13]. Biases can also result if cases who are carriers tend to volunteer for the study more readily than cases who are noncarriers, as might happen if a potential proband knows he or she is in a family with carriers. It can be difficult to model the dependence on carrier status of competing hazards of death, hazard of death from the cause of interest following disease incidence, and tendency to volunteer. Nonetheless, Chatterjee et al. [6] modeled the dependence of competing risks on carrier status to account for the joint effects of BRCA1/2 mutations on ovarian and breast cancer risk.

6 Discussion

We have reviewed how cohort, population-based case–control, case–control family and kin–cohort designs can be used to estimate pure cumulative risks for carriers of a mutation. These pure risks can be adapted by taking competing causes of mortality into account to estimate absolute (or "crude") risks. These ideas extend readily to computation of genotype-specific risks for nondominant genetic diseases.

We have emphasized that for counseling and many other applications, we are interested in the genotype-specific risk for a randomly selected member with that genotype from the target population. Studies in which relative risks are estimated by comparing affected with unaffected family members yield family-specific genetic relative risks that may exceed the marginal relative risks of interest, in the presence of substantial random familial effects. Nonetheless, multiplying estimates of relative risk times population-based estimates of baseline age-specific disease hazard, as in the population-based case–control design, can provide a robust approach to estimation of genotype-specific hazards, pure cumulative risk, and absolute risk. If the disease-conferring genotypes are rare, age-specific disease hazards from the general population can usually be taken as baseline age-specific hazard rates. When the disease genotypes under study are not rare, the population hazards need to be muliplied by $1 - AR(t)$, where $1 - AR(t)$ is an estimate of age-specific population attributable risk to obtain the needed baseline hazard.

References

1. Anderson P, Borgen O, Gill R, Keiding N (1991) Statistical Models based on counting processes. Springer-Verlag, New York
2. Begg C (2002) On the use of familial aggregation in population-based case probands for calculating penetrance. J Natl Cancer Inst 94:1221–1226
3. Benichou J, Gail M (1990) Estimates of absolute cause-specific risk in cohort studies. Biometrics 46:813–826
4. Benichou J, Gail M (1995) Methods of inference for estimates of absolute risk derived from population-based case-control studies. Biometrics 51:182–194
5. Chatterjee N, Wacholder S (2001) A marginal likelihood approach for estimating penetrance from kin-cohort designs. Biometrics 57:245–252
6. Chatterjee N, Hartge P, Wacholder S (2003) Adjustment for competing risk in kin-cohort estimation. Genetic Epidemiol 25:303–313
7. Chatterjee N, Kalaylioglu Z, Shih J, Gail M (2006) Case-control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. Biometrics 62:36–48
8. Claus E, Risch N, Thompson W (1991) Genetic analysis of breast cancer in the cancer and steroid hormone study. Am J Human Genetics 48:232–242
9. Claus E, Risch N, Thompson W (1994) Autosomal dominant inheritance of early-onset breast cancer. Implications and risk prediction. Cancer 73:643–651
10. Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65:141–151
11. Cornfield J (1951) A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. J Natl Cancer Inst 11:1269–1275
12. Ford D, Easton D, Stratton M, Narod S, Goldar D, Devilee P, Bishop D, Weber B, Lenoir G, Chang-Claude J, et al (1998) Genetic heterogeneity and penetrance analysis of the brca1

and brca2 genes in breast cancer families. the breast cancer linkage consortium. Am J Human Genetics 62:676–689

13. Gail M, Chatterjee N (2004) Some biases that may affect kin-cohort studies for estimating the risks from identified disease genes. Springer, New York
14. Gail M, Brinton L, Byar D, Corle D, Green S, Schairer C, Mulvihill J (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879–1886
15. Gail M, Constantino J, Bryant J, Croyle R, Freedman L, Helzlsouer K, Vogel V (1999) Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 91:1829–1846
16. Gail M, Pee D, Benichou J, Carroll R (1999) Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped-proband designs. Genetic Epidemiol 16:15–39
17. Gail M, Pee D, Carroll R (1999) Kin-cohort designs for gene characterization. J Natl Cancer Inst Monogr 26:55–60
18. Gail M, Pee D, Carroll R (2001) Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. J Stat Plan Infer 96:167–177
19. Genest C, Mackay R (1986) The joy of copulas: bivariate distributions with given marginals. Am Stat 40:280–283
20. Hooper J, Southey M, Dite G, Jolley D, Giles G, McGredie M, Venter DED (1990) Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in brca1 and brca2. Australian breast cancer family study. Am J Human Genetics 8:813–826
21. Hsu L, Prentice R, Zhao L, Fan J (1999) On dependence estimation using correlated failure time data from case-control family studies. Biometrika 86:743–753
22. Langholz B, Goldstein L (1996) Estimation of absolute risk from nested case-control data. Biometrics 53:767–774
23. Li H (1998) Analysis of age of onset data from case-control family studies. Biometrics 54:1030–1039
24. Liang K, Zeger S, Qaqish B (1992) Multivariate regression-analyses for categorical-data. J Royal Stat Soc Ser B 54:3–40
25. Moore D, Chatterjee N, Pee D, Gail M (2001) Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. Genetic Epidemioly 20: 210–227
26. Oakes D (1989) Bivariate survival models induced by frailties. J Am Stat Assoc 84:487–493
27. Prentice R, Pyke R (1979) Logistic disease incidence models and case-control studies. Biometrika 66:403–411
28. Prentice R, Kalbfleisch J, Jr AP, Flournoy N, Farewell V, Breslow N (1978) The analysis of failure times in the presence of competing risks. Biometrics 34:541–554
29. Self S, Prentice R (1988) Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Statistics 16:64–81
30. Shih J, Chatterjee N (2002) Analysis of survival data from case-control family studies. Biometrics 54:1115–1128
31. Struewing J, Hartge P, Wacholder S, Baker S, Berlin M, McAdams M, Timmerman M, Brody L, Tucker M (1997) The risk of cancer associated with specific mutations of brca1 and brca2 among ashkenazi jews. New Engl J Med 336:1401–1408
32. Wacholder S, Hartge P, Struewing J, Pee D, McAdams M, Brody L, Tucker M (1998) The kin-cohort study for estimating penetrance. Am J Epidemiol 148:623–630
33. Whittemore A (1995) Logistic regression of family data from case-control studies. Biometrika 82:57–67
34. Zhao L, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice R (1998) Combined association and aggregation analysis of data from case-control family studies. Biometrika 85:299–315

Processing Large-Scale, High-Dimension Genetic and Gene Expression Data

Cliona Molony, Solveig K. Sieberts, and Eric E. Schadt

Abstract The now routine generation of large-scale, high-throughput data in multiple dimensions (genotype, gene expression, and so on) provides a significant challenge to researchers who desire to integrate data across these dimensions in hopes of painting a more comprehensive picture of complex system behavior. This type of integration promises to elucidate networks that drive disease traits associated with common human diseases like obesity, diabetes, and atherosclerosis. However, to effectively carry out this type of research not only requires the generation of large-scale genotype and molecular profiling data but also requires the development and application of methods and software in addition to a computing infrastructure capable of processing the large-scale data sets. Mastery of the methods and tools and having access to an appropriate computing environment capable of processing large-scale data will be critical to maintaining a competitive advantage, given future successes in biomedical research will likely demand a more comprehensive view of the complex array of interactions in biological systems and how such interactions are influenced by genetic background, infection, environmental states, life-style choices, and social structures more generally. In this chapter, we detail the methodological and computing issues associated with carrying out large-scale genome-wide association studies on tens of thousands of phenotypes, where the aim is to identify those phenotypes that are intermediate to DNA variations and disease phenotypes. This type of analysis can provide insights into the molecular networks that are perturbed by DNA and environmental variations, and as a result, induce changes in disease associated traits, providing a path to interpret genome-wide association study data as well as uncover networks that drive disease processes.

1 Introduction

The availability of low-cost, high-throughput technologies for genotyping hundreds of thousands of DNA markers has led to unprecedented success in human genetics

C. Molony (✉)
Rosetta Inpharmatics, LLC, a wholly owned subsidiary of Merck & Co., Inc., Seattle, WA 98109, USA
e-mail: cliona_molony@merck.com

S. Lin and H. Zhao (eds.), *Handbook on Analyzing Human Genetic Data*,
DOI 10.1007/978-3-540-69264-5_11, © Springer-Verlag Berlin Heidelberg 2010

over the past 2 years, with highly replicable associations identified for a number of common human diseases, including age-related macular degeneration [1–3], diabetes [4, 5], and obesity [6]. While these and the coming discoveries from a slew of genome-wide association studies currently under way provide a peek into pathways that underlie disease, they are usually devoid of context, so that elucidating the functional role such genes play in disease can linger for years, or even decades, as has been the case for genes like ApoE, an Alzheimer's susceptibility gene identified nearly 15 years ago [7]. Even in cases where an association to disease has been well localized to a given locus representing just a single gene, in the absence of experimental support the gene cannot be definitively claimed as the disease susceptibility gene. This problem is exacerbated in experimental murine cross populations derived from inbred mouse strains, where in addition to the problem of inferring the function of positionally cloned genes and of determining the mechanistic underpinnings of disease from the genetic data alone, the extent of LD operating in such populations makes positional cloning a difficult and time consuming process.

An alternative to the forward genetics approach to dissecting complex traits like disease is the construction of molecular networks that drive disease, where such networks are constructed from molecular phenotype data scored in populations that manifest disease. The information that defines how variations in DNA lead to variations in complex traits of interest flows through molecular networks that actually define the complex traits. Therefore, characterizing the molecular networks that underlie complex traits like disease can provide a more comprehensive view of disease, and this in turn can lead to the direct identification of key genes underlying disease processes, as well as providing a rich biological context within which to infer the functional roles played by these key genes. Recent studies characterizing gene expression networks have not only demonstrated that gene expression traits are significantly heritable, they have also demonstrated how genetic loci associated with gene expression traits can be combined with clinical trait data to infer causal associations between genes and disease traits [8–14]. Because complex biological processes that lead to disease are often system and context dependent, leveraging DNA variations as a systematic source of perturbations on molecular networks and clinical traits facilitates studying complex biological processes at the systems level, in addition to studying gene function at the level of individual pathways [15, 16].

However, carrying out studies to uncover sub-networks that drive disease traits associated with obesity, diabetes, and atherosclerosis in human populations involve a number the generation of large-scale genotype and molecular profiling data, novel methods and software to integrate these types of data with clinical data to elucidate networks driving disease, and then a computing infrastructure capable of processing the massive amounts of data being generated today. Mastery of the methods and tools and having access to a computing environment capable of processing large-scale data will be critical to maintain a competitive advantage, given future successes in biomedical research will likely demand a more comprehensive view of the complex array of interactions in biological systems and how such interactions are influenced by genetic background, infection, environmental states, life-style choices, and social structures more generally [17, 18]. This holistic view requires embracing complexity in its entirety, so that complex biological systems are beginning to be seen as dynamic, fluid systems able to reconfigure themselves as conditions demand [19–21]. An alternative to these artificial, more simplistic perturbations are naturally occurring genetic variations that segregate in populations such as those obtained from experimental crosses or that occur naturally. Not only are these types of perturbations naturally occurring, possibly making them more relevant for identifying key nodes in networks that drive disease, but they also occur in a multi-factorial context that enables the study of additive effects and epistatic interactions. Because most common human diseases are thought to manifest themselves as a result of many weak contributors rather than a few dominant factors, it is important to study complex traits in this more realistic setting.

In this chapter, we detail the methodological and computing issues associated with carrying out large-scale genome-wide association studies on tens of thousands of phenotypes, where the aim is to identify those phenotypes that are intermediate to DNA variations and disease phenotypes.

2 Data Management, Access and Workflow

While the primary purpose of a genome-wide association study (GWAS) is biomedical discovery, along with all of the long-term effort that goes into collecting phenotype data and pedigrees or subject information, and in addition to the effort required for genotyping and final statistical analyses, data management issues are one aspect of a genetic study that are often overlooked. When large-scale data is stored in spreadsheets or files, keeping it organized, documented, and ensuring that data integrity is maintained, can require significant resources. Efficient and safe sharing of data within a research group is often difficult, and preparing data for genetic analysis programs requires significant manual intervention that is tedious and, as a result, highly error prone.

Human genetics researchers will also be required to plan for and manage an efficient flow of data and information in a scale not typically encountered in most biomedical research settings. It is likely that individuals with expertise in informatics, large-scale data management, and possibly software and/or web development skills will be needed for a GWA study or genetics of gene expression (GOGE) program. Point-to-point transfer of genotype and/or gene expression data can be managed efficiently by a Laboratory Information Management System (LIMS) or by establishing shared access points. However, the actual storage and ability to access large-scale, raw data efficiently can be a complicated task. A relational database management system (RDBMS) is a system that manages data using a relational data model via a collection of relations (generally called tables). For example, an RDBMS is an efficient way to retain and access the meta-data of individual studies along with the annotation for individual samples and markers, as well as the individual data points for genotype and gene expression data. Almost all RDBMS employ

Structured Query Language (SQL) as their query language (primary user interface to manipulate the data). While SQL-compliant RDBM systems are designed to be inherently efficient, the design and execution of an RDBMS in many biomedical research settings is usually performed by a temporary member of a research group who has some technical experience but does not have the required level of expertise to design an optimal system. In some cases, researchers may even find that they require many more people to manage the data than to analyze it.

For genotype and phenotype data management, a number of supported SQL-based public [22] and commercial options do exist (http://www.progenygenetics.com), and over the long term these types of solutions may be more sustainable and allow for better use of the biomedical researcher's time. In practice, even when data management solutions are designed and executed by technical experts, researchers with multiple large-scale high-dimensional data sets may eventually encounter difficulties in managing and manipulating the files on which they wish to perform analysis. Researchers should also consider using alternative data format structures for both storage and analytical solutions that require a fraction of the disk space, RAM, and processing time for manipulation. For example, a number of customized binary formats have been employed in commonly used analysis packages such as PLINK [23] and HelixTree (GHD) [24]. Additionally, plug-in technologies can be incorporated into genotype provider software to actually create the compatible binary formats from the very start [24]. Standardized binary structure libraries exist and use of a common format will permit the sharing of data across a wide variety of computational platforms using applications written in different programming languages. For example, Hierarchical Data Format, http://hdfgroup.org/index.html (HDF) is a library and multi-object file format for the transfer of graphical and numerical data between computers. HDF is freely available and commonly used by organizations in the public and private sectors when the data challenges being faced push the limits of what can be addressed by traditional database systems, XML documents, or in-house data formats. Analysis tools such as R and Matlab® can easily handle this format. HDF also supports several different data models, including multi-dimensional arrays or tables, holding a mixture of related objects that can be accessed as a group or as individual objects. Combining the compact and efficient binary formats with a streamlined RDBMS to manage multiple data sets is an attractive path forward that will permit efficient exploration within and between complex data sets. A number of groups are pursuing this approach as a long-term strategy and some of these integrated tools, like Syllego (http://www.rosettabio.com/products/syllego) are already available.

Carrying out quality control (QC) on large-scale data may occur at many points in the execution of a GWA study. Here, we highlight a couple of these points with technical considerations. Genotype calling algorithms need to be robust and efficient, and many are undergoing rapid development and improvement as high-density genotyping becomes more accessible [25–30]. While it is not plausible to interrogate each SNP on a manual basis to ensure quality, identify single anomalies, or systematic biases, researchers should regularly refer to current literature on this subject and be prepared for QC investigations that may extend beyond the technology provider's software and lead to novel genotype calling algorithms. CHIAMO (www.stats.ox.ac.uk/∼ marchini/software/gwas/chiamo.html) is one such example developed during the WTCCC study [31]. Large-scale GOGE population studies permit researchers to identify inconsistencies based on an even broader panel of information. Some examples include using large panels of marker data to infer population substructure [24] or even IBS mapping between individuals to identify more closely related individuals available in PLINK [23]. Small panels of gene expression and sex-linked markers can be used to confirm gender identity, ethnicity, or highlight sample swapping. It is also important to plan for seamless comparison of data generated across platforms. This may require instituting mechanisms that normalize data to some common reference set of ideals before loading into a central repository. For example, markers genotyped on different platforms may use different reference strands to call the same SNP marker, and in some cases the base-pairing and reference alleles for the SNP may result in spurious inferences when results from different data sets differ. While this example is well known, technical safeguards should be put in place to permit one to look across large data/results repositories and be certain that data is represented in a uniform fashion in all cases.

3 Analysis Issues with High-Dimensional Data

3.1 Power

Estimating power is a tricky proposition even for a single trait. Add to that the high dimensionality realized when analyzing tens of thousands of traits, and the task of estimating power becomes downright daunting. For studies involving the genetics of gene expression there are two relevant dimensions to consider: (1) the genomics dimension, where power is affected by the number and characteristics of the markers chosen and (2) the trait dimension, where, conditional on the marker set, the power is affected by the number of traits considered as well as the number of samples. For discussion on marker set properties, see Chapter "Population-Based Association Studies".

Perhaps, the biggest challenge to estimating the power of a high-dimensional study is identifying the appropriate significance threshold given the multiplicity of markers and traits, accounting for the correlation structure among both the markers and traits. When available, prior information can be used to help guess at an appropriate significance threshold for a given type I error rate. Use of HapMap data gives an empirical distribution from which genotype data can be simulated. Simulating from the real marker data maintains the correlation structure observed in human populations. It is important to note that since these correlation structures differ from population to population, the HapMap population most closely resembling those in your study should be used in simulation. For example, in a study of primarily Caucasian Americans, the Utah CEPH HapMap population would be most appropriate. In cases where the population cannot be matched exactly, it is sufficient that the distribution of allele frequencies and patterns of LD be appropriate.

Prior data is also useful in accounting for correlation among traits. The ideal situation is when a dataset is available on the same platform, assay (gene set), tissue, and ascertainment (e.g. disease), which might be the case if a pilot profiling effort had been done. In the absence of a pilot dataset, sometimes an appropriate publicly available dataset can be identified (http://www.ncbi.nlm.nih.gov/geo/ or http://www.ebi.ac.uk/arrayexpress/). Given a well-matched dataset, simulations may be performed directly from the empirical distribution of the data, keeping the correlation structure fixed. In other words, the sampling should not be done independently across traits as this would increase the effective number of traits simulated. It is precisely the correlation among traits that reduces the effective number of independent traits and reduces the stringency by which we have to adjust for multiplicity.

To identify an appropriate significance level cutoff, data can be simulated under the null distribution. At this point, some consideration of the amount of errors tolerated should be taken into consideration. For a large scale study, restricting to an expected number of false positives of 0.05 over the entire study will be quite stringent. Likewise, a pointwise expected number of false positives of 0.05 is too loose. The typical approach to controlling error in large data sets is through the *false discovery rate* (FDR). In other words, the FDR is the expected proportion of false discoveries. This not only depends on the number of false discoveries but also on the number and effect size of (and ultimately the power to detect) the true signals in the dataset, on which speculation is difficult. For the purposes of calculating power, setting a criteria based on the *family-wise error rate* (FWER) is a viable approach to identifying a reasonable significance cutoff. The FWER represents the probability of realizing more than a given number of false positives, m. While it is common to consider the FWER when m is zero, given the dimensionality in the context large-scale gene expression data, choosing a larger m is prudent.

Another strategy to identify power (and analyze data) in the context of gene expression data might involve distinguishing between *cis* and *trans* regulation. In genetics of gene expression studies it is of interest to distinguish between DNA variation within a gene region that associates with the expression of that gene (cis effects) and DNA variation significantly outside a gene region that associates with the expression of that gene (trans effects) [32]. By restricting attention to the *cis* regulatory effects, the number of marker-trait pairs examined is significantly reduced, which in turn reduces the stringency required for these discoveries. This will reduce the required sample sizes to detect cis effects, but at the cost of reducing your ability to discover *trans* effects.

In the absence of any appropriate prior data, a good strategy might be to power the study for a GWA of a single disease phenotype. If a study is not minimally designed to detect genes contributing to the primary clinical phenotype of interest, its usefulness is questionable. Since in many cases, the effect sizes for gene expression traits are larger than those for complex disease traits, due to the fewer number of contributing factors, these effects may be captured even after adjusting for multiplicity. In all cases, a comprehensive power calculation package such as QUANTO may be useful for computing power for various study designs. For more flexibility, basic power functions are available in the R package Powerpkg.

3.2 Data Trends and Unaccounted for Heterogeneity

Many factors can affect gene expression levels and patterns including genetics, environment, demography, and technical factors. Some technical factors like dye bias and primer distance are well known and typically de-trended from the raw data by the processing software [33]. However, other factors may exist that are not well modeled or characterized, and many other nontechnical factors may be unknown or unmeasured. When these unknown factors are confounded or associated at random with the primary scientific hypothesis (genetic association), spurious associations may be seen, resulting in a loss of power to detect real associations. One approach to removing such trends from expression data is called *surrogate variable analysis* (SVA) [34]. In this approach, linear transformations of sets gene expression traits are used as surrogates for unobserved predictors. Typically, this approach is applied conditional on a primary predictor of interest, however, in the case of genetic studies, there are potentially hundreds of thousands to millions of predictors of interest. Therefore, this approach would have to be applied separately for each marker-gene pair. In this situation, SVA could be applied in the absence of any one predictor of interest, although it would then be important to perform the association analysis both with and without the SVA adjustment, since the adjustment could potentially wipe out real trends like hotspots or gene-set enrichments.

3.3 Outliers and Transformations

When employing parametric models, outliers can violate model assumptions, alter power and type I error, and, especially in the case of markers with relatively rare minor allele frequencies, lead to spurious associations. Therefore, it is important to include a strategy to deal with outliers by either removing them, “normalizing” them or employing models that minimize their influence. When tens of thousands of traits are involved, removing outliers by hand is not feasible. In the case of gene expression data, high-quality algorithms for the processing and QC of raw results will minimize trends and outliers caused by known technical factors. However, occasionally outliers remain. Standard statistical approaches to detecting outliers may be applied to flag potentially affected traits. Failing that, ad hoc criteria like identifying traits with gaps of greater than some arbitrary number (say 3) of standard deviations between ordered values can be used to identify potentially problematic traits. Other strategies for dealing with outliers or heavy tails include normalizing traits using the inverse-normal distribution, or employing nonparametric models such as

Kruskal–Wallace or robust regression methods that minimize the influence of heavy tails.

4 Implementing a Standard First-Pass Analysis Pipeline

The utility of high-throughput technologies like gene expression microarrays is the ability to simultaneously measure thousands of traits, which in turn allows us to look at patterns created by the traits, instead of just the individual traits themselves. More advanced approaches such as multivariate analyses and expression networks can incorporate multiple traits simultaneously. These approaches can be powerful, but they are computationally costly. Some of these will be treated later in this chapter. Another approach to identifying patterns in data is to, instead of focusing on patterns in the raw data, focus on patterns that arise in the results of the individual trait analyses. While this does disregard some information, the approach does have the advantage of simplicity and can yield valuable and meaningful information as a first pass.

Previous chapters have provided thorough details regarding approaches and computational resources for association analyses. The type of analysis and software chosen should, of course, be appropriate to the data design. Given a high-dimensional data set, computational tractability should be taken into consideration. For regression-based analyses, statistical packages such as R, SAS, and Matlab are convenient, user-friendly platforms; however, this convenience comes at a significant computational cost over more specialized code written in a more fundamental programming language like C or FORTRAN. The overhead realized from the use of packages like R, SAS, and Matlab can be a factor of more than 15 times greater than what can be achieved using C or FORTRAN. Access to powerful computing resources like large clusters of processors can reduce the criticalness of having highly optimized code, but to assess significance via permutation methods, computation will need to be efficient enough to be repeated. With the availability of software packages such as PLINK which have been specifically designed for whole-genome analysis, including functionality to deal with all aspects of analyses from basic QC to advanced issues like stratification and haplotype analysis, there are good-quality options for stand-alone analysis packages.

4.1 The Model – Common vs. Individual

When analyzing thousands of traits, it is not possible to choose an individual model for each, at least not by hand. Some expression traits might have sex, age, or environmental-factor specific patterns [35,36], while others may be relatively robust across conditions. While a failure to model these factors can lead to spurious associations if there is correlation between the unmodeled covariate and any of the marker genotypes, it is more likely to reduce power to detect true associations [37]. Including nonsignificant covariates, can add additional noise which in turn can reduce power to detect true signals. This effect, however, is typically small compared to the power lost by not including appropriate covariates. In some instances, sex-specific QTLs have been identified [35, 36]. To appropriately model QTL that are mediated by covariates, like sex, require including an interaction term in the genetic model. This increases the degrees of freedom required for the test of association, and in this case, falsely including the interaction term can reduce power [37].

Another approach to modeling covariates in a large number of traits is to employ an automatic model selection procedure using AIC, BIC, or p-value criteria. Two algorithms are available in R to perform stepwise regression: *step* and *stepAIC* in the MASS package. One caveat to remember when using model selection is that without the use of cross-validation, the resulting nominal p-values can be inaccurate due to the potential to overfit data. In this case, permutation is critical to assess the adjusted significance.

4.2 Estimating Heritability

Gene expression traits have been shown to be heritable [11, 12, 35], although the degree to which expression is heritable may differ from tissue to tissue [35]. Given a study design that incorporates phenotypic measurements on related individuals like case-parent-trios, sib-pairs, and pedigree-based studies, heritabilities can be estimated. For a more simple study, designs such as case-parent-trios and sib-pair designs, in which pairs (or trios) of related individuals are all related in the same way (e.g., all siblings or all parent-offspring), regression using any statistical package can be used. For example, the slope of the offspring's trait value regressed on the mean of the two parents' trait value is an accepted estimate of heritability (in the narrow sense). Alternately, when only one parent is available, twice the slope of the parent–offspring regression should be used. For full sibs, the estimate is also twice the slope of the regression of one sib on the other, or alternatively, 2Cov(sib pairs)/Var(trait). However, it should be noted that due to additional correlation arising from shared environmental factors, this is typically an upwardly biased estimate of heritability.

For more general pedigrees, where estimating heritability requires more involved modeling than simple linear regression, genetics packages are available. Linkage analysis software employing variance component methods (Almasy and Blangero, 1998) such as SOLAR can be used to estimate the polygenic component in the absence of a major gene. In other words, the same methodology used for linkage analysis can be used in the absence of marker data to estimate heritability. Many of the same principles apply to the estimation of heritability as apply to other genetic models. Covariates should be modeled when appropriate and available. Adjustments should be made for multiplicity just as it is in association analysis. Since heritability analysis is done in the absence of marker data (i.e., only once per trait), it is more computationally tractable and permutation (of trait values) can be used to assess significance and adjust for multiplicity.

4.3 Ethnicity and Substructure

Even after adjustment for demographic and environmental variables, some genes exhibit differential expression with respect to ethnicity. Failure to account for ethnicity can cause spurious associations for these traits at markers whose allele frequencies differ between populations. As with any association study, strategies to minimize the effects of substructure should be employed either through design (case-parent-trio or use of an isolated population) or appropriate analysis. Tools such as STRUCTURE [24] can identify population structure existing in a given sample, and various methods for structured association and genomic control can be used to adjust for it. These methods are discussed in detail in the chapter "Markov Chain Monte Carlo Linkage Analysis Methods". Principle component analysis based methods are a technically easier alternative to MCMC methods with reduced computation requirements and little cost in accuracy (Price et al. 2006).

4.4 Multiplicity

Genomewide association analysis of gene expression data requires testing tens of thousands of genes at hundreds of thousands to millions of markers. Failure to adjust for multiplicity can result in billions of false positives. On the other hand, simple Bonferroni p-value correction is too stringent to retain any power to detect true associations. For high-dimensional data, false discovery rate (FDR) is a less conservative approach to controlling false positives [38]. It allows some expected number of false positives, as long as that number is low compared to the number of true positives. For further discussion of general issues surrounding FDR analysis see the chapter "Multiple Comparisons and Multiple Testing Issues", though we will discuss strategies specific to the analysis of expression data.

The choice of FDR method depends on whether storage or computation is more limiting. The asymptotic approach implemented in the *qvalue* package [39] of R requires that the p-values from all tests be saved, which may not be possible without significant storage space and significant modification to the existing package to accommodate large data sets. For example, in a recent study we tested greater than one million SNPs for association to roughly 40,000 gene expression traits, resulting in 40 billion tests. In this instance, even if only the p-values, SNP ids, and trait ids for each test were saved, nearly one terabyte of storage would be required, and even with such storage capacity the scale of data would be beyond R's current capability to handle. A final drawback to the asymptotic approach is that, because of departures from model assumptions, often the null distribution is not actually uniform as the asymptotic approach assumes. Therefore, while permutation analysis carried out to evaluate the FDR requires more computation, it saves on storage by only storing results for tests significant at a more liberal significance level.

Because some of the known mediators of gene expression, like transcription factor binding sites, enhancers, and silencers, are known to reside proximal to the coding gene, we might hypothesize that eQTL are more likely to be near the gene than elsewhere in the genome [14]. Whether or not this is true on an absolute scale, when considering the relative size of the proximal region we might be more inclined to believe a marginal result near the gene than one of the same size residing in a random location in the genome. For this reason, it is reasonable to apply a more liberal threshold to identify eQTL near genes (putative cis-acting eQTL) than those that are far away or unlinked with the gene itself (trans-acting eQTL). However, the degree to which the threshold is loosened should be data-driven. In other words, the FDR criteria should remain constant between the proximal and distal sets. If, in fact, the density of eQTL is higher near the gene, the p-value corresponding to the fixed FDR will be less stringent.

The procedure for multiplicity adjustment based on this approach is as follows: (1) carry out the association analysis on all traits over all markers storing only results at a liberal significance threshold (e.g., 0.1); (2) perform a reasonable number of permutations by permuting the individual ids that link the individual genotype data to the individual gene expression data such that the marker–marker and trait–trait correlation structures are preserved, again storing only results that meet a liberal significance threshold; (3) for both the observed and permuted data results, identify the proximal and distal trait–marker pairs; (4) separately in the proximal and distal sets, identify the p-value that corresponds to a set ratio of false positives to total positives, where the false positives are estimated from the permuted set, adjusting for the number of permutations. Unlike studies of single traits where many permutations must be carried out to get a handle on the FDR, in the context of tens of thousands of traits the number of p-values computed under the null hypothesis of no association is extreme (in the earlier example greater than one billion p-values are computed), so that even for a small number of permutations the p-value distribution will be very stable. Therefore, as few as five permutations may result in stable FDR estimates as we have shown earlier [40].

5 High-Performance Computing

High-performance computing (HPC) resources have long been used in a number of different research and development settings such as aerospace design, climatology, transportation systems, commerce, particle physics, and protein structure determination. In biological research, the need for large-scale HPC resources accelerated with studies centered on assembling and deciphering various genomes over the past 15 years. Human genetics, however, is now at a similar point where the ability to generate terabytes of data in single experiments is now possible. There are a multitude of open-source, public, and commercial software available for genetic analysis with one or a handful of traits such as R (http://R-project.org), SAS/Genetics$^{\mathrm{TM}}$, HelixTree®, and PLINK [23]. Each of these programs permits varying degrees of customization and optimization in an HPC setting. For studies on GOGE, the number of single point tests can easily exceed 40 billion [40], thus requiring considerable forethought on the execution strategy for even a single pass of association testing, not to mention the execution of genome-wide empirical testing to determine significance which may need to be derived for each trait separately. To increase the computational efficiency of these study types, access to an HPC cluster is ideal.

A cluster is simply two or more computers, usually called nodes, that work together to perform a particular task or set of tasks. In general, there are four major categories of clusters:

- **Storage clusters:** Designed with a common image of the file system across multiple servers. This allows the servers to read/write to a single shared file system and eliminates redundant files and applications.
- **High availability clusters:** Designed to provide continuous availability of service, even in the case of a node failing mid-operation.
- **Load balancing clusters:** Designed to match the number of nodes according to cluster job load.
- **High-performance cluster:** Designed to permit jobs to work in parallel over some number of the nodes, usually to perform concurrent calculations. This enhancement of an analytical application is ideal for large-scale analyses, especially those approaches requiring exhaustive search such as in the analysis of epistatic interactions.

While a single cluster may be designed with one of these categories as its primary focus, most clusters are now designed to reflect some combination of these functions in greater or lesser degrees.

The key component of a cluster is a set of multiple standalone computers, generally UNIX boxes, PCs, or servers. However, additional requirements include cluster-specific operating systems, high-performance interconnects and lots of cable, middleware, parallel programming environments, and applications suited to a particular cluster instance. While there are a number of resources that can provide much greater detail with respect to the appropriate design and components of a cluster [41] and the guiding principles behind some of the different flavors of cluster design (http://www.beowulf.org, http://now.cs.berkeley.edu/, http://hpvm.sourceforge.net/), it is worth making mention here of a least one type of software that when incorporated into the analytical techniques used in large-scale genetic mapping efforts can provide tremendous improvements in the performance for both computational calculation management and speed.

Along with the emergence of the cluster as a viable parallel computing platform, with many HPC clusters now competing directly in the supercomputer space, leaps in efficiencies and performance have been enabled by the simultaneous emergence of message passing libraries. These libraries, such as Message Passing Interface (MPI) [42, 43] and Parallel Virtual Machine (PVM) [44], enable the mapping of parallel algorithms onto large clusters in a portable way. It is generally accepted that these two libraries provide different solutions to the same problem (mapping of parallel algorithms), however, much investigation and discussion has occurred in the computational sciences field as to what each of these libraries actually do and how [45]. Understanding the subtleties that have been uncovered may influence the choice of one software over the other. In the end, these libraries permit a programmer to divide a task, usually a large number of computations or a very complex computation, among a group of networked computers, and then assemble the results of this processing into a coherent set of results. A common misconception is that arbitrary software will run faster on a cluster. Standard analytical programs speed-up will generally scale linearly with batching of jobs and the number of nodes utilized, unless the analytical software is modified to take advantage of the cluster. Specifically, by making direct use of MPI or PVM libraries, modified software can perform multiple independent parallel operations (including the implementation of computationally intensive statistical methodologies) that can be distributed among the available processors, for super-linear gains in speed.

Leveraging HPC resources and code optimization has been explored for QTL mapping in mice, recognizing increases in efficiencies and speed for advanced analyses [46, 47]. Parallelized approaches have allowed for the simultaneous search for multiple QTL in mouse studies [46], while in human GWA studies involving the genetics of gene expression studies, distributing calculations across multiple processors and improving the way in which calculations and data are distributed over multiple processors, have been employed to run large-scale association testing as well as obtain empirical significance thresholds via permutation [40]. Adaptation of available technologies to a framework for large-scale GWAS is necessary, in particular for genetics of gene expression studies, where reconstruction of comprehensive genetic regulatory networks must incorporate the identification of genetic interactions.

6 Further Recommendations for Efficiency Gains in GOGE Studies

Recent large-scale GWAS initiatives have made gains by employing economies of scale in instituting centralized SNP genotyping, data coordination and control centers (http://www.hapmap.org, http://www.wtccc.org.uk/) [48], providing data sets that have undergone common quality control checks and standardized annotation to multiple researchers for individual analysis [31, 49]. Additionally, one of the most recent operational advances is the use of a single large common control population for multiple case–control GWAS studies [31]. The WTCCC effort demonstrates that this is a viable option analytically, and leveraging a common resource for controls, such as Illumina Inc.'s iControlDB (http://www.illumina.com/pages.ilmn?ID = 231) or Affymetrix Inc.'s Control Cohort Initiative (http:///www.affymetrix.com) is certainly an attractive option for financial reasons.

While it is beyond the scope of this chapter to detail a complete roadmap for executing a GWAS, we stress the need to carry out the meticulous capture of the complete data flow, quality checks, and analyses in a tracked, if not automated, fashion. Even on an individual GWA study scale, there are numerous benefits to instituting the approaches discussed herein. Active capture of the complete process will not only aid in the accurate interpretation of the individual study results but will also permit the interpretation of results in a more comprehensive fashion through the integration multiple data sets and results.

7 Constructing Gene Networks to Enhance GWAS and GOGE Results

As discussed, generating a GOGE data set and performing a first-pass analysis on this scale of data is a major undertaking. The identification of or other DNA markers that associate with the expression of one or more genes is a primary goal of a GOGE study. However, if analysis of GOGE data stopped at the identification of SNPs that associate with expression, the true value of these data would not be realized. Genes do not carry out their functions in isolation of other genes, but instead operate in complex networks that together, in a context-specific way, define the complex behavior that emerges from biological systems. Therefore, understanding gene networks in a diversity of contexts will lead to an increased understanding of complex system behavior, including disease.

The reductionist approach to elucidating the complexity of biological systems has motivated straightforward genetic association approaches, where the identification of single genes associated with disease has served as the primary means of getting a foot into pathways for complex phenotypes like disease. However, even in cases where genes are involved in pathways that are well known, it is unclear whether the gene causes disease via the known pathway or whether the gene is involved in other pathways or more complex networks that lead to disease. One example of this is TGFBR2, a recently identified and validated obesity susceptibility gene [13]. While TGFBR2 plays a central role in the well studied TGF-β signaling pathway [50], TGFBR2 and other genes in this signaling pathway are correlated with hundreds of other genes [13, 16], so that it is possible that perturbations in these other genes or in TGFBR2 itself may drive diseases like obesity by influencing other parts of the network beyond the TGF-β signaling pathway. Therefore, considering single genes in the context of a whole gene network may provide the necessary context within which to interpret the disease role a given gene may play.

Constructing gene networks can provide a convenient framework for exploring the context within which single genes operate. A network is simply a graphical model comprised of nodes and edges. For gene networks associated with biological systems, the nodes in the network typically represent genes, gene products, or other important molecular entities like metabolites, and edges (links) between any two nodes indicate a relationship between the two corresponding genes. For example, an edge between two genes may indicate that the corresponding expression traits are correlated in a given population of interest [51–53], that the corresponding proteins interact [54], or that changes in the activity of one gene lead to changes in the activity of the other gene [13]. Interaction or association networks have recently gained more wide-spread use in the biological community, where networks are formed by considering only pair-wise relationships between genes, including protein interaction relationships [19], co-expression relationships [55, 56], as well as other straightforward measures that may indicate association between two genes. In all cases, these networks have been demonstrated to exhibit a scale-free and hierarchical connectivity structures [17, 56, 57], providing higher-level insights into how biological networks may be ordered. The scale-free property exhibited by most biological networks implies that, like the Internet, most genes in a biological system are strongly connected to a small number of genes, while a smaller set of genes (often referred to as hub nodes) are connected to many other genes. The hierarchical property implies that biological networks are highly modular, with genes clustering into groups that are highly interconnected with each other, but not as highly connected with genes in other groups.

7.1 Constructing Weighted and Unweighted Co-Expression Networks

In constructing co-expression networks based on gene–gene interaction strengths, there are two basic approaches: (1) an unweighted network reconstruction approach that involves setting hard thresholds on the significance of the interactions between genes and (2) a weighted network reconstruction approach that avoids hard thresholding. For unweighted gene co-expression networks, gene–gene relationships are encoded in a binary form. That is, two genes in the network are connected by an edge if the correlation coefficient or the significance level of the correlation measure meets some predetermined threshold [57–61]. The drawback of the unweighted approach is that the determination of the hard threshold is somewhat arbitrary and the resulting networks may be sensitive to the threshold selected. More importantly, the binary encoding actually destroys information regarding the interaction strength between two genes, resulting in a loss of power to establish higher-order relationships among the genes in the network. In contrast, the weighted gene co-expression network analysis assigns a connection weight to all pairs of genes by employing soft-thresholding functions whose parameters are estimated based on a biologically motivated scale-free topology criterion [13]. Weighted gene co-expression networks preserve the continuous nature of gene–gene interaction at the transcriptional level and are robust to parameter selection. However, constructing these networks is more computationally intensive as all pairs of nodes are simultaneously considered, so that as the number of nodes grows, the number of pairs to consider grows quadratically.

7.2 Using Genetics in Constructing Co-Expression Networks

Multiple traits driven by common QTL is a central idea that can be leveraged to construct networks. The construction of co-expression networks can be aided by the introduction of genetic data, which at the very least can serve as a filter to help reduce artifactual correlations between expression traits. Significant artifactual correlations can arise in larger-scale gene expression studies because of correlated noise structures between the array-based experiments in such studies. Therefore, one of the more straightforward ways to leverage the eQTL data in this setting is to simply filter out gene–gene correlations in which the expression traits are not at least partially explained by common genetic effects [57]. For example, we can connect two genes with an edge in an unweighted co-expression network if (1) the p-value for the Pearson correlation coefficient between the two genes is less than some prespecified threshold and (2) the two genes have at least one common eQTL. This can be taken a step further by formally assessing whether two expression traits driven by a common QTL are related in a causal or reactive fashion, filtering out correlations driven by expression traits that are independently driven by common or closely linked QTL [13, 62].

One intuitive way to establish whether two genes share at least one eQTL, is to carry out single trait eQTL mapping for each expression trait and then consider eQTL for each trait overlapping if the corresponding LOD for the eQTLs are above some threshold and if the eQTL are in close proximity to one another. The significance of the statistic corresponding to the strength of association between two genes in the co-expression networks is then chosen such that the resulting network exhibits the scale-free property [56, 57, 63] and the false discovery rate for the gene–gene pairs represented in the network is constrained. Beyond this simple, albeit intuitively appealing, eQTL overlap method, we can formally test whether two overlapping eQTL represent a single eQTL or closely linked eQTL by employing a pleiotropy effects test (PET), such as that originally described by Jiang and Zeng [64, 65]. The formation of gene clusters by simultaneously considering gene–gene and marker–gene correlations also promises to provide a more comprehensive characterization of shared genetic effects [66].

7.3 Identifying Modules of Highly Interconnected Genes in Co-Expression Networks

Given the scale-free and hierarchical nature of co-expression networks [17, 56, 57], one of the key problems is to identify the network modules, or functional units, in the network that represent those hub nodes (nodes that are significantly correlated with many other nodes) that are highly interconnected with one another, but that are not as highly connected with other hub nodes. Figure 1 illustrates a topological connectivity map for the most highly connected genes in the liver tissue of a previously described human liver cohort [40]. After hierarchically clustering both dimensions of this plot, the network is seen to break out into clearly identifiable modules. Gene–gene co-expression networks are highly connected, and the clustering results shown in Fig. 1 illustrate there are gene modules arranged hierarchically within these networks.

Ravasz et al. [67] used manually selected height cutoff to separate tree branches after hierarchical clustering, in contrast to Lee et al. [68] who formed maximally coherent gene modules with respect to GO functional categories. Another strategy is to employ a measure similar to that used by Lee et al. [68], but without the dependence on the GO functional annotations, given it is of interest to determine independently whether co-expression modules are enriched for GO functional annotations [57]. An emerging trend for module identification is to uncover alternative network structures such as cores and cliques and high-level organization forms like overlapping modules (communities) [69, 70]. The modules identified in this

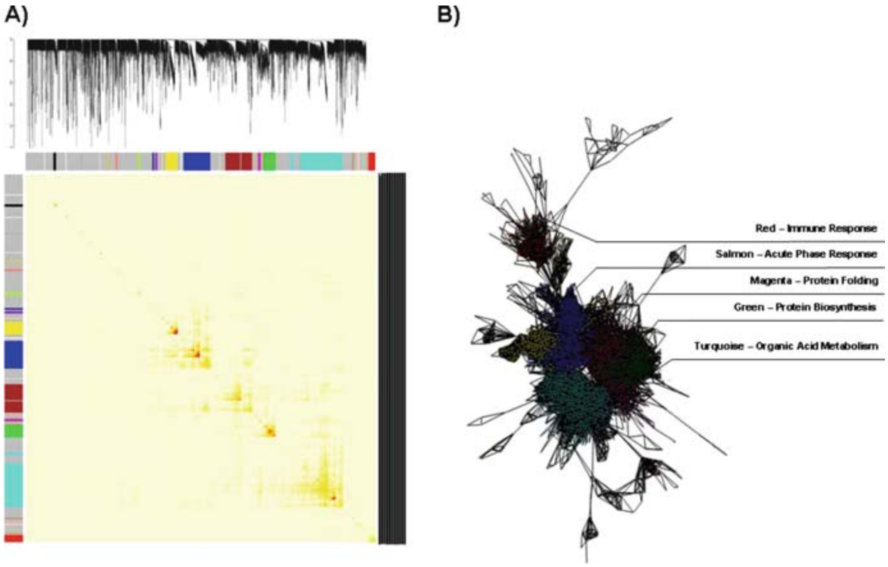


Fig. 1 A co-expression network and corresponding functional modules constructed from a previously described human liver tissue cohort (HLC) [40]. (**a**) The hierarchically clustered topological overlap matrix along with the identified functional modules in the gene co-expression network comprised of the top 25% (10,025) most differentially expressed genes in the HLC. Genes in the rows and columns are sorted by an agglomerative hierarchical clustering algorithm. The different shades of color signify the strength of the connections between the nodes (from white signifying not significantly correlated to red signifying highly significantly correlated). The hierarchical clustering and the topological overlap matrix strongly indicate highly interconnected subsets of genes (modules). Modules identified are colored along both column and row. (**b**) The corresponding graph of the HLC co-expression network. The colors of the nodes represent their module assignments as described in (**a**). The functional categories denoted for some of the modules represent the most enriched GO Biological Process category for the module

way are informative for identifying the functional components of the network that are associated with disease [57]. It has been demonstrated that the types of modules depicted in Fig. 1 are enriched for known biological pathways, for genes that associate with disease traits, and for genes that are linked to common genetic loci [56, 57]. In this way, one can identify those key groups of genes that are perturbed by genetic loci that lead to disease, and that therefore define the intermediate steps that actually define disease states.

8 Looking Toward the Future: Probabilistic Causal Networks

The co-expression networks are a useful construct for understanding the overall connectivity structure of networks that drive complex phenotypes like disease. However, they are still interaction based and so do not provide the detailed resolution needed to understand how any particular gene can induce changes in other genes, and, more generally, get at models that actually predict complex system behavior. The present day challenge is to study the biological functions driven by the different regions of the genome, determining whether such regions encode for a protein or noncoding RNA, the functional role played by a given protein or RNA, the biological processes related to this function, and so on. There are continually growing numbers of systematically generated data, including gene expression (transcriptomics); protein–protein interaction assessed by yeast two-hybrid; protein identification, quantification and post-translation modification identification by mass-spectrometry (proteomics), and more recently metabolite levels measured by NMR or mass-spectrometry (metabolomics). To assess the function of individual genes, compendiums of yeast gene knockout [71] and mouse gene knockout (e.g., DeltaBase) have been constructed, in addition to global synthetic fitness or lethality (epistatic interaction) screens [72]. Further, there are efforts to systematically collect and manually curate knowledge and to represent such data into easily accessible databases, such as KEGG [73], BioCarta (http://www.biocarta.com), MetaCore (http://www.genego.com), and Ingenuity (http://www.ingenuity.com). With all of these efforts, integrating these types of high-throughput data is critical if we hope to construct models that are predictive of complex biological systems. Ideker et al. [74] integrated genomics, gene expression, and proteomics data to study small networks, refining such networks in a trial-and-error manner using experimental approaches. Systematically integrating different types of data into probabilistic networks using Bayesian networks has been proposed and applied for the purpose of predicting protein–protein interactions [75] and protein function [68]. However, these Bayesian networks are still based on association between nodes in the network as opposed to causal relationships. From these types of networks, we cannot infer whether a specific perturbation will affect a complex disease trait or not. To make such predictions we need networks capable of representing causal relationships. Probabilistic causal networks are one way to model such relationships, where causality in this context reflects a probabilistic belief that one node in the network affects the behavior of another. We anticipate these types of networks becoming increasingly important in the human genetics space to gain a mechanistic understanding of how a given DNA perturbation induces changes in one or more genes that go on to affect networks that cause disease. The integration of genotypic and expression and other data have recently been shown, in a Bayesian network framework [76], to enhance the overall accuracy of predictive networks [40, 51–53]. We have also recently demonstrated how this class of network can be used to inform associations identified in GWA studies [40].

9 Summary

The significant challenge we face in the post-genome era is deciphering the biological function of individual genes, pathways, and networks that drive complex phenotypes like disease. The availability of low-cost, high-throughput technologies for genotyping hundreds of thousands of DNA markers has led to a number of successes in identifying associations between these markers and complex traits like age-related macular degeneration [1–3], diabetes [4, 5], and obesity [6], validating the GWAS approach as the best human genetics approach for identifying genetic loci that associate with disease. However, while this approach has now delivered and will continue to deliver loci for almost all common human disease, GWA studies on their own cannot typically elucidate the functional role the underlying gene or genes play in disease, and, in fact, cannot usually lead to a definitive identification of the susceptibility gene or genes, given a lack of experimental support for the functional consequences of a given DNA variation on gene function.

The genetics of gene expression studies discussed herein provides an alternative to the forward genetics approach to dissecting complex traits like disease. The information that defines how variations in DNA lead to variations in complex traits of interest flows through molecular networks that actually define the complex traits. Therefore, characterizing the molecular networks that underlie complex traits like disease can provide a more comprehensive view of disease, and this in turn can lead to the direct identification of key genes underlying disease processes, as well as providing a rich biological context within which to infer the functional roles played by these key genes. Because complex biological processes that lead to disease are often system and context dependent, leveraging DNA variations as a systematic source of perturbations on molecular networks and clinical traits facilitates studying complex biological processes at the systems level, in addition to studying gene function at the level of individual pathways [15, 16]. However, genetics of gene expression studies can involve levels of data management and analysis that go beyond the current capabilities of most human genetics groups. We have discussed a number of issues related to effectively carrying out this type of study and leveraging the results to inform more standard human genetic association studies, including data management, data QC, the need for high-performance computing, analysis issues related to simple associations between SNPs and expression traits, and the construction of gene networks to identify whole networks that associate with disease. These represent the beginning steps that can be taken to leverage GOGE data, and certainly much of the future of human genetic studies will involve more integrative analyses that seek to inform how variations in DNA impact networks that go on to cause disease.

With large-scale molecular profiling, genotypic and clinical data collected from large-scale human and experimental populations, focusing on how a single protein or RNA impacts disease will ultimately give way to how a network of gene interactions impacts disease. The integration of genetic, molecular profiling, and clinical data has the potential to paint a more detailed picture of the particular network states that drive disease, and this in turn has the potential to lead to more progressive treatments of disease that may ultimately involve targeting of whole networks as opposed to current therapeutic strategies focused on targeting one or two genes [77].

Web Resources

Here, we list a number of software resources we have found useful in the analysis of large-scale GWAS and GOGE data. This list is not intended to be exhaustive, but instead offers a handful of starting points to help guide the interested reader. There are many other tools in each of the areas mentioned below that are not listed but that are extremely useful depending on the specific study and/or problem. Therefore, we encourage the reader to use this list only as a jumping off point.

Resources for carrying out genetic analyses (described in the text):

- **PLINK**: http://pngu.mgh.harvard.edu/∼purcell/plink/
- **QUANTO**: http://hydra.usc.edu/gxe
- **SOLAR**: http://www.sfbr.org/solar/index.html
- **STRUCTURE**: http://pritch.bsd.uchicago.edu/software.html
- **SVA package for R**: http://www.genomine.org/sva/

Resources for accessing raw data and results from GWAS and GOGE studies:

- **WTCCC**: http://www.wtccc.org.uk
- **GAIN**: http://www.fnih.org/GAIN2/Overview_description.shtmldb
- **GAP**: http://www.ncbi.nlm.nih.gov/sites/entrez?db = gap
- **Illumina**: http://www.illumina.com/pages.ilmn?ID = 231Can
- **webQTL**: http://www.genenetwork.org/

Resources for gene expression data:

- **The Gene Expression Omnibus** (GEO) houses many of the gene expression data sets from published studies and provides tools to facilitate mining of these data: http://www.ncbi.nlm.nih.gov/geo/
- **ArrayExpress** is similar to GEO in serving as a warehouse for gene expression and associated data: http://www.ebi.ac.uk/microarray-as/aer/

- **SAGEmap** provides tools to carry out differential gene expression analyses on SAGE (Serial Analysis of Gene Expression) data: http://www.ncbi.nlm.nih.gov/sage/
- **The Cancer Genome Anatomy Project** (CGAP) provides access to extensive gene expression data in normal, precancerous, and malignant cells from a number of tissues: http://www.ncbi.nlm.nih.gov/ncicgap/

Resources for annotating gene sets:

- **Kyoto Encyclopedia of Genes and Genomes** (KEGG) provides extensive pathway and gene function information: http://www.genome.jp/kegg/
- **BioCart**: http://www.biocarta.com
- **MetaCore**: http://www.genego.com
- **Ingenuity**: http://www.ingenuity.com
- **The Gene Ontology** (GO) provides a controlled vocabulary for describing genes and the processes in which they are involved: http://www.geneontology.org/
- **The Database for Annotation, Visualization and Integrated Discovery** (DAVID) provides a number of tools for functionally annotating and classifying experimentally derived gene sets: http://david.abcc.ncifcrf.gov/
- **Cytoscape** is a tool to visualize molecular interaction data and to integrate interaction data with molecular profiling data: http://www.cytoscape.org/

BioCarta (http://www.biocarta.com), MetaCore (http://www.genego.com), and Ingenuity (http://www.ingenuity.com).

References

1. Edwards AO et al. (2005) Complement factor H polymorphism and age-related macular degeneration. Science 308:421–424
2. Haines JL et al. (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308:419–421
3. Klein RJ et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
4. Grant SF et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38:320–323
5. Sladek R et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885
6. Herbert A et al. (2006) A common genetic variant is associated with adult and childhood obesity. Science 312:279–283
7. Peacock ML, Warren JT Jr, Roses AD, Fink JK (1993). Novel polymorphism in the A4 region of the amyloid precursor protein gene in a patient without Alzheimer's disease. *Neurology* 43, 1254–1256.
8. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755
9. Bystrykh L et al. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet 37:225–232
10. Chesler EJ et al. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37:233–242

11. Monks SA et al. (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75:1094–1105
12. Morley M et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747
13. Schadt EE et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717
14. Schadt EE et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302
15. Hartwell LH, Hopfield JJ, Leibler SMurray A.W (1999) From molecular to modular cell biology. Nature 402:C47–52
16. Schadt EE, Sachs A, Friend S (2005) Embracing complexity, inching closer to reality. Sci STKE 2005:pe40
17. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
18. Zerhouni E (2003) Medicine. The NIH Roadmap. Science 302:63–72
19. Han JD et al. (2003) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430:88–93
20. Luscombe NM et al. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431:308–312
21. Chen Y et al. (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452:429–435
22. Zhao LJ et al. (2005) SNPP: automating large-scale SNP genotype data management. Bioinformatics 21:266–268
23. Purcell S et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.
24. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
25. BRLMM: an Improved Genotype Calling Method for the GeneChip® Human Mapping 500K Array Set (Affymetrix, 2006)
26. Carvalho B, Bengtsson H,, Speed TP, Irizarry RA (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 8:485–499
27. Hua J et al. (2007) SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. Bioinformatics 23:57–63
28. Liu WM et al. (2003) Algorithms for large-scale genotyping microarrays. Bioinformatics 19:2397–2403
29. Rabbee N, Speed, TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12
30. Teo YY et al. (2007) A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 23:2741–2746
31. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
32. Sieberts SK, Schadt EE (2007) Moving toward a system genetics view of disease. Mamm Genome 18:389–401
33. He YD et al. (2003) Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. Bioinformatics 19:956–965
34. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3:1724–1735
35. Emilsson V et al. (2008) Genetics of gene expression and its effect on disease. Nature 452:423–428
36. Yang X et al. (2006) Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Res 16:995–1004
37. Wang S et al. (2006) Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. PLoS Genet 2:e15
38. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. JRSS B 57:289–300

39. Storey JD (2002) A direct approach to false discovery rates. JRSS B 64:479–498
40. Schadt EE et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6:e107
41. Yeo C et al. (2006) Cluster computing: high-performance, high-availability, and high-throughput processing on a network of computers. In Zomaya A (ed) Handbook of nature-inspired and innovative computing, pp 521-55142. Message Passing
42. Interface Forum. MPI (1994) A message-passing interface standard. Int J Supercomputer Appl 8:165–414
43. Message Passing Interface Forum. MPI2 (1998) A message passing interface standard. Int J High Performance Comput Appl 12:1–299
44. Geist A et al. (1994) PVM: Parallel Virtual Machine—a user's guide and tutorial for network parallel computing, MIT, Cambridge, MA
45. Gropp W, Lusk E (2002). Goals guiding design: PVM and MPI
46. Carlborg O, Andersson-Eklund L, Andersson L (2001) Parallel computing in interval mapping of quantitative trait loci. J Hered 92:449–451
47. Jayawardena M, Ljungberg K, Holmgren S (2007) Using parallel computing and grid systems for genetic mapping of quantitative traits. In Applied parallel computing. State of the art in scientific computing, vol Volume 4699/2007 627–636, Springer, Berlin
48. University of Washington, Fred Hutchinson Cancer Research Center to coordinate National Human Genome Research Institute disease studies (2007)
49. Tanaka T (2005) [International HapMap project]. Nippon Rinsho 63(12):29–34
50. Ramji DP, Singh NN, Foka P, Irvine SA, Arnaoutakis K (2006) Transforming growth factor-beta-regulated expression of genes in macrophages implicated in the control of cholesterol homoeostasis. Biochem Soc Trans 34:1141–1144
51. Zhu J et al. (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res 105:363–374
52. Zhu J et al. (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. PLoS Comput Biol 3:e69
53. Zhu J et al. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40:854–861
54. Kim JK et al. (2005) Functional genomic analysis of RNA interference in C. elegans. Science 308:1164–1167
55. Gargalovic PS et al. (2006) Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proc Natl Acad Sci U S A 103: 12741–12746
56. Ghazalpour A et al. (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet 2:e130
57. Lum PY et al. (2006) Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. J Neurochem 97(1):50–62
58. Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000:418–429
59. Davidson EH, McClay DR, Hood L (2003) Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Sci U S A 100:1475–1480
60. Bergmann S, Ihmels, J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2:E9
61. Carter SL, Brechbuhler CM, Griffin M, Bond A.T (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20:2242–2250
62. Doss S, Schadt EE, Drake TA, Lusis AJ (2005) Cis-acting expression quantitative trait loci in mice. Genome Res 15:681–691
63. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
64. Jiang C, Zeng ZB (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140:1111–1127

65. Zeng ZB (1993) Precision mapping of quantitative trait loci. Genetics 121:185–199
66. Lee SI, Pe'er D, Dudley A.M, Church GM, Koller D (2006) Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc Natl Acad Sci U S A 103:14062–14067
67. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555
68. Lee I, Date, SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306:1555–1558
69. Wuchty S, Almaas E (2005) Peeling the yeast protein network. Proteomics 5:444–449
70. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818
71. Hughes TR et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102:109–126
72. Pan X et al. (2006) A DNA integrity network in the yeast Saccharomyces cerevisiae. Cell 124:1069–1081
73. Kanehisa M et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34:D354–D357
74. Ideker T et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–934
75. Jansen R et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302:449–453
76. Pearl J (1998) Probabilistic reasoning in intelligent systems: networks of plausible inference, xix, p 552, Morgan Kaufmann, San Mateo, CA
77. Schadt EE, Lum PY (2006) Reverse engineering gene networks to identify key drivers of complex disease phenotypes. J Lipid Res 47:2601–2613
78. Almasy L, Blangero J (1998) *Multipoint quantitative-trait linkage analysis in general pedigrees.* Am J Hum Genet 62:1198–211
79. Price AL et al. (2006) Principle components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Index