Statistics for Biology and Health

For further volumes:
http://www.springer.com/series/2848

Nan M. Laird · Christoph Lange

The Fundamentals of Modern Statistical Genetics

Springer

Nan M. Laird
Department of Biostatistics
Harvard University
Boston, MA 02115, USA
laird@hsph.harvard.edu

Christoph Lange
Department of Biostatistics
Harvard University
Boston, MA 02115, USA
clange@hsph.harvard.edu

ISSN 1431-8776
ISBN 978-1-4419-7337-5 e-ISBN 978-1-4419-7338-2
DOI 10.1007/978-1-4419-7338-2
Springer New York Dordrecht Heidelberg London

Printed on acid-free paper

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

To our families, and all of the families whose data we have analyzed.

Preface

Statistical genetics has played a pivotal role for more than a century in the discovery of genes that cause disease in humans. Driven by advances in molecular genetics and medicine and the continuing improvements in genotyping technology, statistical models and methods have adapted over time to the challenges presented by new study designs.

In this book we discuss the statistical models and methods that are used to understand human genetics from an historical perspective. Starting with Mendel's first experiments to more recent genome-wide association studies, we describe how genetic information can be incorporated into statistical models to discover disease genes. While we cover most of the commonly used approaches in statistical genetics (e.g., aggregation analysis, segregation, linkage analysis, etc.), the focus of the book is on modern approaches to association analysis. Our treatment of earlier topics is mainly to help the reader see the larger picture and understand the historical development of methods. We provide numerous examples to illustrate key points throughout the text, both of Mendelian and complex genetic disorders.

Most statisticians, biostatisticians and data analysts are aware of the key role that their disciplines have played in finding disease genes, but have little direct knowledge of how gene discovery via gene mapping works. This book arises from teaching courses to graduate students, with varying levels of statistical preparation, at the Harvard School of Public Health. Our intended audience for this book is largely quantitatively oriented health scientists, including biostatisticians, statisticians, epidemiologists, physicians and molecular geneticists, who want to learn about statistical methods for genetic analysis, whether to better analyze genetic data, or to pursue research in methodology. We assume familiarity with elementary probability, statistical inference and methods, specifically distributions for two or more variables, conditional, marginal and joint distributions, Bayes rule, likelihood methods, hypothesis testing, estimation, correlation and the essential ideas of regression, including linear, log-linear and logistic. However, the book emphasizes concepts and examples, and the exercises include problems for students with a broad range of skill levels. We assume no formal training in genetics, but familiarity with basic concepts in molecular genetics is necessary and will be reviewed in the first chapter.

There are many excellent texts in statistical methods currently available to students and we have used many of them in our teaching. This book shares much with the classic texts of Sham (1998) and Lange (2002), both of which were written with a similar audience in mind. Our book is less focused on linkage and more focused on association analysis than the text by Sham, and provides easier reading for students with less mathematical training than the book by Lange. We also share much with the newer texts by Thomas (2004) and Yang (2000), being less epidemiologically oriented than Thomas, with more emphasis on human disease than Yang. The book by Foulkes (2009) has a stronger emphasis on software implementation while our focus is on statistical theory and methods.

Boston, Massachusetts Nan M. Laird
Bad Godesberg, Germany Christoph Lange

Acknowledgments

This book would not have been possible without our former students, the staff at the Harvard School of Public Health, our colleagues, friends and our families whom we would like to thank for their the support, encouragement and help during the writing process of the book. The book is based on the statistical genetics courses that we have been teaching over the last 10 years here at the Harvard School of Public Health. With their feedback and during enjoyable discussions, many of our former students helped us to improve our courses and thereby the book. Without the active and patient support from the staff at the Department of Biostatistics at the Harvard School of Public Health, especially Jelena Follweiler, we would not have been able to put the book together. We also thank the reviewers, our colleagues, friends and family members for their keen eye during the review process of the book. We are especially thankful to Kaustubh Adhikari, Gourab De, Matt McQueen, Jessica Lasky-Su and Tony Paredo for their help with the exercises, Ross Lazaras for help with figures, and to Wai-Ki Yip, Deborah Blacker, Garrett Fitzmaurice, Jonathan Haines, Alkes Price, Benjy Raby, David Alexander, Lily Altstein, Sharon Lutz, Cory Zigler, Andreas Kraeusling, and several anonymous reviewers for their comments on drafts of the book. We are also indebted to Tyler VanderWeele for contributing Section on Compositional Epistatis and Compositional Gene-Environment Interactions. Last, but not least, we would like to mention John Kimmel for leading us through the book writing process.

Contents

Chapter 1
Introduction to Statistical Genetics and Background in Molecular Genetics

An understanding of the basic ideas of inheritance has been evident throughout the history of mankind, ever since the domestication of animals or the practice of farming began. The Babylonians and ancient Egyptians utilized cross pollination of crops and selection of domesticated animals for breeding, but did not develop a formal theory for the principles underlying the inheritance of traits. Later, ancient Greek philosophers developed elementary theories to explain how inheritance worked in humans, grappling unsuccessfully with the apparent paradox that inherited characteristics can sometimes differ between offspring and parents. Some diseases in humans, such as sickle cell anemia and hemophilia, have been recognized as inherited disorders for centuries and, as the science of medicine developed, so too did the recognition that many diseases are heritable. Yet, bipolar disorder, one of the oldest known disorders in humans, was not widely regarded as heritable until the 1950s.

Although we can document an awareness of the basic concept of inheritance for millennia, most of our current knowledge about inherited human diseases has been acquired only in the last century. As the concept of inherited disease gradually developed, *Genetics*, the science of inherited variation and heritable biological material in living organisms, became an integral part of the search for the origin of disease. Today stories of gene discovery for many diseases dominate the news landscape. Despite centuries of formal and informal observation of patterns of inheritance in humans, the discipline of human genetics is relatively young. Humans are difficult to study because, in contrast to plant and animal genetics, experimental crossings are not possible, environmental factors are hard to control, and humans have small families with many years required for a new generation to develop. Environmental and genetic factors broadly overlap during childhood, making it difficult to separate the relative contributions of the environment and genetics to the development of disease. As a result, much of our understanding of basic genetic principles and how genes affect variation in organisms comes from experimental studies in plants (Mendel's experiments with garden peas in the 1860s) and animals (Morgan's experiments in the 1920s with flies). Mendel's laws were initially largely ignored, but 'rediscovered' by scientists in the 1900s and hotly debated by geneticists, biologists, statisticians and biometricians. Part of this debate centered on the apparent conflict between Mendel's work *Experiments in Plant Hybridization* (1865) and

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_1,

Darwin's theories set forth in the *Origin of the Species* (1859), which was published just prior to Mendel's paper. Darwin used the notion of inherited traits as the basis for natural selection, but he believed that traits in parents were 'blended' in the offspring. Mendel's work verified the inheritance of traits, but he deliberately used discrete traits that were not blended in offspring. Developing models and theories for how Mendel's discrete inherited units could explain variation in continuous human characteristics was a subject of much debate during these early years of statistical genetics. In this text we use the term trait broadly to encompass both measured and discrete characteristics, as well as disease outcomes.

1.1 Basic Concepts in Genetic Disease

Statistical Genetics is a branch of statistics that deals with the analysis of inherited traits and genetic data. We use genetic data loosely here to refer to the biological material that is inherited during reproduction via egg and sperm cells. In early days, statistical genetics was largely dominated by statistics for experimental studies in plants and animals. Galton's statistical work in the 1880s on the inheritance of height in humans is an important exception to this rule. Over the years, the methodological focus of statistical genetics has changed to keep pace with the different kinds of genetic data that technology has made available. Most recently, new technologies arising from the Human Genome Project and HapMap Project have generated a surge of methodological development to address unsolved problems in human genetics. The development of statistical models and methods to explain how genes influence traits continues to be a common goal in plant, animal and human genetics.

When the discipline of statistical genetics was just beginning, we had little understanding of the basic biological underpinning of genetics and inheritance apart from the fact that humans had 'units'–later termed 'genes'–that were inherited from their parents and that 'units' could differ from person to person. Most important from the statistical point of view, there was no standardized way to assay or characterize the genetic information at the molecular level in an individual. The available data for most statistical investigations consisted only of *traits*, also known as *phenotypes*. We use the terms traits and phenotypes here to mean individual characteristics, not observed at the molecular level, which are thought to have a heritable basis. For example, a person's blood type at the ABO locus is a phenotype which depends upon their variants at the ABO gene. A person's phenotype (here blood group) can be obtained without knowledge of their gene variant. However, knowing a person's blood type will imply something about the information encoded in their ABO gene. In these early years, statistical genetics was focused on methods for determining if traits or diseases were inherited and measuring the degree of inheritance (studies of *familial aggregation*), and to determine the underlying genetic model that explains the relationship between the phenotype and the underlying disease (*segregation analysis*). For these analyses, individuals with the disease, called *probands*, were identified; information on relatives of the probands was used to form family or pedigree structures. The term *ascertain* is used when referring to probands to indicate that the selection of individuals for study may depend on their phenotypes; depending upon study objectives, all available relatives of probands may be included in the study regardless of phenotype. The phenotypes or traits of the relatives and their familial relationships were exploited in a segregation analysis to infer the underlying genetic model.

Advances in our understanding of the biology of genetics and in laboratory technology have enabled us to now readily obtain data directly on gene variants, called genetic *markers*, at specific locations in the genome. Having marker data for samples of families has enabled *gene mapping*, which encompasses a variety of methods used to find the chromosomal location of a disease-causing gene. An early statistical approach to gene mapping was to use *linkage analysis* of pedigrees. Linkage analysis uses marker data and the traits of a pedigree; by studying the patterns of co-inheritance of the marker and the disease throughout the pedigree, we can infer how far the disease gene is from the marker. Linkage analysis relies on both Mendel's laws discovered in his pea experiments and the concept of genetic distance elucidated in Morgan's fly experiments. Many variants underlying genetic disorders have been discovered using the basic idea of linkage analysis, e.g., cystic fibrosis, Huntington's disease and rare variants underlying early onset Alzheimer's disease.

The genetic architecture of diseases in humans ranges from diseases that are caused by just a single disease variant in a single gene to settings where multiple variants in multiple genetic loci can contribute to the disease risk, often interacting with environmental factors. Diseases or disorders which are initiated by variants in a single gene are typically rare and severe conditions, e.g., Cystic Fibrosis, Duchenes' Muscular Dystrophy and Sickle Cell Anemia. Such diseases are often referred to as *Mendelian disorders or diseases*. Mendelian disorders and diseases follow simple Mendelian patterns of inheritance in families and generally do not have any other causes other than the genetic disease variant. Linkage analysis has been very successful in finding genes for *Mendelian disorders*.

Most common diseases, e.g., asthma, obesity, Alzheimer's disease, bipolar disorder, etc., fall into the category of *multi-factorial diseases* or *complex genetic diseases*. Here, disease risk is thought to be influenced by a set of genes and environmental factors which may interact with each other. Although this book is focused on the concepts of gene mapping for complex diseases, the basic genetic principles of inheritance of genetic material are the same for both Mendelian and complex diseases. Linkage analysis has been less successful with finding genes for complex disorders.

Today gene mapping involves scanning the entire human genome at hundreds or thousands or even millions of genetic markers in the genomes of large samples in order to look for genetic variation associated with disease traits (*Genome Wide Association Studies* (GWAS)). Such studies have led to new findings for many complex diseases: cancer, diabetes, eye disease, obesity and others. These *Genetic Association* studies are also a component of gene mapping; their current popularity stems from the advances in genotyping and from information about the structure of genetic variation captured in the HapMap Project. Genetic association analysis is distinct from virtually all other types of statistical genetics analyses in that it can be carried out using samples of unrelated individuals rather than families or pedigrees. Genetic association analysis using both unrelated samples, and using samples of families, is the main focus of this book.

Population Genetics is concerned with the genetic variation within and between populations, over time and space. This includes modeling variation in genes due to many factors: selection of certain variants due to response to environmental conditions, in- and out-migration, drift occurring in small populations, and mutations, as well as understanding genetic differences in populations. There are some key principles of population genetics, namely Hardy-Weinberg equilibrium, linkage equilibrium and population substructure, which are important in association analysis and will be covered in a short introductory chapter.

Genetic Epidemiology is a branch of epidemiology that deals with both genetic and environmental contributions to disease. Genetic epidemiology uses methods from statistical genetics and epidemiology to understand the interplay between genes, environment and disease. Sometimes data on geographic, spatial, temporal and/or racial, as well as familial, variation in disease rates can provide insight into the genetic nature of disease.

We close this section with two examples of genetic diseases which illustrate some of the ideas discussed in this section.

Sickle Cell Anemia. Sickle cell anemia is a Mendelian disorder that affects red blood cells and is associated with severe morbidity, including pain, hemolytic anemia and infections; without proper medical management, the death rate is high. This disorder is a common textbook example of a genetic disease because it was the first to be labeled a molecular disorder resulting from a genetic mutation. The disorder was widely recognized as an inherited disorder for centuries by indigenous people in sub-Saharan Africa because of the way it occurred in families, but reports did not appear in the Western medical literature until the early twentieth century. By this time, the shape of red blood cells could be seen under a microscope, and scientists discovered that the red blood cells of those suffering from sickle cell disorder had a sickle shape, rather than the normal round shape. This phenomenon was calling ‘sickling’ and individuals with red blood cells which could be made to sickle were labeled ‘sicklemics’.

Segregation analyses of African and African-American pedigrees (see Section 4.4) played an important early role in demonstrating that the disorder was genetic and in understanding its inheritance mechanism. Segregation analyses of African and African-American pedigrees done in the 1920s used the sickling trait, ignoring the fact that not all sicklemics had sickle cell disorder. Some 25 years later, segregation analyses using sickle cell disease as the trait correctly identified the genetic nature of sickle cell disorder. About this time, laboratory studies showed that sickling was due to a genetic variant which changed the molecular structure of hemoglobin, enabling scientists to limit their search to the hemoglobin gene without any linkage studies. A decade later, the specific variant in the hemoglobin gene on chromosome 11 was located.

A second reason for the popularity of the sickle cell example is that it illustrates the phenomenon of *selection*. The variant causing the sickle trait protects individuals against the malaria parasite *Plasmodium falciparum*, which is found largely in India and Africa. This explains the high prevalence of the variant, and the disorder, in those regions and gives rise to the concept of 'heterozygote advantage', which we will discuss in subsequent chapters.

Alzheimer's Disease. Alzheimer's Disease (AD) is a complex disorder with a strong genetic component; it is one of the first complex disorders where multiple genes explaining some of the AD risk were found. It is a brain disorder with progressive destruction of brain cells leading to loss of memory and other cognitive functions, social impairment and eventually death. It is the single largest cause of dementia and there is no known cure. AD was first described in 1906; its hallmark characteristic is the presence of plaques and tangles in the brain at autopsy. Alzheimer's disease is typically a disease of old age, but in a small fraction of cases it occurs as early as the late thirties or forties. Early-onset AD, particularly prior to age 50, is more likely to have a family history consistent with Mendelian inheritance and is often referred to as 'familial' AD. A large number of very rare variants in three genes have been identified which cause early onset AD in a Mendelian fashion, typically at ages earlier than 60. At present, over 200 of these rare variants in these three genes have been reported in only about 500 families world-wide.

Far more common is late onset disease; here advancing age is the primary risk factor, with a risk of nearly 50% in individuals over age 85. Aside from age, family history is probably the strongest risk factor for late-onset Alzheimer's disease. There are also 'environmental' risk factors that appear to enhance the risk of the late onset form of the disease including head injury and a variety of cardiovascular risk factors such as high blood pressure and diabetes.

Many studies involving familial aggregation, segregation, linkage and association have been used in the search for genes contributing to AD. Studies of familial aggregation identified a genetic component to the disorder and provided the justification for molecular studies. During the period around 1990, several linkage studies using AD pedigrees showed evidence for linkage to different chromosomes-14, 19 and 21. In some cases, the linkage results were consistent only in selected families; there was particular controversy and lack of replication for the linkage on chromosome 21, because several early AD families not included in the linkage analysis showed no evidence of linkage to the region implicated on chromosome 21.

Ultimately, the recognition that multiple genes on different chromosomes were involved, each with different variant rates and/or disease risks, enabled discovery of four genes: three genes with rare disease variants which are largely responsible for familial AD and one gene primarily responsible for late onset disease. The three genes involved in familial AD show very predictable patterns of inheritance in large pedigrees, making them ideal candidates for simple linkage analysis. Two genes for familial AD, the Amyloid beta Precursor Protein (APP) gene on chromosome 21 and the Presenilin 1 (PSEN1) gene on chromosome 14, were found via successful linkage analyses. A variant in the APP gene on chromosome 21 was found using linkage with only one large family. A third familial AD gene, Presinlin 2 (PSEN2), was found after a linkage analysis of a small number of families with a common ancestry identified a linked region on chromosome 1. Disease variants in that gene were quickly identified using homologies with the disease variant on the PSEN1 gene on chromosome 14.

The story with the gene affecting primarily late onset AD, Apolipoprotein E (APOE) on chromosome 19, is quite different. With disease onset at the end of the life span, the disease status of many pedigree members is unknown, and there is no clear pattern of inheritance among these late onset families. Connecting APOE gene variants with AD involved newly developed statistical methods for linkage studies of complex disease, biological clues, and serendipity, but most convincing was a series of association studies involving both families, and cases and controls, showing an association between late onset AD and genetic variants of APOE. This gene is sometimes referred to as a susceptibility gene since having a particular variant in this gene enhances risk of AD, but does not determine AD with certainty.

The hunt for additional AD genes continues actively; over 100 genes beyond APOE have been reported to be associated with AD, but until recently none has been consistently confirmed. However, recently two GWAS identified association in the APOJ gene that replicates consistently across several studies.

1.2 Review of Molecular Genetics

This section and the next serve as the basic background material in biology needed for the remainder of the book. Individuals with no prior exposure to the concepts may find these sections difficult to absorb on first reading. It may be necessary to reread these sections while covering later material.

The *human genome* refers to all of the basic biological material that is transmitted from parents to offspring, determining their inherited characteristics. The heritable material is stored on *chromosomes* in the nucleus of every cell. There are 23 pairs of chromosomes in the human genome; 22 of the pairs are *autosomal*, consisting of non-identical copies (i.e., the two copies may have different variants) of the same chromosome while the 23rd pair contains the sex chromosomes (Fig. 1.1). The two non-identical chromosomes in an autosomal pair are referred to as *homologous* chromosomes. For the sex chromosomes, females have two non-identical copies of the X chromosome, while males have one X and one Y. The *centromere* of a chromosome is a region found near the middle of the chromosome; it plays an important role in cell division and reproduction. It also is used to specify genetic locations as it divides each chromosome into a short arm (p for petit) and a long arm (q, next after p in the alphabet). The banded regions shown in Fig. 1.1 can be seen under a microscope after staining; they are also used in specifying genetic locations.

Each chromosome is composed of long strands of *DeoxyriboNucleic Acid (DNA)*. DNA is the basic biological material of inheritance; it determines how proteins are manufactured in the body. DNA is composed of complementary base pairs (Fig. 1.2). There are four distinct bases (A, C, T, G) which compose DNA in pairs. The pairing is obligatory: G and C are always paired, and A and T are always paired.

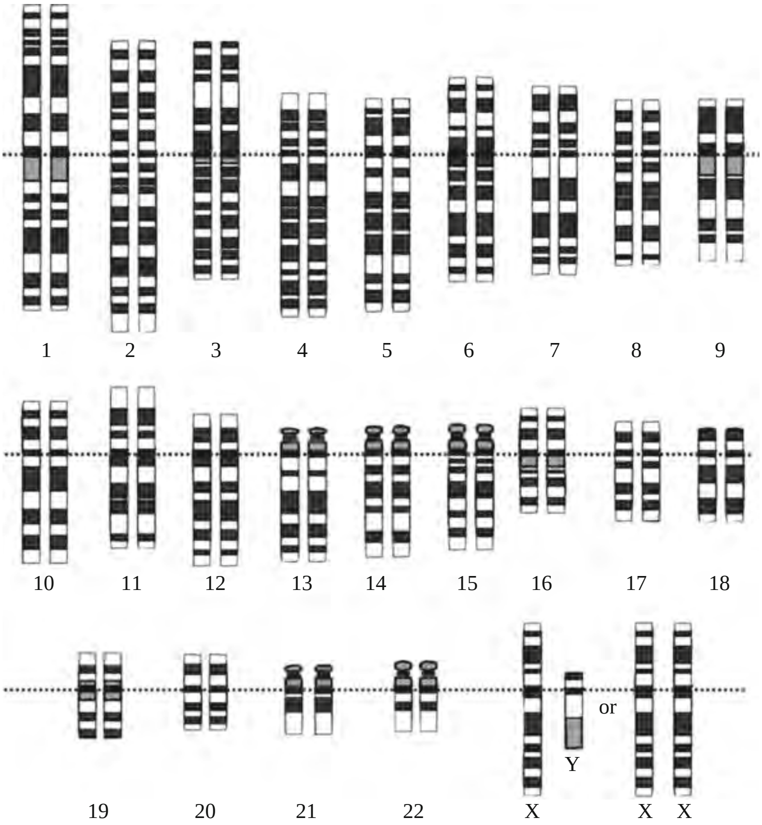


Fig. 1.1 A Graphical representation of the human genome. *Source*: National center for biotechnology information

Genes are largely contiguous stretches of DNA that are responsible for making proteins; the beginning and end of a gene are signaled by specific, short DNA sequences. Current estimates suggest that there are about 20,000–30,000 genes distributed throughout the genome. This estimate has varied wildly over the years, mostly getting much smaller; older books quote 80,000 or even 100,000 genes. The DNA sequence in a gene consists of coding and non-coding regions, or *exons* and *introns*. DNA sequences that make up the exons code for specific proteins determined by the DNA sequence; DNA sequences lying in introns, or sequences lying outside of genes, do not code for proteins, but are thought to play other important roles in regulating the manufacture of proteins. Most of the three billion base pairs in humans are in non-coding regions. Genes vary widely in size, some being as small as a few thousand base pairs, and some containing millions of base pairs. The number and size of both exons and introns also varies between genes. For instance, the APOE gene has four exons and three introns in 3611 base pairs and the gene coding for the human ABO blood type has seven exons and six introns in 5171 base pairs.

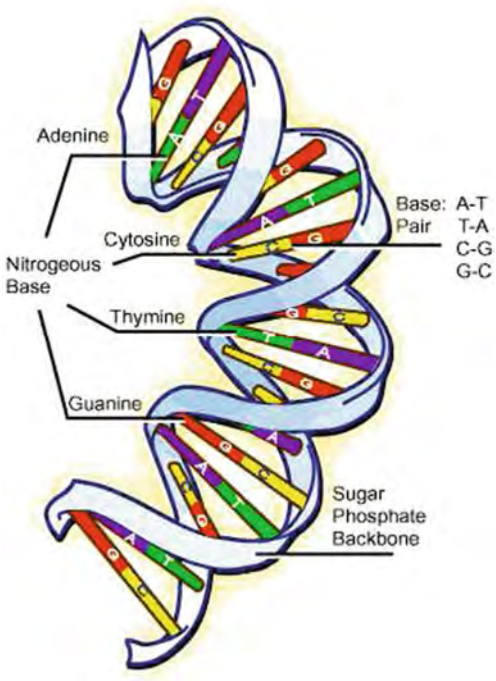


Fig. 1.2 A strand of DNA showing complementary base pairing. *Source*: Courtesy of Jane Wang

The focus of gene mapping has historically been to find the location of one or more of the protein coding regions which have variants affecting disease. However, increased appreciation of the role that non-coding DNA plays in gene regulation and expression, and the many recent association studies that implicate non-coding regions as associated with disease suggest that complex disorders may be influenced by genetic variants in non-coding regions as well.

A *genetic locus* refers to a particular location in a chromosome that is *polymorphic*. Polymorphic means that the data at that locus can have more than one possible variant; a *polymorphism* refers to a polymorphic genetic locus. The different variants at a locus are called *alleles*. Historically the minor allele frequency at a locus was required to have population frequency of at least 1% in order for a locus to be considered polymorphic, but more recently the term is used loosely to indicate any locus where two or more variants are found, regardless of frequency. When there are only two possible variants, it is conventional to refer to them as ‘A’ and ‘a’. When dealing with the autosomes, an individual with two copies of A, one on each of the two chromosomes, is called *homozygous* A, or AA; an individual with one A and one a is called *heterozygous*, or Aa, and aa is homozygous a. The terms *homozygote* and *heterozygote* are also used to denote individuals who are homozygous A or a, or who are heterozygous. The *genotype* of an individual refers to the pair of alleles at a location, i.e., AA, Aa, or aa.

A Comment on Notation: Since Mendel’s paper, it has been conventional to use the capital and lower case forms of a letter to describe the 2 versions of an allele, but this has limitations for more than two alleles. Other conventions which are used include capital letters, A, B, C, etc., or numbers, 1, 2, 3, etc., for different alleles at a single locus. Generally, with a number or letter designation, the choice of label is arbitrary. In some cases a specific designation is given to a disease allele, e.g. S in the Hemoglobin gene which causes sickle cell anemia. With SNPs (see below) we sometimes use base pairs, e.g., G or A to describe alleles. We do not adopt a single labeling convention in this book, but will use the most convenient notation in a given discussion. In most cases, a particular convention will be obvious.

1.3 Types of Genetic Variants

A key feature in the success of gene mapping is having information on genetic variation in humans. Genetic variation means that different copies of homologous chromosomes can have different DNA sequences in specific regions; this definition covers a multitude of possibilities for how DNA sequences differ. In the living cell, DNA undergoes frequent chemical change, especially when it is being replicated. Most of these changes are quickly repaired; those that are not repaired result in *mutations*. All new genetic variation in humans arises as a result of these mutations. Variation in DNA sequence from person to person also arises as a result of the process of reproduction; this will be discussed in Section 2.3. Mutations can arise *de novo* during the process of meiosis, meaning they are present in the offspring but not in either parent, or they can be inherited from parents. Generally we reserve the term mutation for the de novo occurrence; subsequently we refer to it as a variant, or a disease variant if it causes increased disease risk. Gene mapping is concerned with finding inherited disease variants.

We use the term *genetic marker*, or just *marker*, to describe genetic data, observed at the molecular level, at a particular locus that allows us to distinguish genetic differences in individuals. Variants that arise in the coding region of a gene can cause the protein encoded by that gene to malfunction and cells that rely on this protein cannot function properly. This can cause problems for the tissues or organs. Such conditions related to gene mutations, or variants, are called *genetic disorders* or diseases. Genetic variants which cause a genetic disorder are often referred to as disease mutations. A *disease susceptibility locus (DSL)* indicates a gene, or specific genetic locus, which has a variant associated with a disease. This nomenclature has arisen as a convenient way to distinguish the underlying disease gene (usually unknown) which one is searching for, from the marker data used in the search. The term mutation is often used to refer to the event which creates a new variant at the genetic locus and not to the variant itself; we will subsequently adopt that convention here.

Single Nucleotide Polymorphisms (SNP). The simplest type of genetic marker is a *single nucleotide polymorphism (SNP)*. The double helix structure of DNA requires that each chromosome has complementary base pairs at each location, as illustrated for each of the two chromosomes in Fig. 1.3, which shows a SNP at a pair of non-identical but homologous chromosomes. For simplicity, one chromosomal variant

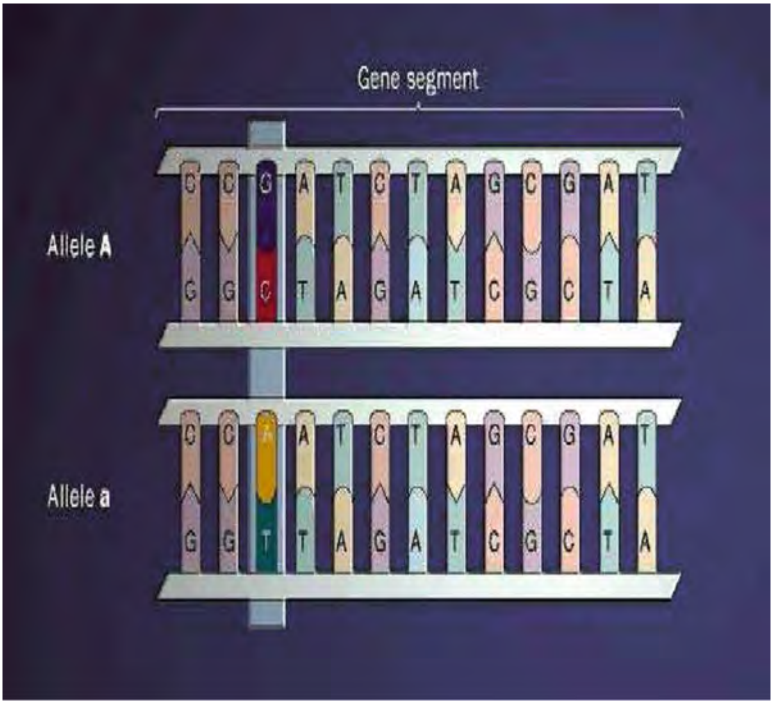

Gene segment
Allele A
C C G A T C T A G C G A T
G G C T A G A T C G C T A
Allele a
C C A A T C T A G C G A T
G G T T A G A T C G C T A

Fig. 1.3 Illustration of a single nucleotide polymorphism (SNP) on a pair autosomal chromosomes. The third base pair of each chromosome shows variation; it can either be G-C or A-T. The labels A and a are used to denote the two variants, or alleles. *Source*: Courtesy of Professor Lyle Palmer

is labeled ‘A’, and one is labeled ‘a’. The A allele should not be confused with the A base in the DNA sequence; rather it is standard notation for an allele and denotes nothing about the underlying biology. In distinguishing the A allele from the a allele, there is a lot of redundant information; a single chromosome is made up of one strand of base pairs. However, for each base in a pair, the other base in the pair is determined by complementarity; thus it is necessary to ‘read’ only one base. In order to unambiguously read a sequence of base pairs, we define a 5’ and a 3’ end according to the asymmetrical bonding of sugar and phosphate residues that form the backbone structure of DNA. By convention, a chromosomal sequence is read, left to right, from the 5’ strand, which is depicted as the top strand in both chromosomes in Fig. 1.3. Thus the sequence for allele A at this location is CCGATCTAGCGAT and the corresponding sequence for a is CCAATCTAGCGAT; they differ only at the third base pair. The two alleles depicted in the figure differ at the third base pair, where an A base is substituted for a G. As we discuss in the next section, whether or not this difference is biologically meaningful depends on where they occur in the DNA sequence and the nature of the letter change.

SNPs play a very important role in modern gene mapping; they occur commonly throughout the genome and the financial cost of genotyping multiple SNPs at different locations is relatively modest, making them very attractive markers for large scale genetic studies. SNPs occur once in every 300 base pairs on average, for approximately 10 million SNPs in the human genome. Most commonly, SNPs are found in intronic or non-coding DNA sequences. SNPs which occur in these non-coding regions have not thus far been shown to have direct genetic effects on disease or traits, but within a coding region, they can be disastrous, as we discuss in the next section.

Indels. Extra base pairs (between 1 and 1000 in number) can be inserted or removed (deleted) in between two specific base pairs in a DNA sequence. Collectively, such variants are called indels. They differ from SNPs in that a SNP is merely a substitution and does not change the number of base pairs in the DNA sequence.

Variable Number of Tandem Repeats (VNTRs). A common type of variation in DNA consists of specific DNA sequences that are repeated immediately adjacent to each other a variable number of times. See Fig. 1.4. Microsattelites are an important class of VNTRs which have a small (1–6) number of base pairs which are repeated. When exactly two nucleotides are repeated, it is called a 'dinucleotide repeat'; when three are repeated, it is called a 'trinucleotide repeat'. Because the number of repeat base pair sequences can vary widely from one person to the next, microsattelites are excellent markers for distinguishing one person from the next. As such, they are widely used in forensic DNA and paternity testing. They also have been used as the basis of most linkage mapping.

Structural Variants. Structural variants include many types of chromosomal changes, including rearrangements, duplications, translocations, inversions, deletion or insertions of genetic material. Many structural variations in chromosomes involving very large segments can be seen under the microscope. These typically arise de novo during the formation of egg and sperm cells and often give rise to substantial disease burden. As these are largely not inherited, they will not be covered in this book. Duplications and deletions involving large segments of DNA that can contain many different genes are usually defined as Copy Number Variants (CNVs) because in such cases, individuals appear to have too many (more than 2) or too

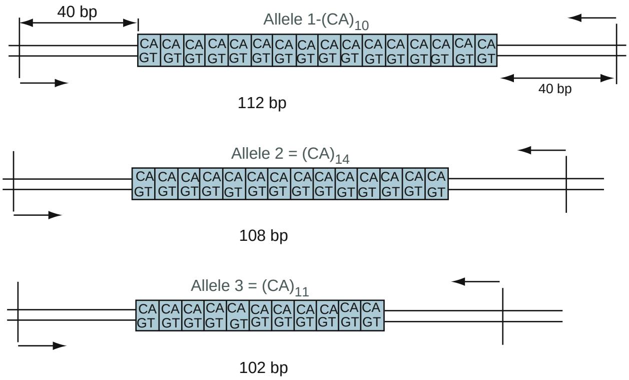


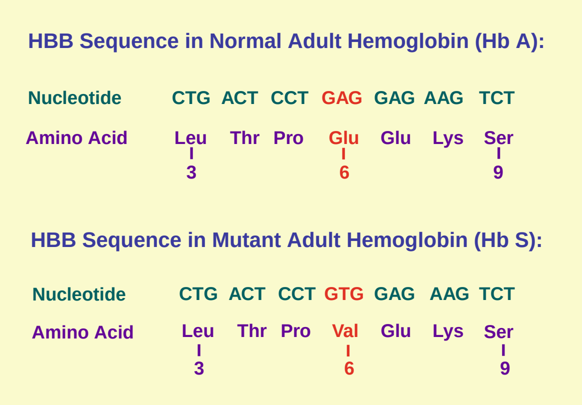
Fig. 1.4 Example of three different alleles at a short tandem repeat of CA

few (0 or 1) copies of the gene or chromosomal segment. The origin of most CNVs present in individuals (de novo versus inherited) is generally unknown. Methods for discovering and detecting CNVs in individuals are now being developed.

1.4 Effects of Genetic Variants on Disease

Many genetic variants have no known effects on disease or disorders in humans, but all types of variants can interfere with normal biological functioning and cause diseases of varying levels of severity. SNPs occurring outside a coding region are thought not to play a role in disease, and even SNPs occurring in a coding region may not have any biological effects because of some flexibility built into coding sequences. As depicted in Fig. 1.5, amino acids are encoded by *codons*, which are three base pair sequences. Most codons have a many-to-one relationship with an amino acid, that is, several three base pair sequences can code for the same amino acid. For example, if the third base in the TCT codon for serine is changed to any one of the other three bases, e.g., TCA, serine will still be encoded. Such variants are said to be *silent* or *synonymous* because they cause no change in their product, but they can still be useful as genetic markers.

Sickle cell anemia is caused by a single base pair change in the hemoglobin gene on chromosome 11. Figure 1.5 shows the coding sequence of the normal hemoglobin gene (A) and the sickle hemoglobin gene (S). The two sequences differ by a change of an A base in the normal Hemoglobin sequence which codes for glutamine, to a T base. The sickle allele (called S for Sickle) changes the sequence coding so that it codes for valine instead of glutamine. This is an example of a *missense mutation* as it changes the DNA sequence to code for a different amino acid. Individuals with SS genotype develop sickle cell anemia; AS and AA individuals are not affected by the disease. AS (and SS) individuals have a better resistance to malaria because some of their hemoglobin is type S. SNPs can also cause *nonsense*



HBB Sequence in Normal Adult Hemoglobin (Hb A):

Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu 3	Thr	Pro	Glu 6	Glu	Lys	Ser 9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu 3	Thr	Pro	Val 6	Glu	Lys	Ser 9

Fig. 1.5 A variant in the hemoglobin gene causing sickle cell anemia

mutations, for example, by changing a coding sequence for an amino acid into a stop sequence which can result in too little protein being produced.

Tandem repeats can also cause diseases and disorders. The non-coding region on the human X chromosome contains a locus where the triplet CGG is repeated (CGGCGGCGGCGG, etc.) in individuals from 5 to 100 times without causing a harmful phenotype. These longer repeats tend to grow longer still from one generation to the next (to as many as 4000 repeats). This causes a constriction in the X chromosome, which makes it quite fragile, and leads to Fragile X Syndrome. Males who inherit such a chromosome show a number of harmful effects including mental retardation. Females who inherit a single fragile X chromosome are only mildly affected.

Huntington’s is another disease characterized by excessive short sequence repeats in the coding region of the Huntingtin gene. The Huntington gene (*HTT*) is located on the short arm of chromosome 4. It contains a sequence of three DNA bases, CAG, repeated multiple times (i.e., ...CAGCAGCAG...) on its 5’ end. If the number of repeats of CAG is less than 27, normal protein is produced, but with more that 36 repeats, a form of protein is produced that increases the rate of neuron decay in the brain and elsewhere, causing the onset of Huntington’s disease symptoms.

Deletion of a sequence of DNA which interrupts a coding sequence can also have effects on disorders. For example a deletion of 32 base pairs in the cytokine receptor-5 (CCR5) gene disables receptors on the surface of cells and disrupts the ability of the HIV-1 virus to infect the cell. In this case, the deletion is beneficial to humans exposed to the HIV-1 virus. Structural variants involving large segments of DNA can cause substantial disease burden. For example, Down’s syndrome arises as a result of errors in meiosis, causing an extra copy of chromosome 21. Although characterized as a genetic disorder, it is not heritable. However, many structural variants can be inherited, and can cause increased disease burden. At the present time, some associations have been found between various complex disorders and CNVs, although none can be considered definitive as yet. Many CNVs appear to be benign, and it is not clear to what extent CNVs are heritable, or largely de novo.

Chapter 2
Principles of Inheritance: Mendel’s Laws and Genetic Models

It is difficult to overstate the impact of Mendel’s research on the history of genetics; indeed, his research in genetics has been credited as one of the great experimental advances in biology (Fisher, 1965). Prior to the publication of his results on experimental hybridization in plants, the concept of inheritance of physical ‘units’ (later called genes) was accepted, and scientists had reported on many hybridization experiments in both animals and plants. Yet no one had set forth principles of inheritance which could be used as a universal theory to explain how traits in offspring can be predicted from traits in the parents. Mendel provided an explicit rule for how the genotypes of the offspring can be predicted from the genotypes of their parents, and he also established models for how genotypes were related to traits. This is nothing short of astonishing in view of the fact that genes and genotypes were not observed; rather their existence was inferred from the phenotypes that were observed. Needless to say, the underlying biology of cell division and the process of formation of sperm and egg cells was not then known; otherwise the derivation of Mendel’s laws would be more straightforward.

Part of Mendel’s success was due to his implicit introduction of the concept of a *genetic model*. A genetic model specifies a probability distribution for the trait, conditional on the underlying genotype at the hypothesized disease locus. Mendel’s genetic models were very simple forms for dichotomous traits that lead to deterministic outcomes. Genetic models underlie most analyses used in statistical genetics. In order to formalize the process of localizing disease mutations and measuring their effect sizes, we often translate the problem to the framework of statistical hypothesis testing and estimation of parameters in the genetic model.

2.1 Mendel’s Experiments

Mendel’s work is known largely through a single research paper, ‘Experiments in Plant Hybridization’ published in 1865. It reported on eight years of experimentation with the garden pea. Mendel made several deliberate choices for his experiments which were crucial in enabling one to infer the laws of inheritance in his series of experiments, essentially examining very simple, now called Mendelian,

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_2,

forms of inheritance. In describing Mendel’s experiments we use the terms gene and genotype to refer to the genetic locus underlying the traits, although the word gene came into use only after Mendel; following Mendel, we refer to the two alleles of a gene as A and a.

Mendel laid out several principles of good experimentation: using large enough samples of crosses, avoiding unintended cross fertilization, choosing hybrids with no reduction in fertility, etc. Here we focus only on those features of Mendel’s experiments bearing on genetics. First is the importance of choosing simple, dichotomous traits for study which are easily recognizable and reproducible. (Mendel studied seven different dichotomous traits.) He called these ‘constant differentiating characteristics’, meaning that two forms of the trait, e.g., green or yellow pods, could be differentiated in plants, and that the same two forms appeared unchanged in offspring. Mendel excluded traits which produced ‘transitional or blended’ results in offspring, or quantitative traits generally. Using dichotomous traits enabled him to use simple genetic models to demonstrate laws of inheritance. It took many decades for scientists to develop models which allowed them to apply Mendel’s laws to continuous traits.

Second was the use of self-pollinating plants which could also be cross-pollinated; both self-and cross-pollination were used in his experiments. See Fig. 2.1. Cross-pollination was used to form the first generation hybrid plants (called

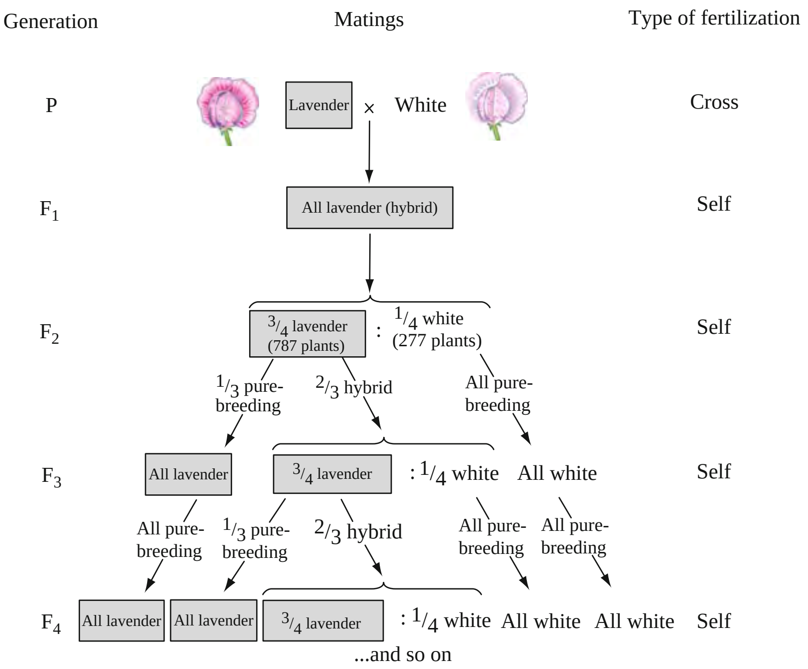


Fig. 2.1 Representation of Mendel’s basic experimental design for the law of segregation. *Source*: Mange and Mange (1999)

F_1 in Fig. 2.1); self-pollination was used to develop the parental pure forms (called P in Fig. 2.1), and to infer the genotypes of subsequent crosses. Mendel started the hybridization with the mating of 'pure' forms (inbred forms of plants which always yielded the same form of the phenotype, e.g., plants always having either yellow pods or green pods); underlying the experiments was the implicit assumption that there were two genetic variants, say A and a, one for each of the two forms of each trait. The use of pure parental forms assured that the experiments always started with the mating of two homozygous parents, either AA or aa, so that the first generation crosses between two pure forms (F_1 hybrids) were always heterozygous Aa.

The result of crossing two different plants showed that only one of the two possible phenotypic forms (purple flowering plants in Fig. 2.1) was observed among the F_1 hybrids. This he termed the dominant form, and the form which disappeared among the first generation hybrids was the recessive. Implicitly, Mendel started with the simple genetic model for homozygotes:

$$P(\text{recessive form of trait}|aa) = 1$$
$$P(\text{recessive form of trait}|AA) = 0$$

$$P(\text{dominant form of trait}|AA) = 1$$
$$P(\text{dominant form of trait}|aa) = 0.$$

Today we usually refer to dominant alleles rather than dominant forms of traits, but the general concept is the same. That is, the A allele is dominant because the Aa genotype has the same phenotype as the AA genotype. Note that these models are deterministic; given a genotype, the form of the trait is determined to be either recessive or dominant with probability 1.

It had already been shown by others that the mating of pure forms led to hybrids with only the dominant form of the trait, but Mendel's contribution was to insist on careful self breeding of successive generations in order to deduce their underlying genotype. He found that the offspring of F_1 hybrids, called F_2, had both recessive and dominant trait forms, in the ratio of 1:3, with the recessive form showing no evidence of contamination by the dominant form. The reappearance of the recessive form allowed him to conclude that the gene for the recessive form was present intact in the F_1 generation, although latent. From the results of the F_1 and F_2 generations we can conclude that

$$P(\text{dominant form}|Aa) = 1$$
$$P(\text{recessive form}|Aa) = 0.$$

Subsequent self fertilization over several generations of F_2 hybrids showed that (1) those plants manifesting the recessive form in the F_2 generation produced only recessive forms among their offspring, and (2) self fertilization of dominant form could be divided into 2 groups: 1/3 produced only dominant offspring as in pure forms, but 2/3 again produced both recessive and dominant forms in the same ratio seen in the F_2 generation of 1:3. These phenotypic ratios are idealized in Fig. 2.1. This led Mendel to deduce the following about the genotypes: $1/4$ of the F_2 hybrids were of the parental recessive form (aa), $1/4 = 3/4 \times 1/3$ were of the parental dominant form (AA), and $1/2 = 3/4 \times 2/3$ were the same as the F_1 generation. From this it follows that the genotypes AA, Aa, aa are in the ratio 1:2:1 in the F_2 generation. This allows us to infer Mendel's first law:

Mendel's First Law (Segregation): One allele of each parent is randomly and independently selected, with probability $\frac{1}{2}$*, for transmission to the offspring; the alleles unite randomly to form the offspring's genotype.*

In summary, the phenotypic ratio for Aa × Aa matings is 3:1 (for dominant to recessive forms) and genotypic ratios are 1:2:1. From Mendel's law of segregation, one can extend the results to a crossing of arbitrary genotypes, as is shown in Table 2.1. The law of segregation underlies the concept of *Mendelian transmissions* of alleles from one generation to the next generation; it is a fundamental and universal concept that forms the basis for many genetic analyses discussed in this book.

Mendel's second law concerns independent inheritance of different traits. We will not examine these experiments in great detail; they are fundamentally not different from the first set of experiments, although more complicated because of the large number of possible outcomes that can be observed when many traits are examined. In addition, as we discuss in the last section of this chapter, not all genes are transmitted independently, so that Mendel's second law is not always true. We now know that genes underlying several of his traits are on the same chromosome and they are not inherited independently. However, Mendel's sample sizes were not sufficiently large to pick up modest departures from independence.

To consider two traits, Mendel considered pure strains for each trait, say AABB and aabb, meaning that one parent always had dominant forms in each trait, and the other parent always had recessive forms for both traits. Experimental crossing gave rise to hybrids with Aa and Bb, which showed only dominant forms for both traits. However, the F_2 hybrids raised from F_1 seed showed four phenotypically different

Table 2.1 Distribution of offspring's genotype conditional upon parental genotypes

Father's genotype	Mother's genotype	Offspring's genotype		
		dd	dD	DD
dd	dd	1	0	0
dd	dD	$\frac{1}{2}$	$\frac{1}{2}$	0
dd	DD	0	1	0
dD	dd	$\frac{1}{2}$	$\frac{1}{2}$	0
dD	dD	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
dD	DD	0	$\frac{1}{2}$	$\frac{1}{2}$
DD	dd	0	1	0
DD	dD	0	$\frac{1}{2}$	$\frac{1}{2}$
DD	DD	0	0	1

plants: those with both dominant forms, plants with one dominant and one recessive form (2 kinds) and plants with two recessive forms, in the approximate ratio of 9:3:3:1 (see exercise 2 of Section 2.4). Subsequent self-pollination of the F_2 generation allowed him to deduce 9 genetic forms among the F_2 hybrids: AABB, AABb, AAbb, AaBB, AaBb, Aabb aaBB, aaBb and aabb in the ratio 1:2:1:2:4:2:1:2:1. These ratios exactly coincide with what one would expect if inheritance of the two traits is independent, for then, with F_2 hybrids,

$$\begin{aligned} P(AA \text{ and } BB) &= P(AA)P(BB) \\ &= (1/4)^2 = 1/16 = 1/(1+2+1+2+4+2+1+2+1), \\ P(AA \text{ and } Bb) &= P(AA)P(Bb) = (1/4)(1/2) = 1/8 = 2/16 \text{ etc.,} \end{aligned}$$

when describing the result of a double heterozygote mating.

Mendel's Second Law (Independent Assortment): *The alleles underlying two or more different traits are transmitted to offspring independently of each other; the transmission of each trait separately follows the first law of segregation*.

Fisher (1936) noted that many of Mendel's statistics were generally too close to their expectations, thus χ^2 statistics comparing observed numbers offspring with a given phenotype to those expected assuming his laws of segregation were true, were often too small, suggesting some data manipulation. This, and the lack of generality of his law of independent assortment (see exercise 3 of Section 2.4), has not diminished the value of his contributions. The lack of independent transmission of different genes is, in fact, fortuitous, as it provides the basis for mapping disease genes by linkage analysis, as will be described in Section 2.3, and in Chapter 11.

2.2 A Framework for Genetic Models

A *genetic model* describes the relationship, usually probabilistic, between an individual's genotype and their phenotype or trait. In Genetic Epidemiology, phenotypes will typically be affection status and we distinguish only between *affected* and *unaffected* subjects in the statistical analysis. Such binary traits can be coded by Y, where $Y = 1$ denotes affected and $Y = 0$ denotes unaffected. For other dichotomous traits such as those that Mendel used, this labeling is arbitrary. For complex diseases, e.g., Asthma, Chronic obstructive pulmonary disease (COPD), Obesity, etc., affection status is often defined by a set of *intermediate phenotypes* or *endophenotypes* which are quantitative measurements that can be more reproducible assessments of the disease features. They can also provide additional insight into the nature and severity of the disease. Standard intermediate phenotypes are body mass index (BMI) as an assessment of obesity, forced expiratory volume in one second (FEV1) for asthma, etc. In some cases, e.g., Alzheimer's disease, the phenotype affection status can be refined by selecting age-of-onset as the target phenotype in the statistical analysis. In general, the selection of the target phenotype is a key question in the planning of the study and the statistical analysis. The phenotype choice will depend on the disease, the possible study designs, statistical power considerations and the necessary adjustments for confounding factors. We will use the variable Y as the variable that describes the phenotype or trait of interest, whether dichotomous or measured.

An individual's genotype at a marker is given by the combination of their two alleles at that locus; we use the notation G to denote an individual's genotype. In the majority of scenarios that we will consider, the marker locus will have only two distinct alleles, e.g., alleles 'A' and 'a'. In the literature such genetic loci are called di-allelic or bi-allelic. Typically, the "small-letter" allele 'a' is assumed to be the more frequent allele of the two and is referred to as the wild type or normal allele. The less frequent allele is labeled with the capital-letter 'A' and referred to as the minor allele. This differs from Mendel's designation of the capital allele as representing the allele associated with the dominant form, because most of the genetic loci we study do not have any known associated dominant or recessive phenotypes, hence today the capital letter usually refers to the less common allele. Under the assumption that the genetic locus is bi-allelic, each of the two chromosomes has to carry either an 'a' or 'A' allele, and, consequently, only three different genotypes are possible: the two homozygous genotypes, AA and aa, and the heterozygous genotype Aa. Order does not matter, so Aa is the same as aA. Thus G can take on only three values in a di-allelic system. With three alleles, there are 6 possible genotypes, etc. Genotypes are inherently categorical but can always be recoded in the form of numerical or indicator variables, as we will discuss at the end of this section.

If the genetic locus is a Disease Susceptibility Locus (DSL), it is conventional to use the D/d designation, as opposed to A/a or B/b; the D-allele is then sometimes referred to as the *Disease Variant* or *Disease Susceptibility Allele*. In formulating genetic models for disease outcomes, we assume the DSL has a direct effect on the phenotype through some biological mechanism. Genetic models can either be deterministic, i.e., the genotype determines the phenotype exactly (*Mendelian Disease*, or, in most cases, probabilistic, i.e., the genotype influences the probability of disease. Conditional upon the individual's genotype G, the probabilistic effect of the locus on the phenotype Y is described by the *penetrance function* which is a set of conditional probabilities, or density functions for continuous phenotypes, which model the distribution of the phenotype/trait, i.e., $P(Y|G)$. If the genetic locus under consideration has no effect on the phenotype of interest, the penetrance probabilities for all three genotypes will be equal regardless of the individual's genotype, i.e., $P(Y|G = dd) = P(Y|G = dD) = P(Y|G = DD)$.

The specification of penetrance probabilities will depend on the type of the disease phenotype. If the phenotype of interest is dichotomous, the penetrance function specifies simple probabilities between zero and one for each genotype, with $P(Y = 1|G) + P(Y = 0|G) = 1$, for each G. When Y denotes disease status, the penetrance probability for $Y = 1$ defines the probability of disease conditional on the genotype of the individual. Mendel considered only two simple genetic models for dichotomous traits: recessive and dominant. The dominant model is

$$P(Y = 1|DD) = P(Y = 1|Dd) = 1 \text{ and } P(Y = 1|dd) = 0, \tag{2.1}$$

and the recessive is

$$P(Y = 1|DD) = 1 \text{ and } P(Y = 1|Dd) = P(Y = 1|dd) = 0. \qquad (2.2)$$

Note that here D is the disease allele (the variant), and $Y = 1$ refers to disease, so that the two models are different. If disease is recessive, it requires two variants, but a dominant disease requires only one. However, if the dominant model holds for the disease outcome, then the recessive model holds for the non-disease outcome, $Y = 0$. This is why Mendel used the terms dominant and recessive to describe possible trait outcomes.

Apart from rare genetic disorders, deterministic models are not very reasonable. Variations of these basic models are constructed by considering stochastic versions which lead to *reduced penetrance* and *phenocopies*. A model is said to be of reduced penetrance if the probability of disease, $P(Y = 1|G)$, is less than 1 for values of G where it is one in the Mendelian models. That is, for the recessive model, $P(Y = 1|DD) = a$ for some $0 < a < 1$, and similarly for the dominant model. The Mendelian models are called *fully penetrant* in contrast to reduced penetrance models, because the probability of disease is either zero or one. The idea behind phenocopies is that the disease could also be caused by another genetic locus, or possibly some non-genetic variable, so that $P(Y = 1|G)$ is positive for those values of G where it is zero in 2.1–2.2. For the dominant mode, for example, $P(Y|dd) = b$ for some $0 < b < 1$. In other cases, the heterozygotes might be intermediate in disease risk between the two homozygotes, suggesting that the number of mutations influences disease risk. Figure 2.2 shows a possible choice for such a penetrance function which allows for both phenocopies and reduced penetrance. Probands with the genotype dd have a 10% chance of being affected. For probands with the genotype DD, the probability of being affected is 7 times higher.

One of the earliest non-Mendelian genes found was APOE for AD. Here there are two mutations giving rise to 3 major alleles (other alleles in the gene are very rare): $E2$, $E3$ and $E4$. The risk of late onset AD increases with an increasing number of $E4$ alleles, but having an $E2$ allele appears protective. Generally $P(Y = 1|G)$ is

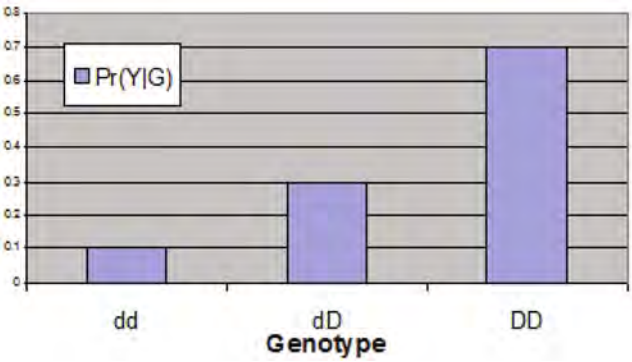


Fig. 2.2 Penetrance function for a dichotomous trait

a complex function of G, but never reaches 1 or 0 for any genotype at the APOE locus.

One publication from the popular press (Pamela McDonald, *The APOE Gene Diet: A Breakthrough in Changing, Cholesterol, Weight, Heart and Alzheimer's Using the Body's Own Genes*) lists the risk for AD as a function of selected APOE genotypes: 20% for 33, 50% for 24, 60% for 34 and 92% for 44. In reality, penetrance functions for AD as a function of APOE genotype are difficult to quantify because they also depend on sex and age. With six possible genotypes, large prospective samples will be required to quantify risk as a function of age and sex with much precision.

For quantitative traits, a natural choice for the penetrance function is a normal density, with a mean that depends upon the genotype while the variance does not. Thus we assume the density function of Y is given by $f(y|\mu_G, \sigma^2)$, where $f(y|\mu_G, \sigma^2)$ denotes the normal density with mean μ_G and variance σ^2; μ_G indicates that the mean depends on the genotype G. For other types of traits, e.g., age-of-onset, the penetrance probability can be selected to be trait-type specific density functions as are used in standard statistical models to describe the relationship between traits and a covariate. Figures 2.3 and 2.4 show examples of penetrance functions for a quantitative trait and for age-at-onset. Again, the notion that the D-allele is the risk allele is echoed in both figures, where we assume larger values of the quantitative trait are deleterious. In Fig. 2.3, the number of D-alleles is correlated with an increased likelihood for larger phenotypic values of Y. Figure 2.4 shows empirical survival curves for AD as a function of APOE genotype, estimated from a large study of individuals free of AD at age 60. Even with this large study, genotype groups have been combined because of sparse numbers at older ages and the low number of subjects with the 4/4 genotype.

Apart from recessive and dominant models for dichotomous traits, thus far we have specified only general probability models which allow the distribution of Y to depend upon G in some unspecified way. The term *Mode of Inheritance* refers to exactly how parameters of the distribution of Y depend on the number of disease alleles. Sometimes the term genetic model is used to describe only the mode of inheritance, and not the entire distribution, but we use genetic model to refer to the penetrance function specifying the entire distribution, and we generally use the mode of inheritance to indicate how the parameters of the penetrance function

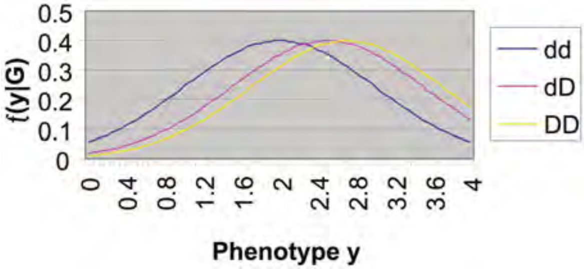


Fig. 2.3 Penetrance functions for a continuous trait

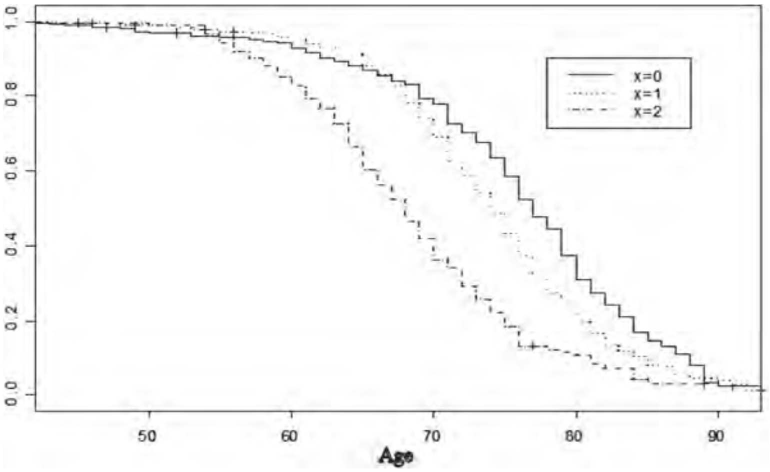


Fig. 2.4 Empirical survival curves for AD as a function of APOE genotype in the NIMH Genetics Initiative Alzheimer's Disease (AD) Sample. The genotype variable x counts the number of $\varepsilon 4$-alleles at the locus

depend on the number of disease alleles. There are four modes of inheritance that are commonly used: *recessive*, *dominant*, *additive* and *codominant*. When only one copy of the disease allele is required to induce an effect on the disease phenotype, $\Pr(Y = 1|dD) = \Pr(Y = 1|DD)$, the mode of inheritance is called *dominant*. However, if 2 copies of the disease allele are required to elevate the disease risk, we speak of a *recessive model* or *recessive mode of inheritance*. Depending on the 'scale', with an *additive mode of inheritance* the penetrance probability of heterozygous genotype is mid-way between the penetrance probabilities of both homozygous genotypes, e.g., $P(Y = 1|Dd) = 0.5 * (P(Y = 1|DD) + P(Y = 1|dd))$ on the linear scale, or $P(Y = 1|Dd) = \sqrt{P(Y = 1|DD) * P(Y = 1|dd)}$ on the log (multiplicative) scale. The *codominant mode of inheritance* makes no assumptions about the relationship among the three penetrance functions, only that they are different. The *heterozygote advantage* model specifies that heterozygotes have the lowest (or highest for a heterozygote disadvantage model) risk of disease; it is occasionally used, especially in plant breeding. We do not use it since it is a special case of the more general codominant model.

Note that with dichotomous traits, $P(Y = 1|G)$ can be equivalently expressed as $E(Y|G)$, and likewise for the continuous trait, $\mu_G = E(Y|G)$. *Generalized Linear Models (GLM)* provide a convenient way to express the dependence of the trait mean on G without specifying the entire distribution of Y. A generalized linear model is similar to an ordinary linear regression model, except it allows the mean of Y to depend on covariates, X, in a non-linear way as:

$$g(E(Y|X)) = \beta_0 + X'\beta_1. \tag{2.3}$$

The link function, $g(\cdot)$, depends on the type of trait. For affection status outcomes, the logistic link:

$$\log[E(Y|X)/(1 - E(Y|X))] = \beta_0 + X'\beta_1, \tag{2.4}$$

or log(relative risk) link:

$$\log[E(Y|X)] = \beta_0 + X'\beta_1, \tag{2.5}$$

models are commonly used in epidemiological work; in genetics, linear models in the probabilities themselves are also commonly used.

Here X is a coding of the genotype that reflects the mode of inheritance; it can be a vector or a scalar, depending on the genetic model. By proper choice of X and link function $g(\cdot)$, all four modes of inheritance can be expressed by equation (2.3); β_0 is an intercept parameter, specifying $E(Y|G)$ when $X = 0$; β_1 gives the additional model parameters which specify how $E(Y|G)$ depends on G. Often the right-hand side of equation (2.3) is written as $X'\beta$ where β is a vector incorporating β_0 and β_1, and X is a vector with the first element always one; here we keep the parameters separate since a test of whether or not the gene affects the trait uses $H_0 : \beta_1 = 0$. Acceptance implies no relation between the gene and the trait. The coding of the genotype for each mode of inheritance is given in Table 2.2. From Table 2.2, we see that β_0 always specifies $E(Y|dd)$ and for the recessive model, it specifies $E(Y|Dd)$ as well. For the recessive, dominant and additive models, β_1 is a scalar and defines the 'effect size' in the chosen scale; for the codominant model, β_1 is a vector of length two that gives the effect of the DD and Dd genotypes compared to dd. Although more complex models can be constructed, these simple generalized linear models suffice for most analyses that we consider in detail.

Table 2.2 Coding the genotype (G) as X to specify the mode of inheritance

Recessive		Dominant	
X	G	X	G
1	DD	1	DD or Dd
0	dd or Dd	0	dd

Additive		Codominant		
X	G	$X1$	$X2$	G
2	DD	1	0	DD
1	Dd	0	1	Dd
0	dd	0	0	dd

2.3 The Biology Underlying Mendelian Inheritance

Today Mendel's Laws can be derived directly from our understanding of *Meiotic cell division* or *Meiosis*, which is the cell division that produces **gametes**, either sperm or ova; the union of a sperm and ova produces the fertilized egg cells (called *zygotes*). Meiotic cell division is in contrast to the standard cell division, *mitosis*, that serves the purpose of cell growth, development, repair and replacement of worn-out cells. While mitosis results in cells that are genetically identical (or clones), the purpose of meiosis is to introduce further genetic diversity by creating gametes, either egg cells or sperm cells, that are genetically different from the parent cells.

The nucleus of every cell contains two copies of each *chromosome* inherited from the parents, one maternal copy and one paternal copy. Such cells are called *diploid* because they have two copies of each chromosome (except for males who have one X and one Y for the sex chromosomes). Meiosis consists of two rounds of cell divisions, each following a meiotic division (Fig. 2.5) ending with four *haploid* cells containing only one copy of each chromosome.

In the beginning of the first meiotic division, both parental copies of the chromosome are duplicated; Fig. 2.5 illustrates the first meiotic division for a single parent in the top panel and the result of the second meiotic division in the bottom. Each parental chromosome is first duplicated as illustrated after the first arrow in the top panel. The duplicated chromosomes are called a pair of *sister chromatids*. The two duplicated chromosomes undergo a separation process; during this process, the arms of the chromosomes may overlap and segments of non-duplicate chromatids can be exchanged between the duplicated chromosomes, as illustrated after the second

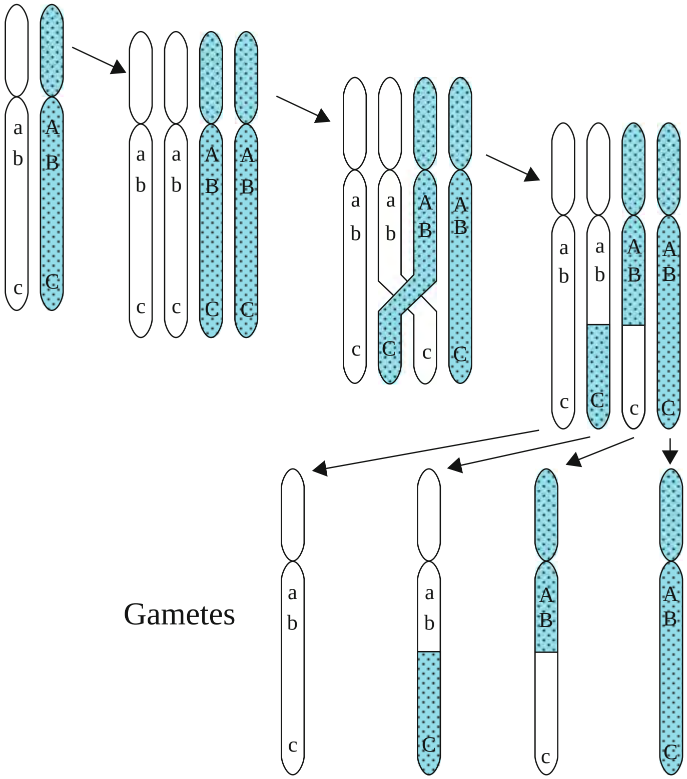


Fig. 2.5 Crossing-over and recombination during the formation of gametes (germ cells) or meiosis

arrow in Fig. 2.5. The exchange of material between two non-sister chromatids is called a *crossover event*. After the third arrow in Fig. 2.5, we see four chromatids. Two are identical to the one seen in the parent, but the other two are a mixture of the two chromosomes in the parent. Notice an important feature of crossing over: it allows each of the four gametes to be a mixture of the genetic material inherited from two grandparents, either maternal or paternal. Thus meiosis is not simply randomly choosing one of two parental chromosomes randomly but rather, it allows for creation of additional genetic diversity by mixing of grand parental information within a single chromosome. Each person inherits approximately $1/4$ of their genetic material from each of their four grandparents.

In the second meiotic division, the chromatids are separated and the final cell division forms two new cells around each chromosome, for a total of four haploid gamete cells. By crossing over, each gamete cell contains a different chromosome, however as a result of the first cell division, at each specific locus there are two gametes with the same maternal allele and two gametes with the paternal allele. A zygote requires one sperm and one ovum (egg cell); assuming that gametes unite randomly to form zygotes, it is then clear that the transmission of each parental allele occurs with probability $1/2$ since the two alleles are represented equally in the gamete cells.

Mendel's law of independent assortment states that alleles at different genetic loci are transmitted independently from one generation to the next. If they are on different chromosomes, this is naturally the case since each pair of chromosomes undergoes the process of meiosis independently. This creates a substantial amount of genetic variation, even without crossing-over; with crossing-over, the possible combinations are essentially infinite.

Crossovers are random events in the sense that they cannot be predicted with certainty; however they do not occur uniformly or independently along the chromosome. Rather, crossover rates can vary by sex, chromosomal region as well as chromosome number, individual and temperature. Crossing over is relatively rare at the centromere and at the ends of a chromosome. *Interference* can create dependencies in the occurrence of successive crossovers. For example, the occurrence of a crossover in a region decreases the chance of a second crossover in an adjacent region, nearly to zero if the regions are very close. Overall the entire genome, the average number of crossovers is about 55 in males, and about 50% greater in females. The average number of crossovers on a chromosome depends upon its length. Thus despite the fact that crossovers do not occur uniformly, they have served as a useful measure of distance for linkage mapping as described in Chapter 5.

Crossovers are inherently unobservable, so we use the concept of recombination to describe crossovers. If we obtain data at two or more loci on a parent and their offspring, then we can infer something about crossovers occurring between the loci provided the parent is heterozygous at the loci. Referring to Fig. 2.5, the parent is heterozygous at three locations, with alleles Aa, Bb and Cc. The set of alleles lying on the same chromosome is called the *haplotype*. Here the two haplotypes are ABC and abc. Note that these haplotypes have been inherited from the two parents of the parent, i.e., the grandparents of the offspring whose gametes are displayed. Suppose that the first gamete, abc, is inherited from the parent. There is no evidence of crossing over here because one parental chromosome is identical abc, and the other parental chromosome shares none of these alleles. In this case we say there is no *recombination* between either the A to B locus, or the B to C locus (or A to C either). Suppose the offspring inherits the second gamete, abC. In this case, the offspring's haplotype differs from either of the parent's haplotypes, thus a crossover must have occurred between the B and C locus, but not the A and B. Thus we say no recombination has occurred between A and B, but a recombination occurred between B and C.

There is not a one-to-one relationship between recombination events and crossing over because recombination refers only to what can be observed between the two specific loci, whereas crossing over refers to events that can occur anywhere in the interval. If no crossover has occurred between two loci (as between the A and B loci in Fig. 2.5) then we will not see a recombination. However, it is possible for two crossovers to occur in an interval; in this case, we may see no recombinant between two markers flanking the interval, i.e., there may be segments of grand-maternal material at the ends of the interval, with grand-paternal material in the middle. The formal definition of the recombination fraction θ is given by P(recombination occurs between two loci).

Crossovers between two loci very close to one another are rare. In this case, the probability of a recombination between the two loci is very small. For example in Fig. 2.5, considering loci A and B, among the four gametes, we observe two ab gametes and two AB gametes: thus among these gametes, the probability of A or a (or B or b) is always $\frac{1}{2}$ by Mendel's law of segregation, but P(A allele and B allele) $=$ P(a allele and b allele) $= \frac{1}{2}$ and P(A allele and b allele) $=$ P(a allele and B allele)$=$ 0. This is contrary to what we would expect by Mendel's law of independent assortment, which would specify a probability of $\frac{1}{4}$ for each of the four possible gametes.

Between loci B and C, the situation is different because we observe a recombination. Again, among the four gametes, P(B) $=$ P(b) and likewise for C and c, but now P(b and c) $=$ P(B and c) $=$ P(b and C) $=$ P(B and C) $= \frac{1}{4}$, which corresponds to independent assortment. In general, the distribution of gametes over many meioses will depend upon the number of crossovers between them. If the two loci are close, θ is small, and the alleles at two loci tend to be inherited together, so that the law of independent assortment does not hold.

The relationship between θ and the distribution of crossovers is given by Mather's law:

$$\theta = (1 - P_0)/2,$$

where P_0 is the probability of zero crossovers. Mather's law can be argued as follows. If there are no crossovers, $P_0 = 1$, and there can be no recombination. With probability $(1 - P_0)$, at least one crossover occurs. If at least one crossover occurs, then the probability of a recombination is $\frac{1}{2}$, regardless of the number of crossovers.

To see why, recall that crossovers cannot occur between sister chromatids, but only between non-sister chromatids. It is easy to see from Fig. 2.5 that one crossover will create two recombinant gametes and two non-recombinant gametes. With two crossovers, the same two non-sister chromatids can be involved in both crossovers (and the number of recombinant gametes is zero) or both sister chromatids of each pair cross over once with their non-sister chromatids, in which case all four gametes are recombinants. Since these two possibilities are equally likely, the average proportion of recombinants is $\frac{1}{2}$. The last possibility, that one sister chromatid crosses over twice with two different non-sister chromatids, gives 2 recombinant and 2 non-recombinant gametes. It is straightforward to argue the probability of a recombinant is also $\frac{1}{2}$ for three crossovers, and so on.

If two loci are very far apart, there are likely many crossovers between them; P_0 approaches one in the limit and the recombination fraction approaches $\frac{1}{2}$. The upper limit of θ corresponds to what we might expect if two loci are on different chromosomes, since by the law of independent assortment, if the parent is heterozygous at both loci, the four gametes will carry the four possible combinations, AB, Ab. aB, and ab with equal probability.

2.4 Exercises

1. Verify lines 1–3 of Table 2.1 using Mendel's first law.
2. Assume two genes with alleles A/a and B/b, controlling two different traits. Assuming that Mendel's second law holds (the alleles underlying the two different traits are inherited independently), and starting with the pure strains as in Mendel's experiments:

   (a) Verify the 1:2:1:2:4:2:1:2:1 ratios for the 9 possible genotypes inferred in the F_2 generation.

   (b) Verify the 9:3:3:1 ratio for 4 possible traits observed in the F_2 generation.

3. In the early 1900s, scientists William Bateson and R. C. Punnett studied inheritance in two genes of Sweet Peas: one affecting flower color (P, purple, and p, red) and the other affecting the shape of pollen grains (L, long, and l, round). Capital letters denote dominant forms, as in Mendel's paper. They crossed pure lines PP · LL (purple, long) × pp · ll (red, round), and self-fertilized the F_1 offspring Pp · Ll heterozygotes to obtain an F2 generation. The table below shows the counts of each phenotype in the F_2 plants.

	Number of progeny	
Phenotype (and genotype)	Observed	Expected from 9:3:3:1 ratio
purple, long (P/– · L/–)	4831	3911
purple, round (P/– · l/l)	390	1303
red, long (p/p · L/–)	393	1303
red, round (p/p · l/l)	1338	435
	6952	6952

(a) Verify the Expected column for testing goodness of fit to the 9:3:3:1 ratio.

(b) Show that the chi-square goodness of fit test exceeds significance.

Note: As a possible explanation for the lack of fit, Bateson and Punnett proposed that the F_1 had actually produced more P × L and p × l gametes than would be produced by Mendelian independent assortment. Because these genotypes were the gametic types in the original pure lines, the researchers thought that physical coupling between the dominant alleles P and L and between the recessive alleles p and l might have prevented their independent assortment in the F1. However, they did not know what the nature of this coupling could be.

(c) What is another possible explanation for lack of fit?

4. How many genotypes are possible with a 3-allele marker? With K alleles?
5. Early onset Alzheimer's disease is very rare; for illustrative purpose, assume it is 0.1% among adults aged 30-60. Rare variants in 3 genes, APP, PSEN1 and PSEN2 have been identified as causing early onset AD in a dominant fashion, with P(AD | any of the three variants) = 1. Early onset AD can also be caused by head injury; many other non-genetic factors have been suggested. In a series of 101 cases of early onset AD, only 7 (or approximately 7%) were found to have these variants in APP, PSEN1 or PSEN2; that is, the attributable risk due to the three rare variants is low. For simplicity, assume that the probability of variants in these 3 genes is so rare that we can assume P(no variant in any gene) $\approx$ 1. Let the disease allele D symbolize a variant in any one of the three genes, d is no variant, and Y $=$ 1 means AD present.

   Estimate the probability of a phenocopy, $P(Y = 1|dd)$ (also known as phenocopy rate) for these genes combined, using the data given and Bayes Rule.
6. Consider a recessive Mendelian disease, where in the population, P(an individual has 2 disease variants) $=$ 0.000001.

   (a) What is the probability that a randomly selected person is affected? Suppose that the randomly selected person is affected. What does that imply about the probability that their sibling is also affected (you can assume that having either one or two parents with two variants is so rare that you can ignore them)?

   (b) Now answer both of these questions assuming the penetrance is only $\frac{1}{2}$, i.e., P(disease | 2 variants) $= \frac{1}{2}$, but the phenocopy rate is still zero.
7. Suppose we are dealing with a quantitative recessive trait, which is distributed as $N(\mu, 1)$ when there are two variants, and $N(0, 1)$ otherwise. Calculate the probability that a randomly selected person with two variants has a trait higher than a person with one or no variants, when $\mu = 0.5$, and when $\mu = 2$.
8. Suppose we observe a quantitative trait which seems to show variation in both the mean and the variance as a function of genotype. Give one example of a genetic model which allows for this.
9. One of the dichotomous traits that Mendel studied, length of plant stem, was actually dichotomized from the measured length. He selected plants with a 6–7'

long axis to have the dominant trait and plants with a $3/4'$ to $1.5'$ long axis to have the recessive trait. Mendel commented that in fact, "…the longer of the two parental stems is usually exceeded by the hybrid… Thus for instance, in repeated experiments, stems of $1'$ and $6'$ in length yielded without exception hybrids which varied in length between $6'$ and $7.5'$." What would be an appropriate (non-deterministic) Gaussian penetrance function model for axis length as a continuous trait? Mendel also noted that there is very little variation in stem height within genotype class. What does that imply about your Gaussian model?

10. Consider the Generalized Linear Model given in equation (2.3) Suppose you wish to include covariates, such as sex or age. Suggest how you might do that in the context of the GLM.
11. Verify the statement concerning two crossovers: If one paternal chromatid crosses over twice with two different maternal chromatids, this gives 2 recombinant and 2 non-recombinant gametes.

Chapter 3
Some Basic Concepts from Population Genetics

The study of allele frequencies and how they vary over time and over geographic regions has led to many discoveries concerning evolutionary history, migration, gene flow, and the correlation between allele frequencies and disease rates across populations. This chapter covers only a few concepts from population genetics, emphasizing those most relevant to gene mapping: allele frequency estimation, population substructure, Hardy-Weinberg Equilibrium (HWE) and Disequilibrium (HWD), which are frequently used in the analysis of genetic data. Other concepts, e.g., Linkage Disequilibrium and Linkage Equilibrium, will be introduced in later chapters as the need arises.

3.1 Estimation of Allele Frequencies

Recall that each person has two copies of each autosomal chromosome, so at any specific locus, each person has two alleles, one inherited from each parent. Consider estimation of the population proportion of a particular allele, A, at a locus; for now, we let all other alleles be denoted by 'a'. The allele proportion in the population is defined as the proportion of chromosomes carrying that allele, regardless of the pairing within individuals. Suppose that we have a sample of size n from a population with a proportion, p, of A alleles. Then to estimate p, we simply count the number of chromosomes carrying the A allele and divide by $2n$, the number of chromosomes. Box 3.1 illustrates this calculation.

Box 3.1 Calculation of Estimated Allele Frequencies from a Sample of n Subjects

Genotype counts from the sample:

n_{AA} = number out of n with genotype AA
n_{Aa} = number out of n with genotype Aa
n_{aa} = number out of n with genotype aa

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_3,

where $n_{AA} + n_{Aa} + n_{aa} = n$. The sample proportion of A alleles,

$$\bar{p} = (2n_{AA} + n_{Aa})/2n, \tag{3.1}$$

estimates the population proportion of A alleles. With a two allele system, the proportion of a alleles is $\bar{q} = 1 - \bar{p}$, as can be verified by exchanging a with A in formula (3.1).

A comment on notation: It is typical in genetics to refer to $\bar{p}$ as the 'A allele frequency', even though it is a proportion, and frequency usually refers to a count.

Note that $\bar{p}$ is an ordinary proportion, but the sample size is $2n$, the number of chromosomes. It is easily seen to be unbiased for the population frequency p provided we have a random sample with equal probability sampling, even if the sample contains relatives. Equal probability sampling requires that everyone in the population has the same probability of being included in the sample. In practice, what we need is that the probability of selection into the sample does not depend upon an individual's genotype or any phenotype related to the genotype. For example to estimate the 3 allele frequencies at the ABO blood group locus, we must genotype sample individuals without regard to their blood group membership (A, B, AB or O). However, the usual standard error for a proportion, $\sqrt{\bar{p}(1-\bar{p})/2n}$, may not hold as this formula assumes independence of the $2n$ sampled chromosomes. We defer discussion of this to Section 3.3 when we take up Hardy-Weinberg Equilibrium. Extension to more than 2 alleles, say A, B, C, etc. is straightforward:

$$\bar{p}_A = (2n_{AA} + n_{AB} + n_{AC} + ...)/2n, \tag{3.2}$$

and similarly for $\bar{p}_B$, $\bar{p}_C$, etc. We leave as an exercise the estimation of allele frequencies for loci on the X chromosome. Estimation of allele frequencies for the MN blood group is illustrated in Box 3.2.

Box 3.2 Example – estimating allele frequencies for the MN Blood Group

An individual's MN blood group is determined by a gene with two alleles, M and N; they control the amount of M and N antigens on the surface of blood cells. The data below come from two different samples of Eskimos in Greenland. We use the data to estimate the M allele frequency.

Location	MM	MN	NN	Total	$\bar{p}$	$\bar{q}$
South West Greenland	126	53	8	187	0.8155	0.1845
East Greenland	475	89	5	569	0.9130	0.0870

For South West Greenland:

$$\bar{p} = (2 * 126 + 53)/(2 * 187) = 0.8165$$
$$\bar{q} = (2 * 8 + 53)/(2 * 187) = 0.1835 = 1 - 0.8165$$

East Greenland

$$\bar{p} = (2 * 475 + 89)/(2 * 569) = 0.9135$$
$$\bar{q} = (2 * 5 + 89)/(2 * 569) = 0.0870$$

Source: Fabricius-Hansen (1939), Ahrengot and Eldon (1952)

Allele Counting is sometimes used to refer to formula (3.1), and also more generally to a method of estimating allele frequencies when data on genotypes are not available directly, but data are available on Mendelian phenotypes, such as ABO blood types.

3.2 Population Substructure

We use the term population substructure loosely to refer to features of a population which result in variation of expected allele frequencies across individuals in a population. The estimate of allele frequency obtained by allele counting (formula (3.1)) will still be an unbiased estimate of the population allele frequency in the presence of population substructure, provided we have equal probability sampling from the target population. However, the presence of population substructure can mean that not all subjects have the same probability of being represented in the sample, depending on how the sample is selected. If genotype frequencies differ over subgroups and the sampling mechanism favors certain subgroups over others, the sample estimate may be biased. Even if there is no bias, population substructure will influence the variance of the estimate and affect the distribution of test statistics that are computed based on allele frequencies. Problems associated with both bias and variability in the test statistics, and methods for handling these problems will be discussed in Chapter 8. Population substructure can also influence the distribution of genotypes in the population. We now provide a brief overview of three common types of population substructure.

3.2.1 Population Stratification

Population stratification is perhaps the simplest form of population substructure, as it coincides with the intuitive notion that individuals in a population can be

Table 3.1 *Population Stratification*. Distribution of albumin types among selected dog breeds and mongrels. *Source*: Adapted from Christensen et al. (1985)

	Genotypes				Frequency
Breed	SS	SF	FF	Total	of S
Basset Hound	0	2	30	32	0.031
Beagle	3	14	52	69	0.145
Dachshund	2	8	26	36	0.167
Collie	2	21	18	41	0.305
Cocker Spaniel	7	24	20	51	0.373
Labrador Retriever	8	10	10	28	0.464
German Shepherd	36	47	23	106	0.561
Terrier, Tibetan	10	11	3	24	0.646
Newfoundland	35	33	3	71	0.725
Poodle	39	36	6	81	0.704
Boxer	54	14	1	69	0.884
Golden Retriever	53	3	1	57	0.956
Basenji	44	0	0	44	1.000
Other pure breeds	94	57	38	189	0.648
Mongrels	22	41	24	87	0.489
Total	399	321	255	975	
Overall Gene Frequency					$p_s = 0.574$
Genotypic Frequencies	0.409	0.329	0.262		

subdivided into mutually exclusive strata; within each strata the allele frequency is the same for all individuals, but it varies between strata. Typically we assume that the different strata represent different racial, ethnic and/or geographic subgroups. Examples of population stratification are readily available from the plant and animal breeding literature. For example, Table 3.1 shows the distribution of the slow allele (S) at the albumin locus stratified by specific dog breed, pure breeds and mongrels.

3.2.2 *Population Admixture*

Population admixture refers to a situation where individuals in a population have a mixture of different genetic ancestries due to the mixing of two or more populations at a previous point in time. Most admixed populations are the result of a migration of one or more population groups from specific regions into a different geographic location with a previously settled population. If the allele frequencies differ in the original ancestral populations, then the probability that an individual has a particular allele depends upon the mixture of that individual's ancestry. Population admixture is a more realistic model for most modern population groups than is stratification. A good example of an admixed population is the Gila Indian River Community, as illustrated in Table 3.2.

Native Americans in the Pima and Papago tribes have different degrees of American Indian and European Hispanic ancestry. Of interest here is the distribution of

Table 3.2 An Admixed Population: Native Americans of the Pima and Papago Tribes

Indian Heritage	Gm3;5;13;14%	% Diabetes*
0	65.8%	18.5%
4	42.1%	28.6%
8	1.6%	39.2%

* Age adjusted
Adapted from Knowler et al. (1988)

the Gm3;5,15,14 allele which lies on a locus of the human immunoglobulin G gene. Table 3.2 shows allele frequencies and also the percentage with diabetes for the adults in the population, stratified by the number of great grandparents with Indian heritage. For those with the highest degree of Indian ancestry, the allele frequency is almost zero, whereas it is almost 70% for those with no great grandparents with Indian Heritage. Such a marker with strong differences among population subtypes is called an *Ancestry Informative Marker (AIM)*.

Note that the percentage of the population with diabetes shows a strong inverse correlation with allele frequency. It is not uncommon to see variation in disease rates across population strata of ancestry; when allele frequencies and disease rates are correlated, as they are here, spurious associations between disease and marker can occur if ancestry is not taken into account. This will be discussed in some detail in Chapter 8.

3.2.3 Population Inbreeding

Population inbreeding occurs when there is a preference for mating among relatives in a population or because geographic isolation of subgroups restricts mating choices. In either case, there is the possibility that an offspring will inherit two copies of the same ancestral allele. The *inbreeding coefficient*, denoted by F, is the probability that a random individual in the population inherits two copies of the same allele from a common ancestor. In large, randomly mating populations the chances that any two mating parents have a common ancestor allele is low, hence F is negligible and often considered to be zero. Inbred populations have non-negligible inbreeding coefficients. At the extreme, self-breeding populations of plants have inbreeding coefficients of one. To see why, consider a self-fertilizing plant with two alleles. All offspring from this self-fertilizing plant have a probability of $1/2$ of inheriting two copies of the same allele. In the next generation, this probability increases to $3/4$, and eventually one of the allele frequencies goes to 1 in these plants. In real populations, it is difficult to estimate F exactly, as other phenomena may mimic the effect of inbreeding. Inbred populations have higher than expected frequencies of rare recessive disorders, because inbreeding tends to increase the number of homozygotes in the population. These issues are discussed in subsequent sections of this chapter.

3.3 Hardy-Weinberg Equilibrium

In this section, we introduce the concept of Hardy-Weinberg Equilibrium (HWE). If the conditions for HWE are met, the genotype distribution is defined by the allele frequency. In many statistical applications, the presence of HWE simplifies the statistical theory and methods substantially and is, consequently, often assumed. We provide here a derivation of the HWE-genotype distribution and discuss tests for HWE as well as the effects of population substructure on HWE. We will consider the Hardy-Weinberg principle for autosomal loci. For extensions to the X chromosome, see Lange (2002).

In 1908, Godfrey Hardy and Wilhelm Weinberg independently derived a formula relating allele frequency in parents to genotype frequency in offspring. There are many assumptions required for the formula to hold: random mating, no inbreeding, infinite population size, discrete generations, equal allele frequencies in males and females, and no mutation, migration, or selection (meaning that certain alleles do not confer a selective advantage or disadvantage in reproduction). Even though none of these assumptions is likely to hold exactly in any population, the Hardy-Weinberg principle often provides a good approximation for population genotype frequencies.

Let p be the frequency of the A allele in a population satisfying the assumptions given above. Then it is easy to show that the genotype frequencies in the offspring after one round of random mating are given by:

$$\left.\begin{array}{rl} P(AA\text{ genotype}) = & p^2 \\ P(Aa\text{ genotype}) = & 2pq \\ P(aa\text{ genotype}) = & q^2 \end{array}\right\} \tag{3.3}$$

A population is said to be in *Hardy-Weinberg Equilibrium (HWE)* if the genotypes in the entire population satisfy (3.3). Since $(2p^2 + 2pq)/2 = p$, the frequency of A allele among the offspring chromosomes is also p. Thus, with HWE, allele frequencies will not change from generation to generation.

The proof of the HWE formula uses straightforward algebra and Mendel's laws. The simplest proof uses the distribution of alleles in gametes. Recall that gametes are sex cells that have only one of each autosomal chromosome; in the formation of gametes, each of the two parental alleles is equally likely to appear in a gamete, hence the allele frequency among gametes is the same as the allele frequency among chromosomes. The various population assumptions made for HWE imply that the formation of a zygote (a fertilized egg cell) is equivalent to the random union of two gametes, one from mother and one from father. Thus with random mating, the probabilities of the number of A alleles in the offspring generation are given by the binomial formula with probability equal to the A allele frequency and the number of trials equal to 2. As a consequence, the number of A alleles in an offspring is distributed as $B(2, p)$. Further, the number of A alleles in a random sample of size n from the population is $B(2n, p)$. Thus an important consequence of HWE is that the formula for var$(\bar{p})$ from a sample of size n is given by the simple binomial formula, $\bar{p}\bar{q}/(2n)$. Box 3.3 summarizes the basis for inference about $\bar{p}$ when HWE holds.

Box 3.3 Inference About allele frequencies in a sample from a population in Hardy-Weinberg equilibrium

Let i index the individuals in a random sample of n independent individuals from a population with allele frequency p; let X_i $(i = 1, \ldots, n)$ denote the number of A alleles for the i^{th} person in the sample and let X_+ denote the summation of X_i over all n individuals. Then we may rewrite $\bar{p}$ as

$$\bar{p} = \sum_{i=1}^{n} X_i/2n = X_+/2n.$$

Since each X_i is distributed as $B(2, p)$, $E(X_i) = 2p$, $\text{var}(X_i) = 2pq$ and with a sample of independent individuals, X_+ is $B(2n, p)$. It follows that:

$$E(\bar{p}) = E(X_+)/2n = p,$$

and

$$\text{var}(\bar{p}) = \text{var}(X_+)/(2n)^2 = pq/2n.$$

In large samples, $\bar{p}$ is approximately $N(p, \bar{p}\bar{q}/2n)$; large sample tests and confidence intervals use this normal approximation. In particular, with large samples, to test $H_0 : p = p_0$ at the α-level, we reject if the magnitude of

$$Z = \sqrt{2n}(\bar{p} - p_0)/\sqrt{p_0(1 - p_0)}$$

is greater than the $(1 - \alpha)/2$ - percentile $(Z_{(1-\alpha)/2})$ of a standard normal distribution. An approximate $100(1 - \alpha)\%$ confidence interval for the true frequency is given by

$$\bar{p} \pm (Z_{(1-\alpha)/2}))\sqrt{\bar{p}(1 - \bar{p})/2n}.$$

The approximations are reasonably good for $n\bar{p} \geq 5$ and $n(1 - \bar{p}) \geq 5$ for levels of α close to 0.05. With smaller samples and smaller levels of α, exact inference for p is based on the fact that X_+ is $B(2n, p)$. See Rosner (1994) (Section 7.11) for example, for information on exact Binomial inference.

Note that there is no assumption made about the genotype distribution in the parental population, only that the allele frequency is p. The parental population need not follow HWE, that is, the genotypes in the parental population do not follow equation (3.3). However, HWE can be achieved in the offspring in only one generation of random mating, provided the other conditions hold. HWE can also be proved by explicitly considering an arbitrary genotype distribution in the parents, still with allele frequency p, and showing that formula (3.3) holds for the offspring distribution.

3.3.1 Testing for HWE

Tests of HWE are useful in a variety of settings; the basic idea is to compare the observed genotypes in a sample with those which are expected if HWE holds. We estimate the allele frequency from the observed counts using formula (3.1) then the expected genotype counts using equation (3.3) where $\bar{p}$ is substituted for p. With large samples, we can compute the standard Pearson χ^2 goodness-of-fit test with 1 degree of freedom. Box 3.4 illustrates the calculation of the HWE test.

Box 3.4 The Pearson goodness of fit test for HWE

H_0 : HWE holds in the population.
H_A : HWE does not hold.
Given a sample of size n from the population:

	Genotype			
	AA	Aa	aa	
Observed	n_{AA}	n_{Aa}	n_{aa}	n
expected	$n\,\overline{p}^2$	$2\,n\overline{pq}$	$n\overline{q}^2$	n

$\overline{p} = (2n_{AA} + n_{Aa})\,/(2n)$
$GOF\chi^2 = \sum(O - E)^2/E$ is distributed as χ^2 with one degree of freedom under H_0, where summation is over all 3 genotypes

Example: CCR-5 Deletion. CCR-5 is a chemokind receptor which is involved in the human immune system. It enables the HIV virus to infect the CD4(+) T cells in ‘normal’ individuals and is necessary for AIDS to develop. A deletion of 32 base pairs causes the coding of incorrect amino acids, leading to a disruption of the normal functioning of the receptors. Because the mutation protects against HIV, we might expect to see an excess of two deletions in a sample of AIDS-free individuals. The data in Table 3.3 come from a study of 212 men who are free of AIDS, after many years of exposure to the HIV virus. The chi-square test is not

Table 3.3 Testing for HWE at CCR-5 locus in a sample of men at risk of HIV infection

	Genotype			
CCR5 Deletion	++	+−	−−	N
Observed	175	33	4	212
Expected	173	37	2	212
	$\bar{p}$ deletion = 0.097			
	$GOF\chi^2$ = 2.46 p-value = 0.11			

+ : Normal or wild type allele; − : Deletion

significant but the sample is small, especially the number of rare homozygotes; the pattern of observed genotypes is consistent with the idea that AIDS free individuals show an excess of two deletions.

Pearson's chi-square test is a large sample test, and the usual recommendation for its validity is that the expected value in each cell is greater than 5. An exact test which is valid for small samples is based on the idea that the number of heterozygotes will be either too big or too small if HWE fails. We can compute an exact test of HWE based on the number of observed heterozygotes, conditioning on the number of minor alleles that are observed (or on $\bar{p}$) and using the resulting hypergeometric distribution. For the data set in the CCR-5 deletion, the exact p-value is the same as the asymptotic one.

3.3.2 Some Causes of the Failure of HWE

Rejecting a test of HWE provides some evidence that HWE does not hold in the population. The failure of HWE is referred to as Hardy-Weinberg Disequilibrium (HWD). There are numerous reasons why HWE might fail, among them population substructure, selection, and genotyping errors. In general, the rejection of the test does not indicate a reason for failure, but there are some predictable patterns. As mentioned earlier in this chapter, selection of the sample with regard to a phenotype associated with the genotype will likely distort the genotype distribution in the sample. If a minor allele homozygous genotype infers greater risk of a disorder, then a sample of subjects with the disorder should have more rare homozygous genotypes, and correspondingly fewer heterozygotes than expected. In addition, the variability of the sample proportion will not follow the binomial formula. The effect of genotyping errors on HWE will be discussed in the chapter on genome wide association studies.

Exactly what happens to the genotype probabilities and/or var($\bar{p}$) depends on many features of the population, but the method of sampling and/or genotyping can also affect whether or not HWE will hold in the sample. We will now show that with population stratification and inbreeding, heterozygotes tend to be underrepresented relative to HWE, and that var($\bar{p}$) is inflated. We build on the notation in Box 3.3 to derive some general formulas. Let X be defined as the number of A alleles (as in Box 3.3), except we drop the i subscript for simplicity. By definition,

$$\begin{aligned}P(X=0) &= p_{aa}\\ P(X=1) &= p_{Aa}\\ P(X=2) &= p_{AA}. \end{aligned} \tag{3.4}$$

It follows by definition that

$$E(X) = 2p_{AA} + p_{Aa} = 2p$$

and

$$\text{var}(X) = 4p_{AA} + p_{Aa} - 4p^2. \tag{3.5}$$

If we substitute formulas 3.3 for genotype frequencies under HWE into formulas (3.4) and (3.5), we find $E(X) = 2p$ and $\text{var}(X) = 2pq$, as expected. We now use the general formulas (3.4) and (3.5) to show what happens when we have population stratification and inbreeding. Calculations for admixture are similar and will not be shown here.

Population Stratification**:** Assume a population with K strata, with allele frequencies p_k, and strata frequencies s_k, for $k = 1, \ldots, k$. Table 3.4 gives the genotype frequencies assuming HWE holds in each strata.

By definition, the allele frequency in the total population is

$$p = \sum_{k=1}^{K} s_k p_k,$$

and

$$\begin{aligned}P(X=1) &= 2\sum_{k=1}^{K} s_k p_k q_k = 2\sum_{k=1}^{K} s_k p_k(1-p_k) = 2p - 2E(p_k^2) + 2p^2 - 2p^2\\ &= 2pq - 2\text{var}(p_k). \end{aligned} \tag{3.6}$$

Table 3.4 Genotype frequencies by strata in a stratified population

Genotype Frequencies*					
Strata	s	p	AA	Aa	aa
1	s_1	p_1	p_1^2	$2p_1q_1$	q_1^2
2	s_2	p_2	p_2^2	$2p_2q_2$	q_2^2
⋮	⋮	⋮	⋮	⋮	⋮
K	s_K	p_K	p_K^2	$2p_Kq_K$	q_K^2

*Assuming HWE holds within each strata

In a similar matter, using formula (3.5), we find that

$$\mathrm{var}(X) = 2pq + 2\mathrm{var}(p_k).$$

Thus with a stratified population, $\mathrm{var}(X)$ is inflated relative to the binomial variance, and the frequency of heterozygotes $P(X = 1)$ is reduced relative to a population in HWE. When $\mathrm{var}(p_k) = 0$, the allele frequencies do not vary over strata, and there is no variance inflation.

Population Inbreeding**:** With inbreeding, there is a positive probability that an individual inherits the exact same A (or a) allele from both parents, meaning the parents have a common ancestor. Since the inbreeding coefficient, F, is defined as the probability that a randomly sampled individual will inherit the same copy from both parents, with inbreeding, we have

$$\begin{aligned} P(AA) &= P(X=2) = Fp + (1-F)p^2 \\ P(Aa) &= P(X=1) = 2pq(1-F) \\ P(aa) &= P(X=0) = Fq + (1-F)q^2. \end{aligned} \tag{3.7}$$

Note that $E(X) = 2p$, but there is a deficit of heterozygotes relative to HWE because $(1 - F)$ is less than one, and further,

$$\mathrm{Var}(X) = 4[Fp + (1-F)p^2] + 2pq(1-F) - 4p^2 = 2pq(1+F),$$

is inflated relative to the variance of a HWE population. This deficit of heterozygotes (*Loss of Heterozygocity (LOH)*) due to population substructure (stratification, admixture and/or inbreeding) is known as the Wahlund effect. For statistical tests or models that assume HWE in their variance calculations, the Wahlund effect can lead to bias and incorrect inference.

3.3.3 Measuring the Departure from HWE

We have seen that the observed number of heterozygotes tends to be too small by a factor of $(1 - F)$ with inbreeding. Under the simple inbreeding model for $P(Aa)$ given in formula (3.7), an estimate of the inbreeding coefficient is given by

$$\hat{F} = 1 - O/E,$$

where O is the observed number of heterozygotes, and E is the expected number of heterozygotes calculated for the test of HWE. With inbreeding, $\hat{F}$ should be positive since we expect to observe fewer O than E. For the simple model of population stratification given above, $\hat{F}$ also can be shown to estimate the correlation between maternal and paternal alleles induced by population stratification (See exercise 12 of Section 3.4). Box 3.5 illustrates the calculation of the coefficient of inbreeding for two samples from Table 3.1.

Box 3.5 Estimation of the Inbreeding Coefficient

The coefficient of inbreeding can be estimated from observed genotype frequencies in a sample to describe the degree of inbreeding and/or population stratification and admixture in a population. We consider two samples from Table 3.1:

Bassett Hound	O = 2	$E = 32 * 2(0.98)(.02)$	$\hat{F} = 0.51$
Tibetan Terriers	O = 10	$E = 24 * 2(0.64)(.36)$	$\hat{F} = 0.096$

Note that the F allele has disappeared from the Basenji breed, at least in the sample reported in Table 3.1.

3.4 Exercises

1. The allele frequency at a locus on the X chromosome is defined as the proportion of the X chromosome alleles carrying the A variant. Given a random population sample, tell how to estimate the A allele frequency at a locus on the X chromosome. Assuming HWE holds, what is the variance of the estimated A allele frequency and how would you estimate the variance?
2. Assume you observe that the proportion of a population affected with sickle cell anemia is 0.01. Assuming an autosomal recessive disease model and HWE, estimate the frequency of the sickle cell mutation at the hemoglobin locus in this population.
3. Construct a chi-square test using the data on MN Blood Group Frequencies from Greenland to test the null hypothesis that the genotype distributions are the same in the two regions.
4. Assume a rare recessive Mendelian disease caused by a mutation with minor allele frequency p in a population of size N. How many cases of the disease do you expect when HWE holds? How many do you expect when the inbreeding coefficient in the population is F and the genotype frequencies are given by formulas (3.7)? Evaluate these expectations when $p = 0.0001$, $N = 10^6$ and $F = 0.2$.
5. Assume the genotypes AA, Aa and aa have frequencies u, v and w in a randomly mating population. By considering all possible outcomes of all possible

mating types, show that the offspring genotypes follow HWE with the same allele frequencies in both generations.

6. Test HWE for the MN blood group locus separately in the Southwest and the East Greenland samples.
7. Assume that HWE holds at the MN blood group locus for both the Southwest and the East Greenland samples. Use a large sample test to test the null hypothesis that the allele frequencies in the two populations are equal; defend your choice of test.

   Hint: Consider testing the difference in two binomial proportions given two independent samples.
8. Of the different dog breeds given in Table 3.1, calculate F for the Mongrels, the 'Other pure breeds', and the Basenji. Comment on the different values. Why might you expect F for Basenji to be one? Why is F for the 'Other pure breeds' combined higher than F for Mongrels?
9. Show that the Pearson Goodness of Fit chi-square test for HWE has only one degree of freedom by showing that:

   (a) the three $(O - E)$ residuals sum to zero and
   (b) the two homozygote residuals are equal.

10. Verify that $\text{Var}(X) = 2pq + 2Var(p_k)$ for a stratified population.
11. Use equations (3.4–3.5) to derive an estimate for $\text{Var}(\bar{p})$ that does not assume HWE.
12. Show that when random mating fails, $\text{Var}(X) = 2pq(1 + p)$, where p is the correlation between maternal and paternal alleles.

Chapter 4
Aggregation, Heritability and Segregation Analysis: Modeling Genetic Inheritance Without Genetic Data

Aggregation and *heritability* analyses are designed to show that diseases, or phenotypes more generally, have a genetic basis by investigating patterns of phenotypic correlation between relatives; *segregation* analysis is used to find support for a specific genetic model underlying the inheritance patterns observed in families. They all involve modeling phenotypic data on families, or pedigrees, without using any genetic data. As such, all were developed during the time when genotyping was expensive, labor intensive, and not widely available. Today, the general concepts used in aggregation and heritability analysis are widely accepted as useful measures of the degree to which traits are inherited; most researchers would not undertake genetic analysis without evidence of aggregation or heritability of the trait. Using segregation analysis to determine the model of inheritance at the disease locus was essential in planning parametric linkage analyses, as described in Chapter 6, but the current popularity of non-parametric linkage analysis and association analysis has put segregation analysis somewhat on the sideline. Although this chapter can be skipped if the reader's primary interest is association, our coverage of these methods is brief and the concepts are useful to anyone with an interest in statistical genetics. In particular, the approach used to construct a likelihood for pedigree data given in Section 4.1 serves as a basis for other analyses in linkage and association discussed in later chapters.

In general, we will refer to a disease gene when the trait or phenotype of interest is dichotomous, e.g., affection status, as a Disease Susceptibility Locus or DSL. For phenotypes and traits that are measured on a quantitative scale, the corresponding genetic locus is typically called quantitative trait locus or QTL. For either trait/gene type, statistical models can be used to estimate the genetic effect sizes of the loci and understand their mode of inheritance, even without having genotypic data available. We can assess the evidence for the presence of a disease gene purely based on phenotypic data from related individuals, using the concepts underlying Mendel's Laws and the statistical models for the penetrance functions.

Aggregation Analysis (for dichotomous traits): By estimating the correlation or similarity of a phenotype among family members, one can assess whether a phenotype aggregates in families. While a positive result of an aggregation analysis confirms the plausibility of a disease gene, it cannot rule out common environmental effects within families as the origin for the observed correlations. With dichotomous

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_4,

traits, family samples are ordinarily *ascertained* by at least one affected individual (the *proband*), i.e., probands are selected by choosing those individuals with $Y = 1$. As a result, the distribution of phenotypes in the sample does not reflect the population and it is not possible to unbiasedly estimate correlations directly from the sample. We will introduce a measure closely related to the correlation, the *recurrence risk ratio*, and illustrate its properties by examining how it depends on the underlying disease model parameters, i.e., the penetrance probabilities, the mode of inheritance and the allele frequencies, as well as the degree of relatedness. To become more familiar with the statistical models and their 'mechanics', we use simple algebra to derive the recurrence risk ratio as a function of the population attributable fraction, which is a commonly used measure of effect size for dichotomous traits.

Heritability Analysis (for quantitative traits): Similar in character to aggregation analysis, the goal of heritability analysis is to estimate the overall genetic effect of the quantitative trait. This effect is defined as the proportion of the total variability in the phenotype explained by variation in all loci underlying the qualitative trait. This proportion is typically referred to as heritability. Samples drawn from a population without regard to their phenotype are referred to as *not ascertained*. In these samples, the heritability can be estimated by examining the phenotypic correlations between relative pairs in pedigrees. This concept will be illustrated with examples.

The last step in the analysis sequence is typically *Segregation Analysis*. By examining the inheritance patterns of the disease phenotype and the transmission of disease from one generation to the next generation within one family, a formal statistical model is fit to the observed pedigree data, generally using maximum likelihood theory. The likelihood function depends on the Mendelian transmission probabilities and the unknown disease parameters, e.g., minor allele frequencies of the disease susceptibility loci, the penetrance probabilities and the mode of inheritance. Various likelihood models, e.g., different mode of inheritance, number of disease loci, are estimated and compared. In the case of nested models, likelihood-ratio tests are used to compare the model fit. For non-nested models, standard model comparison/selection criteria such as AIC or BIC are used (Burnham and Anderson 2004). Besides the formal statistical test for the presence of a disease gene or QTL, the importance of segregation analysis stems from the parameter estimates for the disease that we acquire by fitting the likelihood models. A segregation analysis provides estimates for the number of possible loci, their penetrance probabilities and their allele frequencies. Estimates for these parameters will be required later on for parametric linkage analysis of genetic data.

4.1 Preliminaries

Underlying all of the methods discussed in this chapter is the need to describe the joint distribution of phenotypes and genotypes of individuals in a family, taking into account a disease model and their sharing of the disease alleles. In this section, we describe the general approach, and apply it in the following sections. For simplicity, we will assume a nuclear family with two parents and two offspring (siblings), but the general approach extends readily to more complex pedigrees. Let D and d denote the disease and non-disease (sometimes called wild type) alleles, respectively, and let p denote the frequency of the D allele in the population. We denote an individual's genotype by the number of disease alleles, i.e., each individual's genotype can take on the values 0, 1, or 2, depending on the number of disease alleles, dd, dD, or DD. Let X_1, X_2, P_1 and P_2 denote the genotypes of the two siblings (X_j) and the two parents (P_i), for $i, j = 1, 2$. Following convention, the capital letters denote random variables, and the lower case denotes the values that the random variables take on.

It is conventional to specify allele frequencies and use them to calculate genotype frequencies assuming Hardy-Weinberg Equilibrium holds in the population. Assuming HWE allows us to show that for any individual in the population, parent or child, the probability of having 0, 1, or 2 alleles is $B(2, p)$, or equivalently, each allele inherited from a parent is independent with $P(D \text{ allele}) = p$. As discussed in Chapter 3, this is a rather restrictive assumption, but modest departures are unlikely to have much effect in the likelihood calculations we discuss. Assuming HWE also implies random mating, so that the parental genotypes are independent, i.e.,

$$P(P_1 = g_1, P_2 = g_2) = f(g_1, g_2) = f(g_1)f(g_2),$$

where $f(.)$ denotes a probability density function, either joint or marginal, i.e.,

$$f(g_i) = P(P_i = g_i),$$

and g_i is the genotype of the ith parent. Assuming only two alleles at the DSL, we can let the range of both x_j and g_i be 0, 1, 2, denoting the number of D alleles that each individual in the family has. Further, the genotypes of the offspring are independent conditional on the parental genotypes, and each follows Mendel's first law. Thus we have the joint distribution of genotypes in the family is given by

$$f(x_1, x_2, g_1, g_2) = f(x_1|P_1, P_2 = g_1, g_2)\, f(x_2|P_1, P_2 = g_1, g_2)\, f(g_1)\, f(g_2), \tag{4.1}$$

where the conditional density functions for X_1 and X_2 are completely known and given by Mendel's first law (Table 2.1). Note that although X_1 and X_2 are conditionally independent given parents, marginally they are not, i.e., $f(x_1, x_2)$ does not factor when we sum equation (4.1) over g_1 and g_2. As we show, the lack of unconditional genotypic independence induces correlations among the phenotypes of siblings.

In this chapter, we assume that no genotype data are observed; rather we work with the joint distribution of phenotypes in a family summing over the unobserved genotypes. For simplicity, we make the assumption of *phenotypic independence*. Phenotypic independence implies that the phenotypes of individuals in the pedigree are independent of each other, given their genotypes. This is a reasonable assumption for Mendelian disorders or monogenic disorders, with a single DSL and limited environmental effects. For complex disorders with multiple DSLs and environmental factors, it is desirable to use more complex models, which take into consideration shared environmental factors. We also make the commonly used assumption that, conditional on an individual's genotype, their phenotype does not depend on the genotype of any other family member. Letting Y_j denote the phenotypes for the two offspring, $j = 1, 2$, we thus have, for example

$$f(y_1, y_2|x_1, x_2, g_1, g_2) = f(y_1|x_1)\, f(y_2|x_2). \tag{4.2}$$

Finally then, the probability density for the offspring phenotypes and genotypes, and the parental genotypes is:

$$f(y_1, y_2, x_1, x_2, g_1, g_2) = f(y_1|x_1)\, f(y_2|x_2)\, f(x_1|g_1, g_2)\, f(x_2|g_1, g_2)\, f(g_1)\, f(g_2). \tag{4.3}$$

In order to obtain the density of just the observed offspring phenotypes, $f(y_1, y_2)$, we sum equation (4.3) over all possible genotype values of both offspring and parents; see for example equation (4.5).

4.2 Aggregation Analysis

Aggregation analysis is based on the principle that the genetic material within a family is inherited following Mendel's laws. Consequently, two related individuals are likely to share more genetic material at any given locus than two unrelated individuals from the general population. For example, letting a mother's two alleles be A_1 and A_2, when we consider two full siblings, the probability that they share the same maternal allele is $P(\text{both sibs inherit } A_1) + P(\text{both sibs inherit } A_2) = (\frac{1}{2} * \frac{1}{2}) + (\frac{1}{2} * \frac{1}{2}) = \frac{1}{2}$. The same holds for the sharing probability of the paternal allele, i.e., two siblings share the paternal allele with probability $1/2$, hence the probability of two sibs sharing both parental alleles is $1/4$. Similar calculations can be done for more distant relative pairs (See Appendix A). If the phenotype of interest has a genetic component, the relative of an affected subject will have a higher predisposition to disease than an unrelated subject in the general population, because of the shared genetic material among relatives.

The strength of the genetic aggregation among relatives is generally measured by the *recurrence risk ratio*. It is defined as a probability ratio which compares the probability of a study subject being affected given that a relative is affected to the general risk in the population. The general risk in the population is commonly referred to as the population prevalence of the disease, and denoted by K. For the relative of an affected individual, the recurrence risk ratio is thus:

$$\lambda_R = P(Y_2 = 1|Y_1 = 1)/K, \tag{4.4}$$

where R denotes the relative type, and the variables Y_1 and Y_2 are the affection status of the two relatives, where $Y = 1$ denotes affected and $Y = 0$ denotes unaffected and

$$K = P(Y_1 = 1) = P(Y_2 = 1).$$

Note that we are assuming a very simple model which implies that $P(Y = 1) = K$ regardless of age, relative order (i.e., parent versus child) and any non-genetic effects. These assumptions are not unreasonable for Mendelian disorders present at birth, but for more complex disorders, this analysis is only approximate. As an exercise (exercise 2 of Section 4.5) you are asked to derive the connection between the recurrence risk ratio and the covariance between the two relatives' phenotypes, i.e., $\text{cov}(Y_1, Y_2)$.

For any given disease, in this simple case the recurrence risk ratio depends only on the degree of relatedness of the two relatives, the underlying genetic model and p. For example, we expect that first degree relatives (siblings, parent-offspring pairs) will have a larger recurrence risk ratio than will second or third degree relatives, or mother/father pairs, who will share no genetic material in the absence of inbreeding. Of course monozygotic (MZ) twins should have the highest recurrence risk ratio since they share all of their genetic material, while dizygotic (DZ) twins should have recurrence risk ratios similar to siblings. Table 4.1 illustrates recurrence risk ratio estimated from a sample of families with members affected with schizophrenia. As expected, the risk ratio is highest for relative pairs sharing the most alleles.

To study the dependence of the recurrence risk ratio on the disease model, the recurrence risk ratio has to be expressed as a function of the penetrance probabilities and the allele frequency at the DSL. In order to keep the algebraic derivations simple, we will focus here only on the sibling recurrence risk ratio, i.e., disease sharing among a pair of siblings, and assume only one DSL. As in Chapter 2, the set of penetrance probabilities is denoted by f_0, f_1 and f_2. These penetrance probabilities determine the probability of offspring disease status, given their genotypes, $X_1 = x_1$ and $X_2 = x_2$:

$$f_j = P(Y_j = 1|X_j = x_j) \text{ for } x_j = 0, 1, 2.$$

The recurrence risk ratio can be re-written by including the unknown genotype data of the siblings and the parents in the joint probability, summing over all unknown genotype configurations and then using Bayes rule to re-express the conditional probability $P(Y_2 = 1|Y_1 = 1)$ in terms of the joint probability $P(Y_1 = 1,$

Table 4.1 Observed recurrence risk ratios from a sample of families with schizophrenia. *Source*: Risch (1990a)

Risk Ratio	λ_O	λ_S	λ_M	λ_D	λ_H	λ_N	λ_G	λ_C
Observed	10.0	8.6	52.1	14.2	3.5	3.1	3.3	1.8

Definitions of subscripts: O = offspring; ; S = sibling; M = MZ twins; D = DZ twins; H = half-sibs; N = niece/nephew; G = grandchild; C = first cousins.

$Y_2 = 1$) and the marginal probability $P(Y_1 = 1)$. The recurrence risk ratio can thus be written as

$$\lambda_S = P(Y_1 = 1, Y_2 = 1)/P(Y_1 = 1)^2.$$

By definition, the numerator can be expressed as:

$$\sum_{x_1,x_2,g_1,g_2=0,1,2} P(Y_1 = 1, Y_2 = 1, x_1, x_2, g_1, g_2)$$
$$= \sum_{g_1,g_2=0,1,2} f(g_1) f(g_2) \left[\sum_{x_1=0,1,2} f_{x_1} f(x_1|g_1, g_2) \sum_{x_2=0,1,2} f_{x_2} f(x_2|g_1, g_2) \right] \quad (4.5)$$

where $f(g_i)$ is given by HWE for $i = 1, 2$ and $g_i = 0, 1, 2$; $f(x_i), i = 1, 2$ are the penetrance functions for the two siblings, and $f(x_i|g_1, g_2)$ is defined by Mendels' law. The denominator can be expressed as:

$$K^2 = \left[\sum_{x=0,1,2} P(Y = 1|x) f(x) \right]^2 = \left[f_0(1-p)^2 + f_1 2p(1-p) + f_2 p^2 \right]^2. \quad (4.6)$$

For simple Mendelian models, f_x will be zero for $x = 0$ and possibly $x = 1$ as well, hence many terms drop out of the summation. Thus given values for the penetrance functions and the allele frequency, the recurrence risk ratio is easily computed.

Table 4.2 illustrates how these formulas are used in the calculation of the joint probabilities of a pair of relative genotypes. From Table 4.2 we see that the P(pair of offspring genotypes) is given by

$$P(OG) = \sum_{\textit{all mating types}} P(MT)P(OG|MT).$$

Thus, for example $P(DD, Dd) = \frac{1}{2}(4p^3(1-p) + \frac{1}{4}(4p^2(1-p))$.

Table 4.2 Calculation of parent-offspring genotype distribution for a pair of siblings

Mating type (MT)	$P(MT)$	Offspring genotypes (OG)	$P(OG\|MT)$
DD x DD	p^4	(DD,DD)	1
DD x Dd	$4p^3(1-p)$	(DD,DD) (DD,Dd) (Dd,Dd)	$\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$
DD x dd	$2p^2(1-p)^2$	(Dd,Dd)	1
Dd x Dd	$4p^2(1-p)^2$	(DD,DD) (DD,Dd) (DD,dd)	$\frac{1}{16}, \frac{1}{4}, \frac{1}{8}$
		(Dd,Dd) (Dd,dd) (dd,dd)	$\frac{1}{4}, \frac{1}{4}, \frac{1}{16}$
Dd x dd	$4p(1-p)^3$	(Dd,Dd) (Dd,dd) (dd,dd)	$\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$
dd x dd	$(1-p)^4$	(dd,dd)	1

4.2.1 Estimating Recurrence Risk Ratios

Table 4.1 illustrates estimates of recurrence risk ratios derived from data on families where schizophrenia is segregating. Recurrence risk ratios such as those shown in Table 4.1 can be estimated from pedigrees or using a standard case-control study of familial risk. A sample of unrelated cases and unrelated controls are obtained, and disease history is evaluated for all relatives, most commonly first degree relatives. Where possible, actual clinical diagnoses are obtained for relatives. To obtain recurrence risk ratios for MZ or DZ twins generally requires data from twin registries. Analyzing the cases and controls separately we can obtain estimates of the sibling risk quite simply as the proportion of affected siblings among all siblings of case probands, and likewise for the control probands. We denote these proportions as s_{case} and s_{control}. Then s_{case}/K estimates λ_S for siblings. In the absence of knowledge of K for the population, we can use $s_{\text{case}}/s_{\text{control}}$ to approximate λ_S when disease is rare (so that s_{control} is approximately K), however, Javaras et al. (2010) have shown how K can be estimated from the data in the sample. The analysis we have described is very simple and intuitive but does not allow for adjusting for variable age at onset or environmental factors. For discussion on estimating recurrence risk ratios, see Guo (1998) or Laird et al. (2000a).

Several studies of familial aggregation were used to justify genetic studies of Alzheimer's disease. The recurrence risk ratio in first degree relatives has variously been reported to be between 1.05 and 2–4, but it depends considerably on age-at-onset, with some studies showing almost no increased risk to relatives at very late age of onset.

4.2.2 Further Simplifications

To derive simpler expressions for λ_R that are a function only of allele frequency, we consider recessive, dominant and additive models, which have the property that it is only necessary to define f_0 and f_2, as then f_1 is automatically determined. For a recessive model, $f_1 = f_0$, for a dominant model, $f_1 = f_2$ and for the additive model, $f_1 = (f_0 + f_2)/2$. A measure that is commonly used to obtain standardized genetic effect sizes that can be compared across genetic models is the *attributable fraction* (AF). The *attributable fraction* assesses the genetic effect relative to the disease prevalence and is defined by

$$AF = (K - f_0)/K = 1 - P(Y = 1|\text{no risk alleles})/P(Y = 1),$$

Thus the AF defines the proportion of disease caused by having at least one disease allele. For diseases which have no genetic basis, this proportion will be zero, while for Mendelian disorders which arise solely as a result of mutations at the DSL (thus $f_0 = 0$), this proportion will be one. The attributable fraction and the prevalence are generally more intuitive parameters and easier to specify than the penetrance functions. If we keep the prevalence (K) and the attributable fraction (AF) fixed we can replace f_0 and f_2 (and also f_1) in the expression for λ_S, allowing us to re-express λ_S as a function of AF and p only (see exercise 4 of Section 4.5). As a result, for the three models, we have

Recessive mode of inheritance:

$$\lambda_S = 1 - \frac{(3p+1)(p-1)}{4p^2} AF^2,$$

Dominant mode of inheritance:

$$\lambda_S = \frac{4p(p-2)^2 + AF^2\left(4 - 11p + 10p^2 - 3p^3\right)}{4p\,(2\,(1 - AF(1-p)) - p)^2},$$

Additive model:

$$\lambda_S = \frac{4p + AF^2(1-p)}{4p\left[1 - AF(1-p)\right]^2}.$$

In general, under all genetic models, the recurrence risk ratio will increase with increasing values for the attributable fraction AF and with decreasing values for the disease allele frequencies p. Intuitively, if the disease allele is very common, then unrelated individuals may share the disease allele with high frequency, but with a rare disease allele, only relatives are likely to share.

The sibling recurrence risk ratio varies qualitatively in its dependence on the disease allele frequency under the different genetic models. To illustrate, we assume a monogenetic disease with no phenocopies, i.e., $f_0 = 0$, which means that there is only one DSL, and no environmental causes of disease. The absence of phenocopies implies that the attributable fraction AF reaches its maximum value of 1. Since the additive model is not plausible for most monogenetic diseases with high penetrance probability, we will restrict the considerations here to the dominant and recessive model. The sibling recurrence risk ratios under the recessive and dominant mode of inheritance are given by (see exercise 5 of Section 4.5):

$$\lambda_S = \tfrac{1}{4} + \tfrac{1}{2p} + \tfrac{1}{4p^2},$$
$$\lambda_S = \tfrac{1}{4} - \tfrac{3}{2p} + \tfrac{5}{4p^2} + \tfrac{1}{p^3}.$$

When the disease allele frequency is low, the effect of allele frequency on λ_S is dominated by the term with the highest power of p in the denominator. Thus under a dominant model, the ratio increases at a rate of $1/p^3$ when the disease allele frequencies become small. Under the recessive model, the ratio increases with $1/p^2$. This illustrates that the recurrence risk ratio will be a powerful tool to confirm the existence of DSLs for monogenetic diseases with very small minor allele frequencies, and small disease prevalence. This will work best under a dominant mode of inheritance. This dependence of the recurrence risk on the mode of inheritance and the disease prevalence is also reflected in Fig. 4.1. Here the sibling recurrence risk ratio is displayed as a function of K for a variety of diseases. Figure 4.1 also shows the limitation of the recurrence risk ratio as a tool for inference about existence of disease genes. While all monogenetic diseases have high recurrence risk ratios, regardless of their mode of inheritance, the risk ratio for the complex diseases are

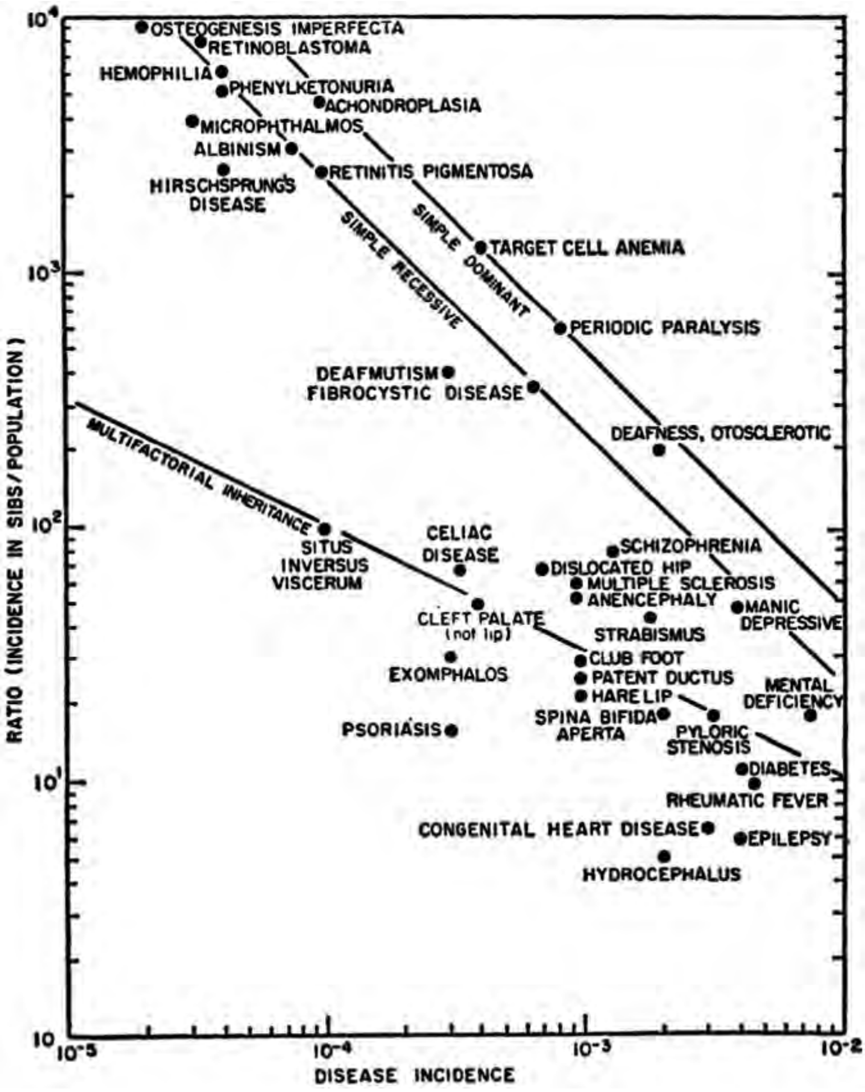


Fig. 4.1 Recurrence risk ratios for Mendelian and non-Mendelian diseases. Relation between disease incidence and relative incidence in sibs of affected individuals for a number of diseases. The lines indicate the expected relationships for simple dominant, simple recessive and Edwards (1963) approximation to multifactorial inheritance (Newcombe 1964)

relatively small which will make it difficult to distinguish genetic effects from environmental correlation between the two siblings.

4.3 Heritability Analysis

We assume here that the trait of interest is measured on a quantitative scale, i.e., height, weight, blood pressure, etc. Using quantitative phenotypic data on relatives, heritability analysis assesses the overall genetic component of a quantitative trait, relative to the total observed phenotypic variation of the trait. In the model for the recurrence risk ratio (dichotomous traits), we assumed the presence of one disease gene. Under a recessive or dominant genetic model, Fig. 4.1 suggests that this is a plausible assumption for many rare diseases that exhibit Mendelian inheritance patterns. However, for common diseases and more generally, quantitative phenotypes, the single disease locus assumption is generally considered to be less plausible and the inheritance patterns of disease within pedigrees support the hypothesis of multiple genes acting jointly. Thus quantitative traits are typically modeled as a function of multiple QTLs. Additionally, environmental variables can easily be incorporated as in linear regression models, but we omit this complexity here. A quantitative phenotype Y can thus be modeled quite generally by

$$Y = \mu + \sum_{m=1,..,M} \{a_m X_m + d_m I\,[X_m = 1]\} + \epsilon, \tag{4.7}$$

where M is the unknown number of QTLs, X_m denotes the number of disease alleles at the i^{th} disease locus, and $I[X_m = 1]$ is an indicator function which is 1 if $X_m = 1$ and 0 otherwise. The parameter μ is the phenotypic mean for individuals who have no disease alleles at any loci ($X_m = 0$ for all i); a_m and d_m are parameters which partition the genetic effect at the m^{th} QTL into its additive and codominant components. The additive effect is simply the increase in Y expected from increasing the number of disease alleles by 1. The codominant component is here understood as the departure of the model from the additive mode of inheritance. This formulation of the mode of inheritance includes all previously discussed genetic models. For example, if $d_m = 0$, the additive mode of inheritance is obtained at the ith QTL. For a dominant or recessive mode of inheritance, the parameter d_m is set to $-a_m$ or a_m, respectively. For monotone penetrance functions, $(-a_m < d_m < a_m)$, for a_m positive, and the reverse for a_m negative.

The random variable ϵ incorporates unspecified environmental/non-genetic influences on an individual's trait into the model. Typically, ϵ is assumed to be normally distributed with mean 0 and variance σ^2. We also often make the assumption of phenotypic independence, i.e., $\text{cor}(Y_k, Y_l) = 0$, conditional on the genotypes of the subjects k and l at all M QTLs, although with normally distributed phenotypes, it is easy to introduce phenotypic correlation. As in previous sections, the X_ms are treated as independent, unobserved random variables, whose distribution is defined by HWE using the allele frequency p_m at each QTL.

In order to assess the overall genetic contribution to the variation in the phenotype, the variance of the phenotype is partitioned into a genetic part and a non-genetic part.

$$\mathrm{Var}(Y) = \mathrm{Var}(G) + \mathrm{Var}(\epsilon) + 2\mathrm{Cov}(G, \epsilon)$$

where $\mathrm{Var}(G) = \mathrm{Var}(\sum_{m=1,..,M}(a_m X_m + d_m I[X_m = 1]))$. In a heritability analysis, one typically assumes that the covariance between the genetic effects and the environment is zero, i.e., $\mathrm{Cov}(G, \epsilon) = 0$. While this assumption is not true in general, it is a reasonable hypothesis in heritability analysis, when the goal is to assess the genetic contributions to the overall variation of the phenotype. In terms of the proportion of explained phenotypic variation, the genetic main effects are typically much larger than gene-environmental interactions, i.e., $\mathrm{Cov}(G, \epsilon)$ is small compared to $\mathrm{Var}(G)$. The *broad-sense heritability* of a trait is defined as the proportion of the overall phenotypic variation in the trait that is attributable to genetic components, e.g.,

$$\mathrm{Var}(G)/\mathrm{Var}(Y).$$

Based on statistical models used in animal and plant genetics that predict the subject's phenotype conditional on the parental phenotypes, the genetic variance can be partitioned into the *Additive Genetic Variance V_A* and the *Dominant Genetic Variance V_G* (Falconer and Mackay (1996)):

$$\begin{aligned} Var(G) &= V_A + V_D \\ V_A &= \sum_m 2p_m(1-p_m)(a_m + d_m(1-2p_m))^2 \\ V_D &= \sum_m (2p_m(1-p_m)d_m)^2 \end{aligned} \tag{4.8}$$

where p_m denotes the minor allele frequency of the mth marker. It can be shown that the additive genetic variance is a function of the average effect of the parents' genes on the offspring's phenotype (the 'breeding value' in animal and plant genetics) and that the additive genetic variance can be estimated based on the parental phenotypes (Falconer and Mackay (1996)). With the exception of situations in which the mode of inheritance is assumed to be heterozygous advantage (i.e., d_m is outside the range $(-a_m, a_m)$), the dominant genetic variance is relatively small compared to the additive genetic variance and it is often reasonable to assume that the total genetic variance $\mathrm{Var}(G)$ is approximately equal to the additive genetic variance.

This leads to the definition of the *narrow sense heritability* (h^2) which is based on the additive variance. The narrow sense heritability (h^2) is defined as the proportion of the phenotypic variance that is explained by just the additive genetic effects,

$$h^2 = V_A/\mathrm{Var}(Y).$$

As indicated above, the advantage of the narrow sense heritability is that it can be directly estimated from the phenotypic data on relatives. We illustrate this feature in Box 4.1 by using trios, i.e., nuclear families with one offspring and both parents, to estimate the narrow-sense heritability.

Box 4.1 Estimation of h^2 Using Parent-Child Trios

Assuming that we have phenotypic data on both parents, we define the mid-parental value of the trait, Y_P as the average phenotypic value of the two parents; we let Y_O denote the phenotype of the offspring. We correspondingly let X_P denote the average of the two parental genotypes, and X_O be the genotype of the offspring. We assume that the genotype probabilities for each parent are independently distributed as $B(2, p)$; the probabilities for each child's genotype, conditional on the parents, are given by Mendel's law. In the first step, we derive the covariance between the offspring phenotype and the mid-parental phenotype, assuming for simplicity $M = 1$ and $d = 0$. In the following derivations, we provide the results of each intermediate step. The algebraic derivation of each step is a homework assignment. It is straightforward to see that for any member of the trio,

$$E(Y_O) = E(Y_P) = E(aX_O) = \mu + 2ap,$$

and

$$V_A = \text{var}(aX_O) = 2a^2 p(1 - p).$$

By properties of the covariance function, we have

$$\text{Cov}(Y_O, Y_P) = a^2 \text{cov}(X_O, X_P) = a^2[E(X_O X_P) - 4p^2].$$

Tedious algebra using Mendel's laws shows that $\text{cov}(X_O, X_P) = p(1 - p)$ and thus

$$\text{Cov}(Y_O, Y_P) = a^2 p(1 - p) = V_A/2.$$

This implies that the phenotypic covariance between the offspring and the mid-parental value is half the genetic variance V_A. Since this covariance and the phenotypic variance can be estimated based on the corresponding empirical estimators, $\text{Cov}(Y_O, Y_P)$ and $\text{Var}(Y_O)$, an estimator for the narrow-sense heritability is given by

$$h^2 = 2\text{Cov}(Y_O, Y_P)/\text{Var}(Y) = 2\rho,$$

where ρ is the correlation between Y_O and Y_P.

Table 4.3 Familial correlations for body mass index derived from four large family studies. *Source* Coon et al. (2007)

Relationship	Framingham heart study	Canada fitness survey	Quebec family survey	Norway study
Spouses	0.19	0.12	0.10	0.12
Parent-Offspring	0.23	0.20	0.26	0.24
Uncle/Aunt-Nephew/Niece	0.08	−0.11	0.14	0.00
Grandparent-Grandchild	NA	0.05	NA	0.07
Dizygotic Twins	NA	NA	0.34	0.20
Monozygotic Twins	NA	NA	0.88	0.58

For other relative pairs, similar estimators for the narrow heritability h^2 can be derived using the sample variances and covariances. A detailed discussion of other relative pairs and the corresponding estimators is provided by Falconer and Mackay (1996). It is important to note that the heritability estimate depends on the allele frequency of the disease allele and the strength of the environmental component ϵ. Since both quantities can be population-specific, the heritability h^2 can vary among populations and change even within one population over time, as allele frequencies vary and environmental influences alter. Heritability does not depend upon the degree of relationship between pairs of relatives, although correlation between relative pairs does. Phenotypes that show high genetic heritability in humans are variables such as height, weight, cardiac QT interval and gene expression profiles. Table 4.3 illustrates variability in correlations across populations for body-mass index (BMI). While there are certainly differences between the four populations, their relative differences are small, suggesting a common genetic component for this trait in these populations.

4.4 Segregation Analysis

The idea of segregation analysis is to test and to compare different statistical models formally, using phenotypic data on related individuals, with the goal of identifying the genetic model that describes the data ‘best’. Genetic model is here understood to incorporate the number of DSLs (or QTLs) and mode of inheritance. Likelihood models are constructed that explain the phenotype or trait as a function of the unobserved disease genes or QTLs, and the ascertainment condition. The models contain as unknown parameters the penetrance probabilities (or functions) of the disease genes and the frequencies of their disease alleles. Using maximum likelihood estimation, different models are fit to the data and formally compared by likelihood ratio tests or by model selection criteria such as AIC, BIC, etc. The model with the ‘best’ fit is then used to provide insight into the number of the underlying disease loci/QTLs and their mode of inheritance. As before, the genotype data are unknown, and must be summed over to form a likelihood. As a consequence, maximum likelihood estimation can be computational intensive. One of the important features of segregation analysis is that it provides estimates for the unknown penetrance probabilities and allele frequencies. Since these parameter estimates are required for parametric linkage analysis, segregation analysis often played an important role in the process of mapping disease genes using parametric linkage analysis. A detailed discussion of segregation analysis is beyond the scope of this book; we focus here on special simple cases that illustrate the general principle. In our discussion we will focus on Mendelian diseases.

4.4.1 Segregation Analysis for Dominant Mendelian Diseases

Mendelian Diseases are often rare diseases that are caused by one or a small number of mutations. Their inheritance and features can best be studied by examining the transmission paths of the disease through multi-generational pedigrees. See Box 4.2 for an explanation of symbols used in drawing pedigrees. In general, one assumes that the frequency of the disease susceptibility allele is small for a rare disease. As a consequence, with a dominant disease, the frequency of the rare homozygous genotype $P(DD)$ is so small that its appearance in affected individuals is considered to be negligible, i.e., all affected individuals are Dd. To keep the consideration simple, we will assume that the gene causing the disease is fully penetrant $(P(Y = 1|dD) = \Pr(Y = 1|DD) = 1)$ and that there are no phenocopies, i.e., $\Pr(Y = 1|dd) = 0$, thus the genetic model is completely known.

Box 4.2 Conventions for Drawing Disease Pedigrees

Males are denoted by squares
Females are denoted by circles
Affected probands are shaded
Left-shaded individuals are carriers, i.e., they carry the D allele, but are not affected themselves
Parents are connected by horizontal lines
Offspring are connected to parents via vertical lines
Double horizontal line represents a consanguineous mating (inbreeding)

Figure 4.2 shows an example for a pedigree with an autosomal dominant disease. The following features that can be observed in Fig. 4.2 are typical patterns of fully penetrant dominant diseases:

1. All affected individuals will have at least one affected parent. Since one copy of the disease allele is sufficient to trigger the disease, the parent who passes on the disease allele to the offspring also has to be affected. The disease appears in every generation.

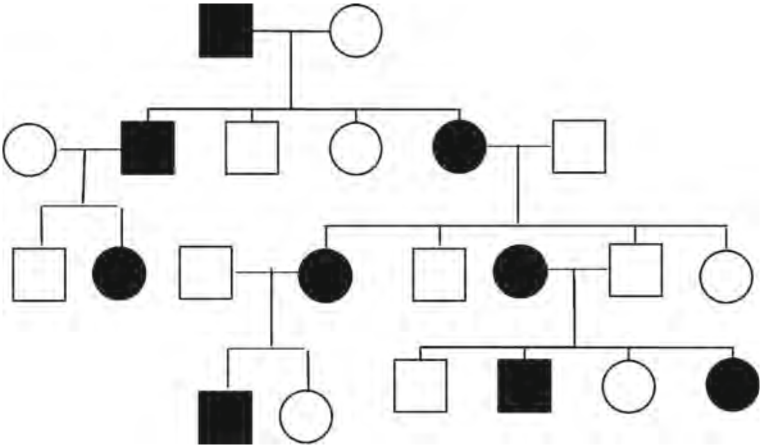


Fig. 4.2 Autosomal dominant inheritance. *Source*: Adapted from Thomas (2004)

2. Offspring of two unaffected parents are also unaffected. Given the full-penetrance of the disease susceptibly allele D and the absence of phenocopies, being unaffected implies that a subject cannot be a carrier of the disease susceptibly allele D and, consequently, cannot pass on the disease to its offspring.
3. Males and females are affected equally, and both transmit the disease to the offspring. Since the disease locus is based on one of the autosomal chromosomes, gender cannot affect the inheritance of the disease.

Given these characteristics and assuming no phenocopies and a fully penetrant DSL, it is easy to test statistically for the presence of a dominant effect, by comparing the likelihood for a dominant Mendelian disease to the likelihood of a more general model. For some of the mating types (e.g., $dd \times dd$), the affection status of the offspring can be predicted with 100% certainty; such mating-types will not be informative for the analysis. Mating types for which the phenotypic outcome in the offspring cannot be predicted with certainty are referred to as segregating mating types versus non-segregating mating types where the outcome in the offspring is deterministic for a given mating type. The mating of 2 unaffected parents will always result in an unaffected offspring and the mating type of 2 affected parents is very rare. These 2 mating types will therefore be ignored in the analysis.

The only common mating type for which the phenotypic status of the offspring cannot be inferred based on the parental affection status is the mating between an affected parent and an unaffected parent. Assuming that the disease allele frequency is rare, the only common genotype combination for this mating type that is consistent with the parental affection status is the mating between an affected heterozygous parent (Dd) with an unaffected homozygous parent (dd). Thus we assume the parental genotypes are known. If the dominant model is true, then probability of being affected for an offspring of this mating type is 50%. For the segregating mating type $Dd \times dd$, we can construct the likelihood of the observed data, by identifying all mating types $(Dd \times dd)$ in the nuclear families of each pedigree in the sample. Let n denote the total number of offspring combined over all (DD×Dd) mating types and n_A denote the number that are affected. Since we condition on known parental genotypes, the genotypes and thus phenotypes (assuming phenotypic independence) of offspring are independent, and the log-likelihood function of the data is

$$n_A \log p_A + (n - n_A) \log(1 - p_A),$$

where the parameter p_A denotes the probability of being affected given that the parental genotypes are $Dd \times dd$. Under the null-hypothesis of a dominant model, the probability p_A will be 0.5. Under the alternative hypothesis, we can estimate p_A by

$$\hat{p}_A = n_A/n.$$

A likelihood ratio test of the null hypothesis, here dominant Mendelian inheritance versus an arbitrary transmission probability, is constructed by taking twice the natural logarithm of the ratio of the two likelihoods, one calculated under the alternative and the other under the null hypothesis. In this case, the likelihood ratio test comparing the dominant model to the unrestricted model is given by

$$2(n_A \log \hat{p}_A + (n - n_A) \log(1 - \hat{p}_A) - n_A \log(1/2) - (n - n_A) \log(1/2)),$$

which is chi-square distributed with one degree of freedom under H_0: dominant mode of inheritance.

***The Sickling Trait*:** The approach to testing for a simple Mendelian dominant mode of inheritance can be illustrated by an application to the autosomal dominant sickling trait discussed in Chapter 1. Recall that the sickling trait simply means that an individual's red blood cells can be made to sickle; these individuals do not necessarily have sickle cell disorder. Figure 4.3 shows a pedigree with the sickling trait segregating, i.e., some family members have the sickle cell trait (S) while others do not (N). Assuming that the DSL is very rare implies no DD individuals in the sample. This allows us to infer the genotype of each family member in this pedigree based on the phenotype absence or presence of sickling cells. Family members who express the sickle cell trait must have the heterozygous genotype. Normal probands without sickling cells have the common homozygous genotype. In the pedigree in Fig. 4.3, we observe 4 informative matings between a sickle-cell carrier and normal subjects. There are 23 offspring originating from these 4 matings; 11 subjects express the sickle cell trait, while 12 subjects are normal. When we compute the likelihood ratio test that is discussed above, we obtain a p-value of 0.42%, providing no evidence for rejecting the null-hypothesis of a fully penetrant dominant disease model.

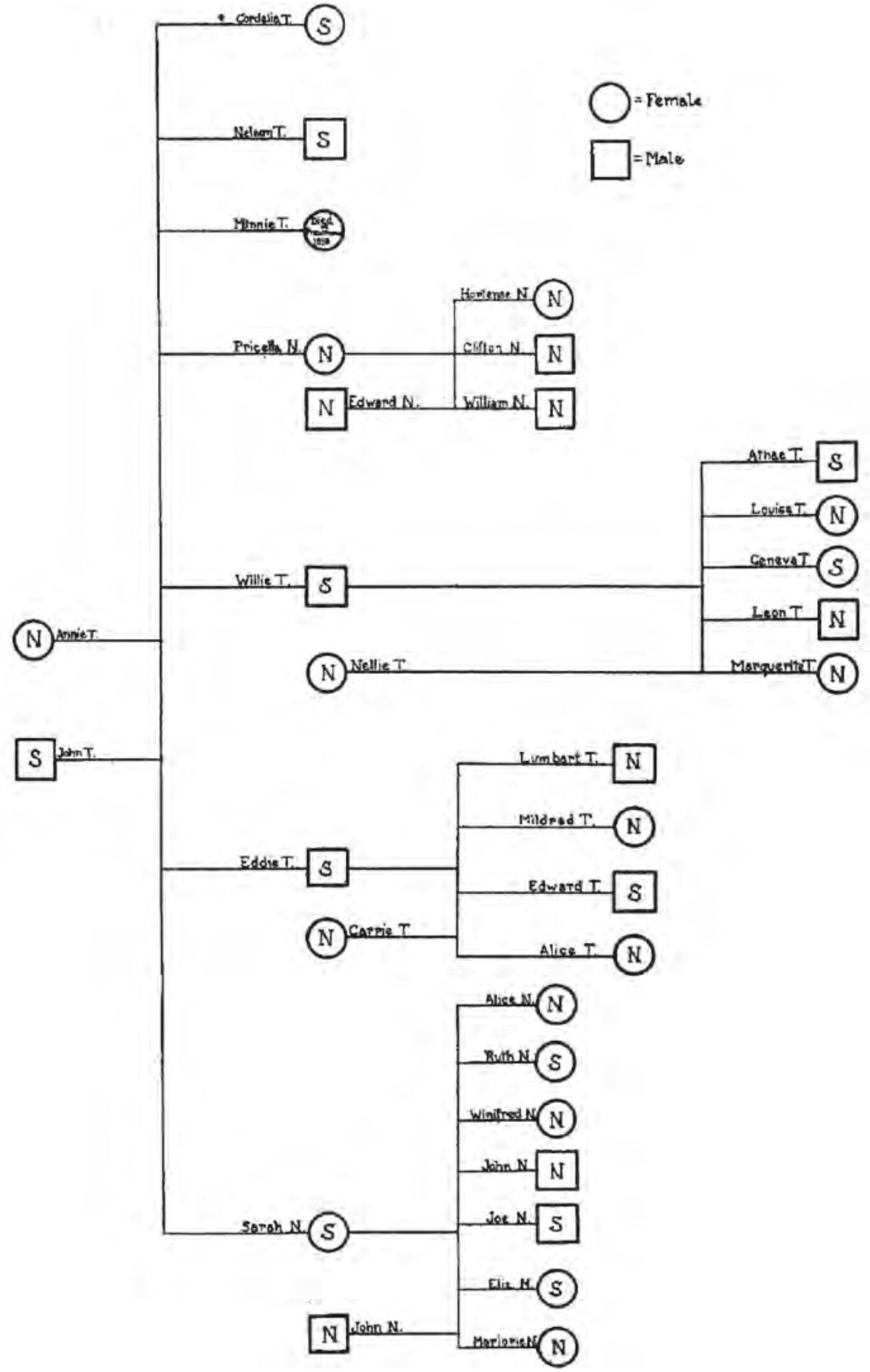


Fig. 4.3 Sickle cell pedigree. *Source* Taliaferro and Huck (1923)

4.4.2 Segregation Analysis for Recessive Mendelian Diseases

The analysis of a recessive Mendelian trait is much more difficult than the dominant, because even with the rare disease assumption, no phenocopies and full penetrance, it is not always possible to identify an unaffected individual's genotype. This inheritance mechanism is illustrated in Fig. 4.4.

Under a recessive model, assuming the absence of phenocopies, two copies of the disease allele are required in order for the subject to be affected. However an unaffected individual can either be Dd or dd. Under a recessive model, the most common segregating mating type is $Dd \times Dd$. Based on Mendelian transmission, it is easy to see that 25% of the offspring of such mating-types will be affected and 75% will be unaffected. However, the parental mating type $Dd \times Dd$ cannot

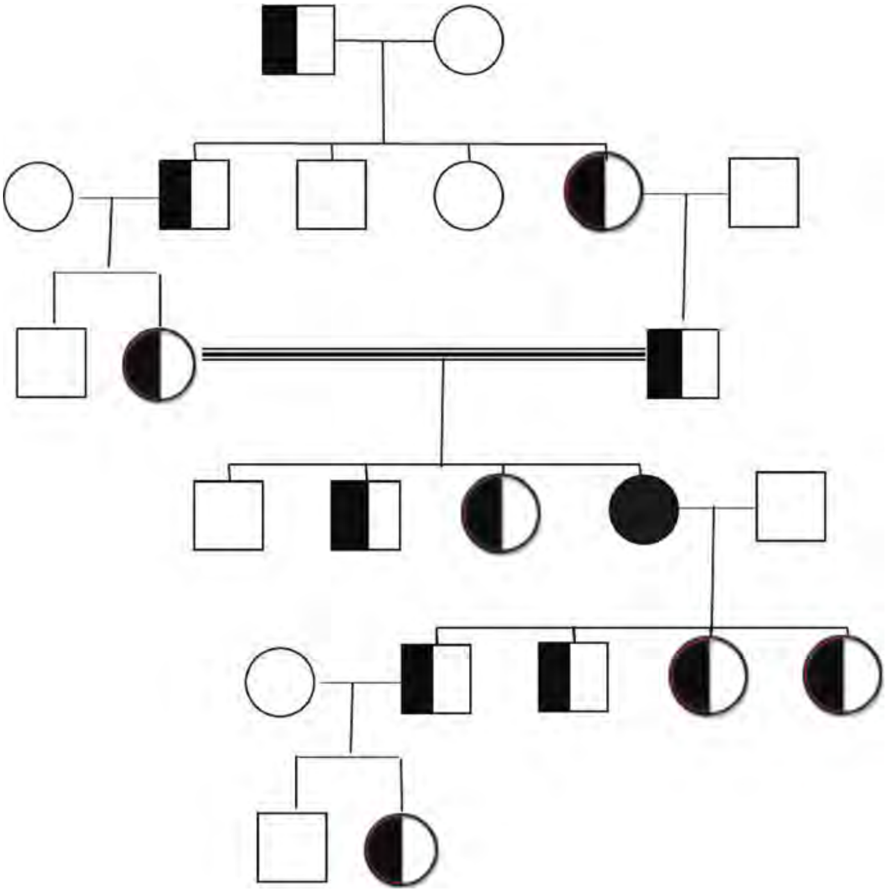


Fig. 4.4 Recessive autosomal inheritance. *Source*: Adapted from Thomas (2004)

be inferred based on the parental phenotypes. The phenotypes of this mating type are two unaffected parents; this is also consistent with the $dd \times dd$ and $dD \times dd$ mating type. Another possibility is to include the ascertainment condition into the likelihood computation. The inclusion of the ascertainment condition, i.e., at least one affected offspring, into the likelihood function requires specifying both the penetrance probabilities and the disease allele frequency. Even assuming known Mendelian penetrance functions, we must specify allele frequency and sum over possible mating types. Thus, the computation of the likelihood ratio test for the recessive model is considerably more complicated.

However, in some cases, the heterozygous genotypes can be actually observed, e.g., subjects with one copy of the disease allele develop a milder or different form of the disease, the previous analysis can still be applied. One example for this is Thalassemia, an inherited blood disorder, similar to sickle cell, which results in a reduced rate of synthesis in one of the globin chains that make up hemoglobin. There are two forms of the disease, a mild form and a severe form. Figure 4.5 (see exercise 14 in Section 4.5) contains a set of pedigrees with both forms of the disease. In this pedigree, cross hatch indicates a deceased person whose affection status is unknown, males are indicated by the Mars symbol and females are indicated by the Venus symbol. The other arrows in the pedigree denote probands. Individuals with the severe form of the disease are denoted by black circles, while the mild form is indicated by half-solid circles. This enables identification of the genotypes of all individuals in the pedigree in this special case. Exercise 14 of Section 4.5 discusses a test of the recessive model for this blood disorder.

4.4.3 Summary

As these segregation analyses show, even in the simplest Mendelian recessive model, a segregation analysis is not straightforward. When we relax the assumptions on full penetrance and phenocopies, and a single DSL, not only are there more unknown parameters to estimate, but the inability to condition on a known or inferred parental genotype makes computation of the likelihoods very complex because of the need to sum over parental and offspring genotypes. In a traditional linkage analysis, one would first carry out a segregation analysis, estimating the parameters of the best fitting model, and use them in the linkage analysis. Nonparametric versions which do not require the specification of a genetic model, avoid the necessity for a segregation analysis, and are more popular for complex diseases.

4.5 Exercises

1. Using formula (4.5) and Table 4.2 in the text, give the probability of a family having two affected sibs, assuming an autosomal recessive mode of inheritance, with disease allele frequency p, and HWE. Use this to calculate the sibling recurrence risk ratio.

Hint: most of the mating types will have a zero probability of an affected offspring and can be ignored.

2. Assuming the dichotomous disease model, with Y_1 and Y_2 being two relatives with relatedness R, show that

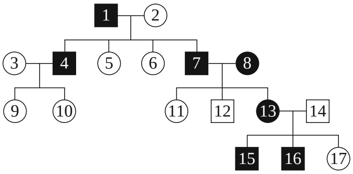
   (a) $Cov(Y_1, Y_2) = P(Y_1 = Y_2 = 1) - K^2$.
   (b) express λ_R as a function of the covariance and K.

3. Why might we expect the recurrence risk ratio in DZ twins to be the same as it is in siblings? In Table 4.1, the observed recurrence risk ratio for DZ twins is bigger than that for siblings. Any possible explanation for that?
4. Verify the expression given in the text for λ_S as a function of allele frequency and Attributable Fraction (AF) for the recessive mode of inheritance.
5. Verify the expression given in the text for λ_S when AF $= 1$.
6. Verify the first three equations in Box 4.1.
7. Assuming only one disease gene ($M = 1$), no phenotypic correlations and no dominance, show that heritability (h^2) can be estimated by the correlation of MZ twins. Hint: $X_1 = X_2$ for MZ twins.
8. If a trait measured in monozygotic twins has a correlation of 0.8, then what is the heritability (h^2) of the trait? Hint: use the result of question 7.
9. Assuming the model used in Box 4.1, give the estimated heritabilities for BMI from the four studies in Table 4.3. Do the non-zero correlations observed between spouses support the assumptions of our model?
10. From the results of homework 7 and 9, estimate the heritabilities from the Norway and Quebec data in Table 4.3 based on MZ twins. Are the heritabilities consistent with those estimated from parent/offspring pairs?
11. Based on your knowledge of Mendelian inheritance of traits, answer the following questions:

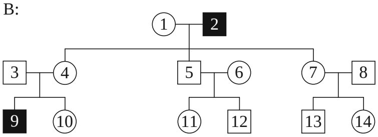
    (a) If a child has an autosomal dominant trait, what can you say about the parents?
    (b) Can autosomal dominant traits skip generations?
    (c) If two parents have an autosomal recessive trait, what can you say about their children?
    (d) Can autosomal recessive traits skip generations?
    (e) If only one of the two parents has an autosomal recessive trait, what can you say about their children?

12. The above answers to question 11 about inheritance can be used to help analyze pedigrees. For each pedigree below, tell if the trait can be autosomal dominant or autosomal recessive (a pedigree can be consistent with both modes of inheritance, or neither). If the pedigree cannot fit a mode of inheritance, tell why. In your answer, refer to specific individuals in the pedigree by number.

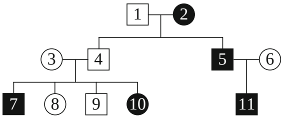
Pedigree A:



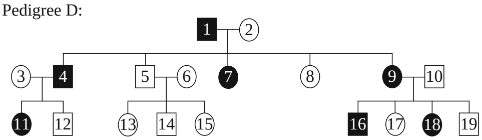
Pedigree B:



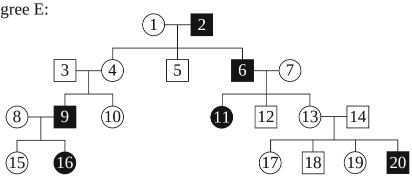
Pedigree C:



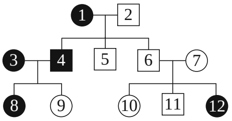
Pedigree D:



Pedigree E:



Pedigree F



13. Refer to pedigree D in question 4.12. Use segregation analysis to test if the disease is consistent with an autosomal dominant disorder. To do so,

    (a) Identify all mating types between an affected and an unaffected parent. How many are there?
    (b) Count all of the offspring from these matings. How many are affected?
    (c) Use either a Pearson goodness-of-fit chi-square or a likelihood ratio test to formally test H_0 model of inheritance is autosomal dominant (fully penetrant) versus an unrestricted model.

    Note, the sample size is really too small here. In practice you would want to combine with other pedigrees.

14. Refer to the pedigrees in Fig. 4.5 showing the segregation of mild and severe forms of thalassemia. Because the genotypes of every subject can be identified by their phenotype, all segregating mating types can be used to test if the severe form of thalassemia is an autosomal recessive, fully penetrant disorder.

    (a) Note that for the severe form, only 5 pedigrees are segregating. Which are they?
    (b) What is the mating type of the parents of all of the affected offspring in every pedigree? Why do you think there is no DD × Dd mating type?
    (c) Estimate the proportion affected from the 5 segregating mating types. What if you omitted the one segregating mating type that has no affected offspring? Would your estimate be unbiased? Explain.
    (d) Use these four families to test if thalassemia severe form is an autosomal recessive disorder (likelihood ratio or goodness-of-fit).

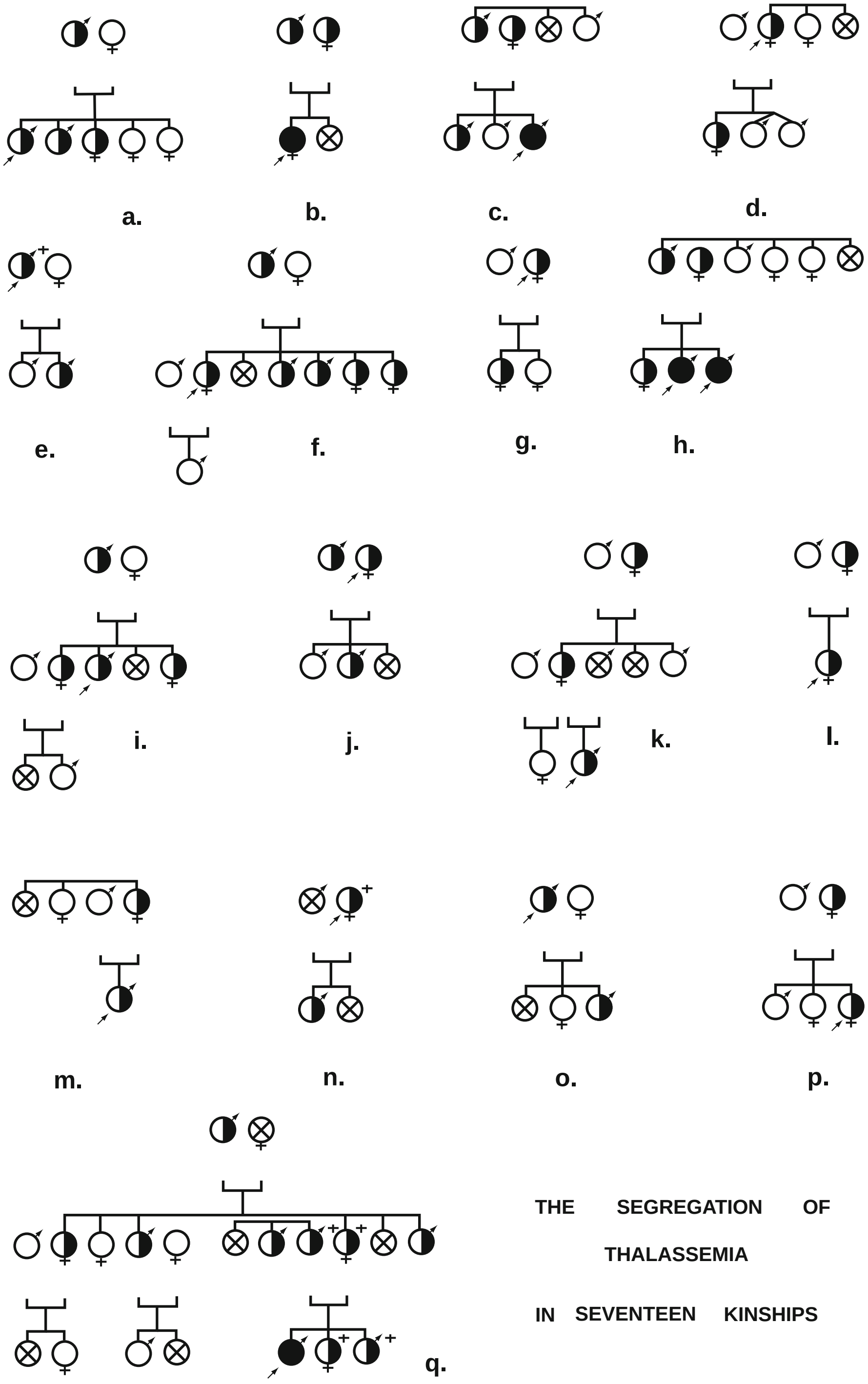


Fig. 4.5 Thalassemia pedigrees, *Source*: Neel and Valentine (1947)

Chapter 5
The General Concepts of Gene Mapping: Linkage, Association, Linkage Disequilibrium and Marker Maps

5.1 Introduction

In the absence of genetic data at the molecular level, the results of heritability, aggregation and/or segregation analysis provided the first hints about the presence of genetic effects and, consequently, the existence of a disease gene. Without information on the etiology of the disease or gene functionality, the next natural question is: ‘Where is the disease locus located in the genome?’ Although disease genes have now been located for most very rare Mendelian disorders, the search for the genomic location of disease genes for complex disorders has proven to be a difficult task.

This chapter provides an overview of basic gene mapping techniques. Gene mapping refers quite generally to the concept of localizing areas of the genome that harbor disease genes or loci, positioning genes at specific locations on the genome, but our focus here is on linkage and association mapping which are two commonly used statistical methods for finding disease genes. We begin with an overview of the concepts of linkage and association mapping, leaving the technical details of the statistical analysis to later chapters. We then formally define Linkage and Linkage Disequilibrium (LD) as well as measures of LD, followed by considerations of map distance and types of marker maps. Finally, we give a brief overview of the Human Genome and HapMap projects which have provided genetic information on markers used to facilitate mapping.

In the best case scenario, a gene mapping approach will be able to identify disease genes (DSLs) that are in the proximity of the genetic markers selected for analysis. Study subjects have to be ascertained so that the non-random relationship between the proband’s phenotype and its genetic information at the DSL extends to the neighboring genomic area where known markers are available. The mapping technique should be designed so that the area which surrounds the DSL and in which the DSL is ‘visible’ contains at least one genotyped genetic marker which will allow detection of the DSL by a suitable statistical analysis method.

With the exception of a few specialized methods, the minimal genomic distance required to permit a mapping technique to detect the presence of a nearby disease gene depends upon the relatedness or the similarity of the study subjects. For study subjects that are first or second degree relatives, the genetic signal is detectable

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_5,

with markers even relatively far away from the disease gene and hence only a small number of genetic marker loci may be needed to cover all the area of interest. This is the basis of linkage analysis. The other extreme is to recruit ‘unrelated’ study subjects. In this case, the genetic signal can be observed with marker data only in very close proximity to the disease gene. While this requires a much larger number of genetic markers to cover the same genomic region, the advantage of this approach will be a more precise estimate of the genetic location of the DSL or the disease gene. This is the basis of association mapping.

The range of detection of genetic effects is determined by recombination events. If the marker locus and the DSL are so far apart that recombination events must have happened with 50% probability in each family member, the genetic signal will not be visible at the marker locus. Besides physical distance, the other factor that influences the likelihood of recombination events between individuals in one given pedigree is their degree of relatedness. These factors can intuitively be understood by looking at Fig. 5.1, which illustrates the concept of Linkage and its relationship to Linkage Disequilibrium (LD).

Figure 5.1 shows the inheritance of chromosomes over many generations and the changes in the chromosomes due to crossovers. Note that the top three rows of the figure depict three generations of individuals who are closely related, i.e., grandparent, parents and grandchildren. The last row shows individuals who are cryptically related in the sense that they share a common ancestor. The concept of linkage is illustrated in the first three rows. In the first generation, the two chromosomes of a single grandparent are illustrated, one white chromosome that carries a common allele at the DSL and one black chromosome, with a cross denoting the

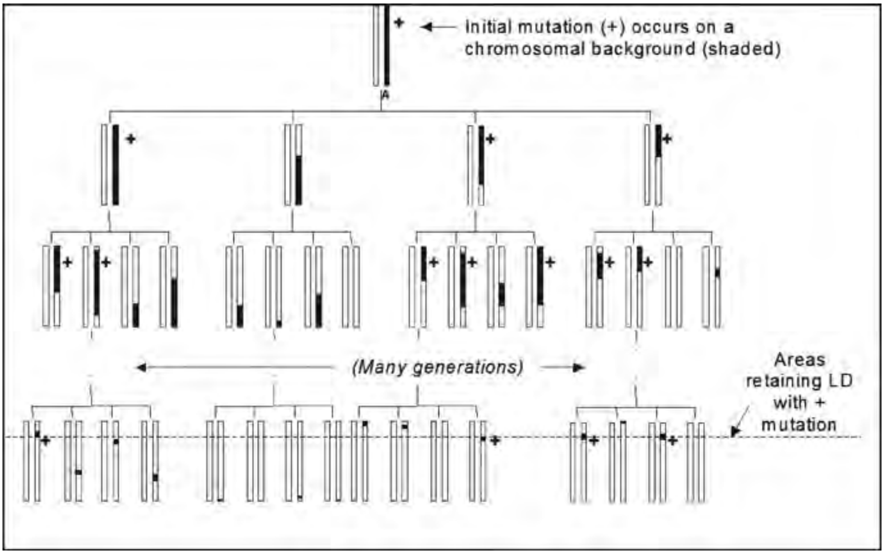

Initial mutation (+) occurs on a chromosomal background (shaded)
A
(Many generations)
Areas retaining LD with + mutation
+

Fig. 5.1 Linkage, association, and linkage disequilibrium

location where a disease variant is located. The second line illustrates the chromosomes that were created during meiotic cell division, when the chromosomes of that parent were copied and transmitted to the next generation. Due to the presence of crossovers during meiosis, parts of the two chromosomes will be exchanged and the resulting chromosomes transmitted to the second generation consist of mixtures of the two (black and white) parental chromosomes.

If we assume for simplicity, that the disease has a dominant Mendelian inheritance, i.e., one copy of the 'cross'-allele will cause the disease and that there are no phenocopies, the grandparent and all offspring in both subsequent generations that carry a 'cross'-allele from the black chromosome will be affected; those without the 'cross'-allele will not be. In this case, the DSL is said to be 'segregating in the family'. A linkage gene mapping approach can now be constructed by assessing recombination events between a known genetic marker and the DSL. The 'cross' allele is inferred from the disease phenotypes. Suppose we observe a marker at the top end of the chromosome. The figure shows no crossovers and thus no recombinations between the top of the chromosome and the 'cross' for all three affected individuals in the second generation (parents). In the third generation, we see crossing over (depending on the exact location of the marker) in only three out of seven affected individuals. In general, if the marker locus is close enough to the disease locus, we expect to see little or no crossing over or recombination between the two markers, i.e., the recombination fraction θ will be close to its minimum of zero. Now consider a marker at the bottom end of the chromosome. Three of the four offspring in the second generation are recombinants, suggesting that the probability of recombination, θ is large. If the marker locus is far away from the disease locus, the transmission of the alleles originally on the black chromosome at the marker locus will not be correlated with the inheritance of the disease and we should see the recombination fraction reach its upper limit of 50%.

Disease-mapping approaches that use the joint transmission of affection status and alleles at the marker locus to localize the disease gene are called *Linkage Analysis*. The term linkage refers to the failure of Mendel's second Law of independent assortment; or more specifically, the situation where the recombination parameter $\theta < \frac{1}{2}$ between two loci. Two loci are said to be unlinked if $\theta = \frac{1}{2}$, and correspondingly, Mendel's second law of independent assortment holds. Formally, to test for linkage, the null hypothesis is $H_0 : \theta = \frac{1}{2}$ (or equivalently, no linkage) and the alternative hypothesis is $H_0 : \theta < \frac{1}{2}$ (or equivalently, linkage is present).

With a large number of offspring in a segregating family, it is possible to obtain statistically significant results in only one family, as illustrated in Fig. 5.2. This figure shows the reconstructed transmissions in a pedigree used to demonstrate linkage of early onset AD to the APP gene. Seven genetic markers, including one in the APP gene which is not the DSL, were genotyped in 13 offspring of two brothers with early onset AD. The absence of any recombination between the four markers shown in Fig. 5.2 in the genetic material transmitted from the two brothers to their six affected offspring suggests that all of the four loci and the DSL are closely linked, but it does not permit pinpointing the location of the DSL (now known to be in the APP gene). However, the two recombinations between the two markers D21S1

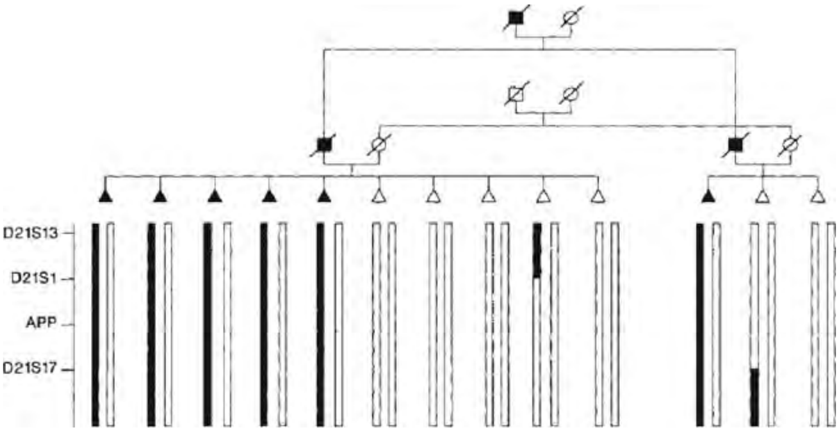


Fig. 5.2 Pedigree in which early-onset AD is apparently inherited as an autosomal dominant disorder. Females are denoted by circles and males by squares; a slash indicates deceased individuals with no genotype data. Triangles used in the last generation preserve anonymity. In generation 2, the spouses of the affected brothers are sisters. Samples were available from the 13 individuals whose chromosomes are illustrated, from a further 19 children and spouses of these individuals and from 7 more distantly related unaffected individuals. Beneath the pedigree are ideograms of the pair of chromosomes for each individual of the third generation, at four loci on the long arm of the chromosome. The linkage data suggest that the black portions of the chromosomes were inherited from the affected fathers. *Source*: Goate et al. (1991)

and D21S17 (whose locations were known) and the DSL in two of the unaffected offspring strongly implicates the region between D21S1 and D21S17 as the site of the DSL. This illustrates that it can be easy to detect linkage, but difficult to be precise about the location of the DSL.

Linkage analysis relies on looking at transmissions of genetic material from parent to offspring over 1–2 generations. In this time frame, only a few recombinations can occur between linked loci. And only a few markers are needed to cover a large region. This is one of the key advantages of linkage analysis and one of the reasons for its popularity during the ‘early’ days of gene mapping, when it was only possible to genotype a handful of marker loci across the human genome. At that time, genotyping more than 20–40 marker loci per chromosome would have been very costly and practically not feasible for most genotyping laboratories. At the height of its popularity, linkage scans of the entire genome used only 400–800 markers.

In samples of unrelated subjects, the ‘genetic signal’ of the disease gene/DSL has a much shorter range in which it can be detected at marker loci, if the genetic variant is old, i.e., it occurred for the first time many, many generations ago. Figure 5.1 can also be used to illustrate this property as well. Two ‘unrelated’, affected subjects whose disease is triggered by the same genetic variant may have had a common ancestor many generations ago in whom the disease mutation initially occurred, making them cryptically related. In Fig. 5.1, the white and black chromosome in the first line can now be considered as the pair of chromosomes from the common ancestor in whom the disease variant (cross allele) occurred for the first time. The

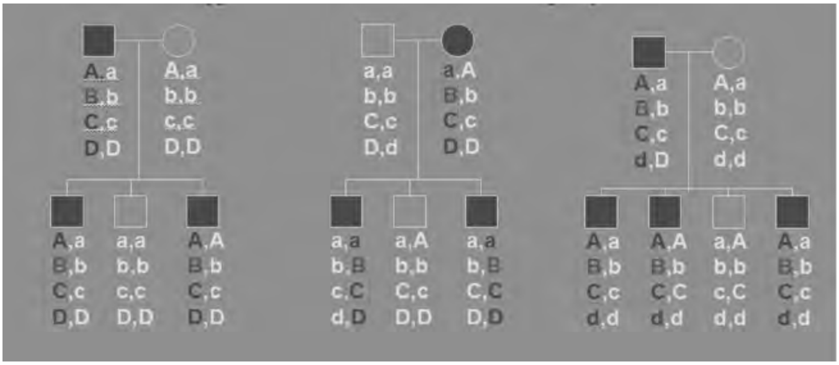


Fig. 5.3 Distinguishing linkage from association. *Source*: Courtesy of Professor Lyle Palmer

last line in Fig. 5.1 shows a population of ‘unrelated’ study subjects whose disease chromosomes originated from the common ancestor with the original disease mutation. Due to the numerous recombination events that took place during the meiotic cell divisions between the generations (middle part of Fig. 5.1), the original chromosomes of the common ancestor have been divided many times and the majority of its parts have been replaced by other copies of the same chromosomal segment. As a consequence, the black areas around the original disease mutation that have remained unchanged are much smaller now, naturally reducing the range in which the genetic signal can be detected. The genetic marker loci have now to be very close to the disease mutation in order to identify the disease gene. However, any markers in the black regions surrounding the ‘cross’ allele at the bottom of the Fig. 5.1 have the important property that they share the same ancestral allele. That is, each diseased person shares the ancestral disease allele at the ‘cross’ location, and they also have the same allele at any marker in the black area surrounding the ‘cross’ allele. In other words, two particular alleles, one from each locus, tend to appear together on the same haplotype in a population. This concept is illustrated in Fig. 5.3 which shows three markers with alleles A,a and C,c and D,d, and a DSL with disease allele B transmitted in three different families. All of the three markers are linked to the disease locus because the same alleles are being transmitted within families at all four locations. However, only the ‘C’ allele is transmitted with the ‘B’ allele in every family. Thus the A and D loci are linked, but only the ‘B’ and ‘C’ alleles will be associated with each other at the population level.

***Constructing a Mapping Approach using Unrelated Individuals*:** Since genetic data on other family members is not available in this context, it is not possible to assess whether the disease locus and a marker locus are physically linked, i.e., the transmissions of alleles at the two genetic loci occur together, or independently. A different mapping concept has to be employed here. Rather than examining recombinations, one relies on the concept of indirect association. The genotype at the DSL and the phenotype of the study subjects will be associated, possibly following one of the genetic models discussed in Chapter 2. The genetic association will also be visible between the marker and the phenotype if a particular allele at the genetic marker tends to appear together on the same gamete with the disease allele at the disease locus. This latter concept, the association of alleles at two loci, is referred to as *linkage disequilibrium (LD)*. The term *allelic association* is also sometimes used to denote the association of alleles at two loci. The absence of association between two genetic loci is referred to as *linkage equilibrium (LE)*. If the marker locus and the DSL are in LE, the phenotype will not be associated with the genotype at the marker locus and, consequently, the 'genetic signal' is not detectible at the marker locus. An observed association between a genetic marker locus and the phenotype of interest suggests the existence of a DSL which is in LD with the marker. This is the basis of *association mapping*. As depicted in Fig. 5.3, in linkage analysis we are only concerned with alleles at two loci being transmitted together from parents to offspring; the particular allele at the marker which is transmitted with the disease variant is irrelevant, and in general will be different for different families. In association mapping, our interest is in the specific alleles associated with disease; except for selected family designs, information on the transmission of alleles from parents to offspring is not used in an association analysis. If the alleles at marker and the disease locus are in LD, an association between phenotype and marker locus can be observed. The marker locus can be used to test for association, but the particular allele associated with the disease may not have any meaningful biological interpretations.

5.2 Genetic Markers and Marker Maps

Regardless of which approach we will select for our analysis, linkage or association, we cannot be successful without marker loci that are sufficiently close to the disease gene. Sets of genetic markers that cover the entire human genome are therefore required. In the very early years of gene mapping, marker loci were usually phenotypes such as blood groups which followed simple Mendelian inheritance patterns. However, advances in technology led to the development of tools to obtain direct information on DNA sequences at the molecular level which could be positioned on a map of the genome. Today the term maker locus is generally taken to refer to data on DNA at a specific location in the genome. The creation of marker maps showing the location of genetic markers was non-trivial. Even until the 1980s, it was only possible to genotype a small number of genetic locations and to identify their locations on the human genome, i.e., to place a marker locus on a particular chromosome, or, even more challenging, to identify its position on the chromosome. The difficulties of mapping a marker locus to a specific location on the human genome originate directly from the genotyping process itself, in which enzymes break down the double-helix structure of each chromosome into much smaller parts which are then selectively amplified in order to make the genetic information accessible.

However, one side effect of this process is that, while the genetic information contained in a small DNA fragment can be identified (so that genotypes can be constructed), the knowledge about its location in the human genome is lost during the genotyping process itself, and must be recovered by a mapping process. As we discuss in the next chapter, the principles of linkage analysis were utilized to identify which markers were in close proximity/linkage with each other and to place them in the proper order on the chromosome. Only with the completion of the human genome project, which provided the sequence for the entire genome, was it finally possible to position markers more precisely to their exact location on a map of the chromosome without having to rely on linkage analysis.

Gene mapping requires a notion of distance between two genetic loci, as well as a map which can show the location of known genetic loci. The two most commonly used measures of distance and their associated maps are genetic (or linkage) and physical. A *linkage map* shows the relative positions of genetic loci on a chromosome as determined by the recombination fraction between them. The unit of distance in a linkage map is based on *Morgans or centimorgans (cM)*, which measures the expected number of crossovers between two loci per gamete. Although originally proposed by Morgan as a measure of distance, the recombination fraction itself is not a particularly useful measure of distance because its maximum is only $\frac{1}{2}$. Further, it has little information about the number of crossovers between two loci unless it is the distance is very small and there is likely only zero or one crossover.

Haldane devised a map function to transform θ into a distance measure by using the assumption that the number of crossovers follows a Poisson distribution with mean 2L. Recall that there are four gametes produced from two pair of sister chromatids. One crossover between two non-sister chromatids produces two gametes with a crossover, and two without, hence distance, say L, measured in expected number of crossovers per gamete is

$$L = E(\#\,\text{crossovers})/2$$

or

$$E(\#\,\text{crossovers}) = 2L.$$

Thus, letting X denote the number of crossovers,

$$P(X = k) = \frac{e^{-2L}(2L)^k}{k!}$$

for k=0, 1, ... Hence

$$P(X = 0) = e^{-2L}.$$

However, we saw in Chapter 2 that

$$\theta = (1 - P_0)/2,$$

where P_0 was defined as $P(X = 0)$. Substituting e^{-2L} for P_0 and solving for L, we find that Haldane's measure of distance is given by

$$L = -[ln(1 - 2\theta)]/2.$$

Here L is measured in Morgans, so if $\theta = 0.01$, then L $=$ 0.0101M or approximately 1 centimorgan (cM). Note that L is approximately equal to θ, when θ is small. For small values of θ, Morgan suggested simply using θ to measure distance in Morgans, i.e., L $= \theta$M, so that distance in centimorgans is L $= 100\theta$cM. Despite the fact the distribution of the number of crossovers in not well captured by a Poisson, because of interference and a varying probability of crossover over the genome, the Haldane map has many good features. For example, a measure of distance should be zero when θ is zero, but essentially infinite when θ is $\frac{1}{2}$. In addition, the distance from A to C should be the sum of the distances from A to B and B to C if marker order is A, B, C (See exercise 10 of Section 5.7). Many mapping functions with different properties have been suggested for transforming the recombination fraction into Morgans or centimorgans (e.g., Morgan, Haldane, Kosambi, etc.). Most of them give very similar distances in cM for small values of θ (say $<$ 0.1), but differ for larger values. A very important feature of linkage maps is that they are all monotone in θ. Thus as θ increases, so does the distance in cM. This point has relevance for distinctions between linkage mapping and mapping based on LD.

A *physical map* gives the locations of identifiable landmarks on DNA (e.g., chromosomal bands, restriction-enzyme cutting sites, genes, etc.). For the human genome, the lowest-resolution physical map (apart from the long and short arms of the chromosomes themselves) is the banding patterns on the 24 different chromosomes; the highest-resolution map is the complete nucleotide sequence of the chromosomes. When loci are close, we usually measure their distance in base pairs. There are approximately 1,000,000 base pairs (bp) in one centimorgan, cM. Table 5.1 shows the estimated lengths of the chromosomes in both cM and Kb. There is generally good agreement with the rule of 1 cM = 10^6 bp. Note that the expected number of crossovers ranges from about 3 on chromosome 1 to $\frac{1}{2}$ on chromosome 21.

Table 5.1 Approximate lengths of human chromosomes measured in cM and in Mb. *Source*: Yang (2000)

Chromosome #	1	2	3	4	5	6	7	8
Length (Mb)	236	255	214	203	194	183	171	155
Length (cM)	293	277	233	212	198	201	184	166
Chromosome #	9	10	11	12	13	14	15	16
Length (Mb)	145	144	144	143	114	109	106	98
Length (cM)	167	182	156	169	118	129	110	131
Chromosome #	17	18	19	20	21	22	X	Y
Length (Mb)	92	85	67	72	50	56	164	59
Length (cM)	129	124	110	97	60	58	198	–
Total (with Y)								
Length (Mb)	3200							
Length (cM)	3702							

5.3 Testing for Linkage or Association: Basic Concepts

Linkage analysis operates on the principle of recombination. Formally, we test the null hypothesis $H_0 : \theta = \frac{1}{2}$ where θ equals $\frac{1}{2}$ implies no linkage. Linkage predated association mapping because relatively few markers are required to cover long distances and the types of markers which are especially suited for linkage analysis (microsatellite) were available much earlier than SNPs, which are well suited to association analysis. For example, the entire distance of chromosome 21 is small enough that any two loci, are linked; using the Haldane map, loci 50 cM apart have a recombination fraction of about $0.316 < \frac{1}{2}$. Thus with sufficient sample size, in principle one should be able to detect linkage to a disease locus anywhere on chromosome 21 with only 1 marker, although clearly, the closer θ is to zero, the higher the power to detect linkage, and the region of uncertainty about the exact location will be very large with only one marker.

Early linkage studies typically tested only a few markers, and evidence for linkage was assessed at each locus separately, as we describe in the following chapter on linkage. As the number of possible markers increased, the concept of multi-point linkage analysis and whole genome linkage scans became popular. In a multi-point linkage analysis, several markers which are linked to each other are analyzed jointly to assess evidence for linkage in the region. This enables one to 'test' each locus in the entire region, even those not typed, as a possible DSL for linkage to the hypothesized DSL. By analyzing markers jointly, one infers linkage at unmeasured loci between the markers to better locate the DSL. A whole genome linkage scan is designed to locate a hypothetical DSL anywhere in the entire genome. A standard whole genome linkage scan requires only 400–800 markers, placed 5 to 10 cM apart. As will be discussed in Section 6.3, by using multiple linked markers all across the genome, it is possible to construct a statistic from these markers that tests globally for linkage between a DSL and any locus on the entire genome. This is a powerful concept, especially because the *α-value* of the test does not depend greatly on the number of the markers. This feature is a result of being able to characterize the joint distribution of the estimated recombination fractions between loci under the null hypothesis and an important advantage of linkage studies over association studies. Since LD extends only for a short genomic distance, a genome scan covering a region requires many more markers and thus association tests than is required for a linkage scan. This means that the burden of multiple testing is much higher for association studies than for linkage studies. A general consequence of this property is that linkage studies require smaller sample sizes than association studies.

The use of only a few markers is both a strength and weakness of linkage studies because linkage studies tend to have low resolution, i.e., they do not give very precise information about where the DSL lies. To see why, consider a recombination parameter of 5% between a disease locus and a marker; this is a relatively small recombination parameter corresponding to a distance of about 5cM, and with a large enough sample it is easy to reject the null hypothesis that $\theta = \frac{1}{2}$. It is much more difficult to find a locus where θ equals zero, which is what we need to pinpoint the location of the DSL. A recombination parameter of 5% translates to about 5,000

Kb. Sequencing 5,000 Kb of DNA to find a variant correlated with disease status is currently expensive and time consuming enterprize. Thus the genetic distance ($\theta = 0.05$) is small, but the physical distance is large. In a typical linkage study, 3–4 markers spanning 20 million base pairs might show evidence of linkage to the DSL. When a large linkage region like this is identified, a typical strategy is to then to cover the entire linked region with SNPs and use association analysis for 'fine mapping'.

In contrast to linkage, the underlying principle of association mapping is LD. Association analysis used in genetic mapping does not differ substantially from any other statistical analysis of association. The goal is to demonstrate a relationship between two variables, in this case, a marker at a genetic locus and a phenotype or trait of interest. In many cases, the basic statistical designs, case-control, case-cohort and population based studies are the same and the methods of analysis are the same, e.g., chi-square tests, ANOVA, and regression analysis (linear, logistic or proportional hazards regression). The main distinguishing feature of a genetic association analysis is that the association works indirectly using LD between the marker and the DSL. If there is LD between the marker and the DSL, we can expect to see a relationship between the marker and the disease. If the marker and the DSL are independent, then we should see no association between the marker and the disease. Thus with an association analysis, the null hypothesis can be framed as H_0: 'no LD between the marker and the DSL'. Figure 5.4 illustrates this concept of 'guilt by association'.

As we show in the next section, the relationship between disease phenotype and the DSL is stronger than the relationship between the disease phenotype and the marker, depending upon the degree of LD. Since we typically will not know the LD pattern between the unknown DSL and the marker, estimates of 'effect' of the marker are distorted versions of the effect of the DSL on the trait, and the focus of a genetic association study is generally on testing and not estimation.

Martin et al. (2000) genotyped sixty SNPs in a 1.5-Mb region surrounding APOE, in 220 cases with Alzheimer's Disease and 220 controls without Alzheimer's

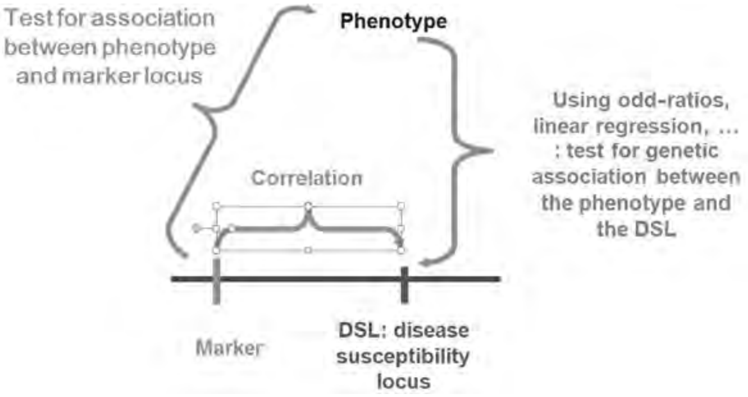

Test for association between phenotype and marker locus
Phenotype
Using odd-ratios, linear regression, … : test for genetic association between the phenotype and the DSL
Correlation
Marker
DSL: disease susceptibility locus

Fig. 5.4 Indirect association: Guilt by association

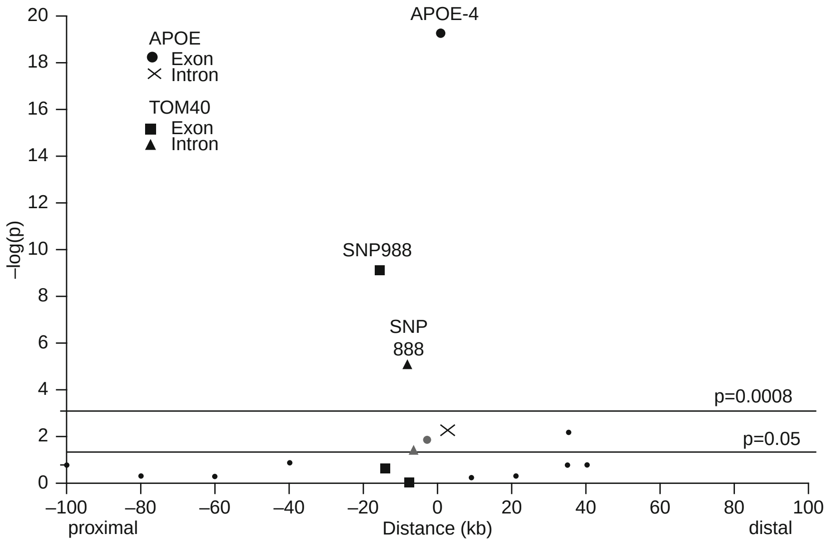


Fig. 5.5 Plot of minus log of P value for testing allelic association between each marker with AD, for SNPs immediately surrounding APOE ($<$100 kb.) *Source*: Martin et al. (2000)

Disease. Standard tests were conducted to look for association of SNP alleles with AD, in cases and controls. Some evidence of association ($p<0.05$) was identified for 7 of 13 SNPs, including the APOE-4 polymorphism, spanning 40 kb on either side of APOE. Figure 5.5 shows a graph of minus log of p-value for testing allelic association between each marker with AD, for SNPs immediately surrounding APOE ($<$ 100 kb). Although there are several SNPs close to APOE with a signal, it is clear that magnitude of the p-value depends on several factors besides physical distance to the DSL.

5.4 A Formal Definition of Linkage Disequilibrium and Related Measures Used to Describe Linkage Disequilibrium

Here we define LD more formally and discuss some measures that are commonly used to quantify LD. We begin with a definition of linkage equilibrium (LE), which characterizes loci on non-homologous chromosomes, or on the same chromosome, but unlinked.

Let the alleles at two markers be denoted A,a and B,b. Let the allele frequencies at each marker be p_A, p_a, p_B, p_b and let p_{AB}, p_{Ab}, p_{aB}, p_{ab} denote the frequencies of the four possible haplotypes. Thus p_{AB} denotes the frequency of a randomly selected haplotype from the population with alleles A and B observed at the two loci, etc. LE implies that the haplotype frequencies are given by the product of the corresponding allele frequencies. The resulting frequencies are given in Table 5.2.

Table 5.2 Population allele frequencies between two loci under linkage equilibrium

	B Locus		
A Locus	B	b	Total
A	$p_{AB} = p_A p_B$	$p_{Ab} = p_A p_b$	p_A
a	$p_{aB} = p_a p_B$	$p_{ab} = p_a p_b$	p_a
Column Total	p_B	p_b	

With LE, the frequency of the haplotype is the product of the allele frequencies at the corresponding loci. Thus LE corresponds to our usual notion of independence in a 2 × 2 table. The haplotype frequency is just the joint probability of A and B being observed on the same haplotype, etc., and the allele frequencies are the marginal frequencies.

When LE fails, the number of alleles at two loci is not the product of the individual allele frequencies. The *LD coefficient*, is usually denoted by D in the literature; it measures the departure from independence. (We will occasionally use the notation δ to distinguish it from the disease allele D.) It is defined as

$$D = p_{AB} - p_A p_B$$

Note that $D = 0$ corresponds to independence in the 2 × 2 table. Since the labeling of alleles is arbitrary, we could just as easily define D in terms of Ab or ab, etc. With only two alleles, any definition gives the same absolute value. This is the same principle as having only 1 degree of freedom in a 2 × 2 table with fixed margins. Table 5.3 uses this expression to define all of the cell probabilities of the 2 × 2 table in terms of D, and the allele frequencies, $p_A,\ p_a,\ p_B,\ p_b$.

A substantial difficulty with using D to measure the lack of independence in the 2 × 2 table shown in Table 5.3 is that D is highly sensitive to the marginal values. Each of the four haplotype frequencies defined in Table 5.3 must be ≥ 0, hence in order to keep p_{AB} and p_{ab} positive when D is negative, we must have

$$D \geq -\min(p_A p_B,\ p_a p_b),$$

and when D is positive, we have

$$D \leq \min(p_A p_b,\ p_a p_B).$$

Table 5.3 Cross classification of haplotype frequencies when independence does not hold

		B Locus		
	A Locus	B	b	Row Total
	A	$p_{AB} = p_A p_B + D$	$p_{Ab} = p_A p_b - D$	p_A
	a	$p_{aB} = p_a p_B - D$	$p_{ab} = p_a p_b + D$	p_a
Column Total		p_B	p_b	

Thus we can define minimum and maximum values for D as:

$$D_{\min} = -\min(p_A p_B, p_a p_b),$$

and

$$D_{\max} = \min(p_A p_b, p_a p_B).$$

If one of the allele frequencies is low, the corresponding haplotype frequency will also be low, as will D. Thus a value of D close to zero may simply reflect low marginal frequencies, and not ‘independence’, as a value of $D = 0$ would suggest.

There have been many proposals for alternative measures of LD which are less sensitive to the marginal frequencies. We will primarily use r, the correlation coefficient, because of its important role in association testing, but we also discuss D', because of its prominent role in statistical genetics. The basic idea behind D' to normalize D by the maximum value (or minimum) to give a fixed range between 0 and 1, regardless of the margins. For D positive, D' is defined as

$$D' = D/D\text{max},$$

and for D negative, we take

$$D' = D/D_{\min}.$$

Thus D' ranges from 0 (at LE) to a maximum of 1, which happens if any cell equals zero.

Despite this improvement in interpretability over D, a high value of D' does not mean that one locus can predict the other with high accuracy. For this, we need the correlation between the two, r, or its square, r^2. The correlation coefficient between the two loci is constructed by assigning a numeric value to each allele at each locus, typically 0 and 1, and then computing the ordinary correlation coefficient. Using this approach, we have that

$$r = (p_{AB} - p_A p_B)/\sqrt{p_A p_B p_a p_b} = D/\sqrt{p_A p_B p_a p_b}.$$

Like D', r^2 is zero if and only if D equals zero, but now $r^2 = 1$ only if a pair of diagonal cells equals zero, i.e., either $p_{AB} = p_{ab} = 0$ or $p_{Ab} = p_{aB} = 0$. In this case, $p_A = p_B$ and $p_a = p_b$. An r^2 of 1 implies perfect predictability; if we know the allele at locus one, we can predict perfectly the allele at locus 2, and vice-versa. This will be important in choosing markers for association studies. Table 5.4 shows a comparison of the measures of LD on a fictitious sample of 100 chromosomes. Note that $D_{\max}$ can be calculated by adding ±2 from each of the cells of the table to give counts of 45, 25, 0, 30. This maintains the margins, but maximizes the value of $D > 0$ over all tables that have these margins.

We now provide a simple derivation to show how LD between a marker and a DSL will induce association between the phenotype and the marker. Consider a case-control study with equal numbers of cases and controls. Let P(A|case) and P(A|control) be the frequency of the disease allele A among the cases and

Table 5.4 Measures of linkage disequilibrium

		B Locus		
	A Locus	B	b	Row Total
	A	43	27	70
	a	2	28	30
Column Total		45	55	100

$$D = (43 - 70 * 45/100)/100 = 0.115$$
$$D_{\max} = \min(70 * 55, 30 * 45)/10,000 = 0.135$$
$$D' = 0.115/0.135 = 0.8581$$
$$r = 0.115/\sqrt{0.7 * 0.3 * 0.55 * 0.45} = .5044$$

controls, respectively, and let a denote the non-disease alleles. As we will discuss in some detail in Chapter 6, a test of association between the DSL and the disease can be framed as no difference in allele frequency among cases and controls, or $H_0 : \Delta_A = 0$, where

$$\Delta_A = \text{P(A|case)} - \text{P(A|control)}.$$

Suppose we do not observe the disease locus, but instead a marker with alleles B and b. Then defining Δ_B as

$$\Delta_B = \text{P(B|case)} - \text{P(B|control)},$$

and assuming that p(disease) does not depend on marker genotype given the genotype at the DSL, we have (Pritchard and Przeworski, 2001)

$$\Delta_B = \Delta_A(\text{P(B|A)} - \text{P(B|a)}). \tag{5.1}$$

Note that in the absence of LD, the alleles at the DSL and the marker are independent,

$$\text{P(B|A)} = \text{P(B|a)= P(B)}$$

hence $\Delta_B = 0$. Thus there will be no association between disease and a marker allele, unless the marker allele is associated with the disease allele.

It is straightforward to show that

$$\text{P(B|A)} - \text{P(B|a)} = \sqrt{p_B p_b}\, r \tag{5.2}$$

where r is the allelic correlation between the two loci, and p_B and p_b denote allele frequencies at the marker. Hence we have

$$\Delta_B = \Delta_A \sqrt{p_B p_b}\, r$$

which implies that

$$\Delta_B^2 < \Delta_A^2.$$

Although the deviation from the null is smaller when using the marker rather than the DSL, the effect on power of the test will depend on allele frequencies at the two loci as well as the effect size. From equation (5.1), and the illustration in Fig. 5.5, if there is no LD ($r = 0$) between the DSL and the marker, then we do not expect to find any association between the disease phenotype and the marker. However, LD is a short range concept. Unlike linkage, we do not expect LD to persist for long distances between two loci. We will make this concept more precise in the next section, but suffice it to say here that a typical candidate gene study might require at least 20 markers, fine mapping a linkage region might require thousands of markers and a whole genome association study 500 K to a million. Rejection of the null hypothesis of no association suggests that the DSL is 'physically close' to the marker. A natural question about association mapping is how close does the marker need to be? This requires a consideration of the origin and maintenance of LD in populations.

***What are the Similarities and Differences Between Linkage, LD and HWD?*:** As the name would imply, linkage disequilibrium is somewhat related to the concept of linkage, and it is also somewhat related to Hardy-Weinberg disequilibrium, but all three terms measure distinctly different concepts. Linkage is a physical concept, describing the physical distance between two loci in terms of recombination events. The minimum value is zero, implying that no recombination ever occurs between the loci; the maximum value is $\frac{1}{2}$ which can be interpreted as the two loci being on non-homologous chromosomes. The recombination fraction does not depend upon an individual's ethnic origin. In contrast, LD is a population concept; it concerns the population probability that alleles at two different loci appear together on the same parental haplotype. In this regard, LD is similar to HWD which concerns the population probability that two alleles at the same locus appear together in an individual's genotype. Like HWD, LD can arise in a population for many reasons, including mutation and close linkage, as well as population substructure. The reason for the term LD has to do with how LD arises, and how it is maintained in a population as was illustrated in Fig. 5.1.

5.5 The Origin and Extent of LD in the Human Genome

LD can arise in a population for many reasons, including mutation, close linkage as discussed in Section 5.1 of this chapter, and population substructure considerations, as discussed in Chapter 3. LD arising due to mutation and close linkage is useful for association mapping. The reason for the term LD has to do with how LD arises, and how it is maintained in a population. As shown in Fig. 5.1, the LD around the mutation was maintained only in a small neighborhood of the locus because linkage over many generations caused the initial association to dissipate. Denoting the LD between a marker and the mutation in Fig. 5.1 by D_0 at the first generation, and D_t after t generations, an approximate formula relating D_t to D_0 and θ is given by

$$D_t = D_0(1-\theta)^t. \tag{5.3}$$

(See exercise 12 of Section 5.7). In the extreme, when $\theta = 0$, the LD does not dissipate over time. When θ is $\frac{1}{2}$, LD dissipates rapidly relative to small values of θ.

Formula (5.3) makes all of the same assumptions that are required for HWE, i.e., random mating, no selection, no mutation, in or out migration and constant allele frequencies. However in practice, LD is affected by many features of the population, including selection, changing allele frequencies and migration and formula (5.3) is not accurate for small values of θ or large values of t. Empirical studies of the relationship between LD and distance show that LD does not decline smoothly with genetic distance. In addition, for association mapping, we are interested in LD in a relative small region which is best measured in base pairs. See Fig. 5.6 which shows the estimated correlation versus physical distance for pairs of markers in five regions of the chromosome, for three different samples. Many estimates, some derived from population genetics models, have been given for how far the range of useful LD (meaning the signal from DSL to marker is still strong enough for detection) extends in terms of base pairs; they range between 50 and 300 Kb. See Kruglyak (2008) for a review. However, as numerous empirical studies have shown, the relationship between distance and LD is not a smooth one.

Although many more markers are used in association mapping than are used in linkage, methods for association analysis which can test for a DSL at every location between two markers are limited by the unpredictable nature of LD from marker to marker. Knowing LD between two loci A and C implies nothing about the LD between A and B and B and C, where B is a marker in between A and C. This situation can be partially alleviated by using external information on LD between a set of typed markers and a set of untyped SNPs whose frequencies are also known (see the HapMap project below) to ‘impute’ marker values at the untyped locations. These ‘imputed’ SNPs can also be tested for association to improve LD coverage in a region (Howie et al. 2009). In order for imputation algorithm to work, one has to assume that the LD-structures in the reference population and the study population are very similar. In any event, in large scale association studies, markers are usually tested separately, leading to issues with multiple comparisons. This will be discussed in Chapter 10.

5.6 The Human Genome and HapMap Projects

The Human Genome and the HapMap projects have been instrumental in providing information on where genes are located in the genome and in providing sets of markers to use in gene mapping studies. The Human Genome Project was a 13 year, multi-national project completed in 2003; the primary objective of the project was to identify all of the genes in the human genome and identify the sequence of all 3 billion DNA base pairs (without identifying the different variants at each base pair). As part of the project, many new technologies were developed for probing

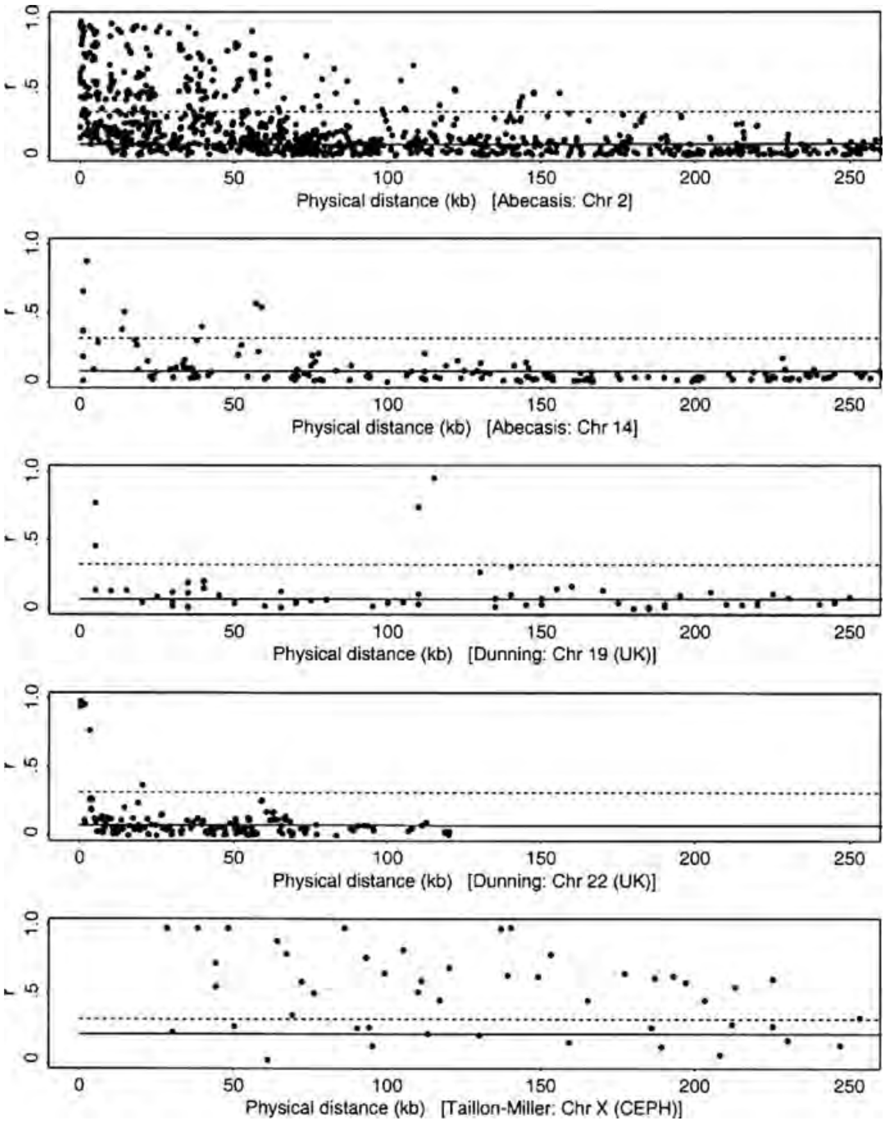


Fig. 5.6 LD (Measured as correlation) versus distance in base pairs *Source*: Pritchard and Przeworski (2001)

the genome. Both genetic maps and physical maps were constructed to aid in gene mapping via linkage and association analysis, including 3,000 microsatellite markers spaced one cM apart used for the linkage maps, and 52,000 Sequence Tag Sites (a short sequence of DNA (200–500 bp) which identifies a unique location in the genome) used for the physical map.

The International HapMap Project was a multinational project begun just as the Human Genome project was ending. One objective of the HapMap project was to provide data that can be used to estimate LD between pairs of loci. Because patterns of LD are specific to populations, the DNA samples used for the initial phases of the HapMap project came from four main populations: 30 parent-offspring trios from the Yoruba people in Ibadan, Nigeria, 30 CEPH trios (largely Western European background), 45 unrelated Japanese from Tokyo and 45 unrelated Han Chinese from Beijing. Additional samples from other populations have been added.

SNPs were selected to create the HapMap for several reasons: there are vast numbers of SNPs; they are located everywhere throughout the genome; and they are relatively easy to genotype. Data on around 10,000,000 SNPs are now available through the HapMap. The basic data produced by the Project are the genotypes of the SNPs of the 270 individual samples and the frequencies of SNP alleles and genotypes in each population. These data are freely available to researchers, as are standard measures of LD, such as D, D' and r^2. In addition, the HapMap describes the common patterns of genetic variation in humans, including the chromosome regions with sets of SNPs in high LD, the haplotypes in those regions, and provides the technology for choosing SNPs in a gene mapping study. It also notes the chromosomal regions where associations among SNPs are weak, suggesting recombination hotspots, areas where a large number of recombinations have led to weak LD between loci on either side of the hotspots.

Figure 5.7 (end of Section 5.7) output from Haploview illustrating the LD structure for 22 SNPs in the APOE gene. The number in a box indicates the D' for the two SNPs which connect the boxes. The shading, from light to dark, indicates the magnitude of D', darkest having the highest D''s. The blocks around sets of SNPs represents sets of adjacent SNPs having relatively high LD. Note that physical proximity tends to go with higher D', but not always. There are not a large number of SNPs in high LD in this display.

5.7 Exercises

1. Briefly explain the concept of recombination (what is a recombinant chromosome?) and define the recombination fraction. What is the relationship between the recombination fraction and linkage?
2. In Fig. 5.2, explain why the two unaffected offspring with recombinations in the region serve to implicate APP as having the DSL. Why are the remaining unaffected offspring uninformative about recombination?
3. Suppose a population of 2000 chromosomes; 1000 carry an A allele at a marker and 1000 carry a. Now suppose a disease mutation (+) arises on one chromosome bearing an A allele, and all the rest of the chromosomes have - at that location.

   (a) What are the marginal frequencies at the marker and DSL?
   (b) Fill in the 2 × 2 table of marker and disease mutation haplotypes.

(c) What are D and D' for this table?
(d) What is the correlation between the marker and DSL?
(e) Repeat the questions above, now assuming only 100 chromosomes, one mutation on the same haplotype as an A allele, and a 50/50 split of A and a alleles.
(f) What is the predicted value of D after 10 rounds of random mating if $\theta = 0.4$? if $\theta = 0.01$?

4. Two polymorphisms (a 23 bp insertion and a 12 bp insertion) near the bovine prion gene are associated with resistance to a neurovegetative disease in cattle (BSE). Data on the presence or absence of variants at these two loci for 350 healthy Holstein cattle are given below:
Both insertions present ($++$ haplotype): 45%
Only the 23 bp insertion present ($+-$ haplotypes): 0%
Only the 12 bp insertion present ($-+$ haplotype): 5%
Neither insertion present ($--$ haplotypes): 50%
Is there evidence for LD? Compute D, $D_{\max}$ or $D_{\min}$, D' and r^2. Given that the $++$ haplotype is associated with disease resistance, what might you expect to see in affected cattle?
5. Show that the maximum value of the correlation $(+1)$ between two biallelic loci is reached when the marginal allele frequencies at locus are the same $P(A) = P(B)$, and the two off-diagonal cells of the 2×2 table are zero. What are the corresponding requirements for an r of -1?
6. Show that the maximum value of D' is 1 when any cell of the 2×2 table is zero.
7. What is a ‘rule-of-thumb’ for relating centimorgans to base pairs? What does 300Kb correspond to in cM?
8. If 2 loci are 2 centimorgans apart, what is their recombination fraction?
9. Suppose the difference in disease allele frequencies in cases and controls is 0.05. Assume we have a nearby marker with minor allele frequency 0.2 and let the correlation between the two loci be 0.8. What is the difference in marker allele frequencies between cases and controls?
10. Using the Haldane map, show that distance is additive, i.e., given three loci A,B and C, the distance from A to B plus the distance from B to C equals the distance from A to C. Hint: assume recombinations between two loci are independent and first calculate P(recombination between A and C) in terms of P(recombination between A and B) and P(recombination between B and C).
11. Verify Equation (5.1).
12. Verify Equation (5.3).

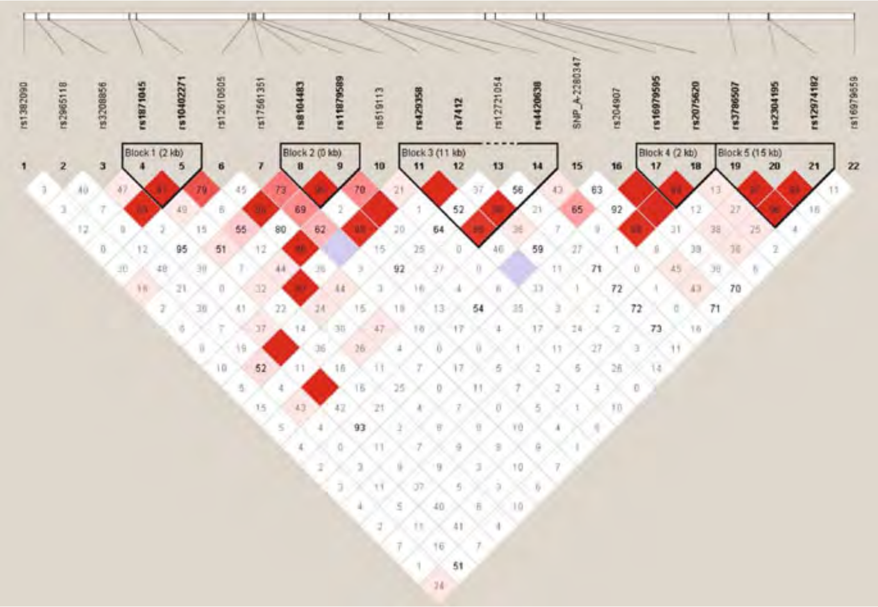


Fig. 5.7 The local LD structure in the APOE gene. The numbers in the squares are the D' values between SNPs that correspond to the squares. *Source*: Coon et al. (2007)

Chapter 6
Basic Concepts of Linkage Analysis

The goal of linkage analysis in human disease gene mapping is to assess whether an observed genetic marker locus is physically linked to the disease locus. This is equivalent to testing the null-hypothesis that the recombination fraction between the marker locus and the disease locus, θ, equals $\frac{1}{2}$. In this case, we say the marker locus and the disease locus are unlinked. It is also possible to estimate θ, which can be used to provide an approximate idea of the location of the DSL relative to observed markers. In this chapter, we discuss the basic concepts of parametric linkage analysis. We explain how linkage between two genetic loci can be utilized to construct long-range mapping approaches that require only a small number of marker loci per chromosome to cover the entire human genome sufficiently. Using fully parameterized statistical models, *parametric linkage* describes the phenotype as a function of the genetic marker locus and its relative distance to the disease locus, i.e., the recombination fraction (Ott (1999)). The simplest case of parametric linkage analysis uses the method of *direct counting*, where θ can be estimated by directly counting recombinant and non-recombinant offspring haplotypes (Ott (1979)). Using the method of direct-counting, we outline the principles of parametric linkage analysis. Advanced topics such as *non-parametric linkage analysis* and *multi-point analysis* (Kruglyak et al. (1996)) are discussed in Appendix A. While the advanced topics that are included in Appendix A are necessary for a thorough grounding in linkage analysis, they are not required for an introduction to association analysis.

The basic approach of a parametric linkage analysis is to construct a likelihood model that describes the phenotype distribution as a function of the unobserved DSL and model the joint transmissions of alleles at the unobserved disease locus and at the observed marker locus as a function of the recombination parameter θ. The likelihood function depends upon the model for the penetrance probabilities, including the mode of inheritance. In simple Mendelian disorders, the penetrance functions are specified as zero or one, and the only unknown parameter is θ. Using standard maximum likelihood theory, the likelihood can be evaluated under the null ($\theta = \frac{1}{2}$) and θ is estimated under the alternative hypothesis via maximum likelihood; a maximum likelihood ratio test for $\theta = \frac{1}{2}$ is constructed by comparing the maximized log-likelihood to the log-likelihood under the null. The validity of the

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_6,

results of the parametric linkage analysis will depend on the validity of the specified model (Clerget-Darpoux et al. (1986); Elston (1998)).

6.1 Basic Approach to Assessing Linkage Between Two Loci

Here we will illustrate the key concepts of parametric linkage analysis using the direct-counting method as an example. In what follows, we first assume that the two alleles at both loci can be directly observed, as it is the case where two SNPs are being mapped to a chromosome. Of course, this assumption is overly simplistic for most diseases, but it allows us here to focus on the key concepts of parametric linkage analysis, i.e., the modeling of the allele transmissions at two loci between two generations in a pedigree.

In any parametric approach, the first step is to understand how the allele transmissions at two genetic loci depend on the recombination fraction θ. This can be done by looking at the production of gametes during the meiotic cell division. In Fig. 6.1, this is illustrated for a parent who has heterozygous genotypes at the two genetic loci, commonly known as a double heterozygote. The alleles at the two loci are Aa and Bb. In this figure, one parental chromosome carries the A-allele at the first genetic locus and the B-allele at the second genetic locus. The second parental chromosome contains an a-allele at the first locus and a b-allele at the second locus. The combination of alleles that reside on the same chromosome is typically referred to as a *haplotype*. A pair of haplotypes in an individual is a *diplotype*. In this example, we observe the haplotypes AB and ab in the parental generation.

For the derivation of the direct counting method, we assume that it is possible to observe the haplotypes in the parental generation and in the offspring generation directly. This assumption will enable us to identify recombination events directly.

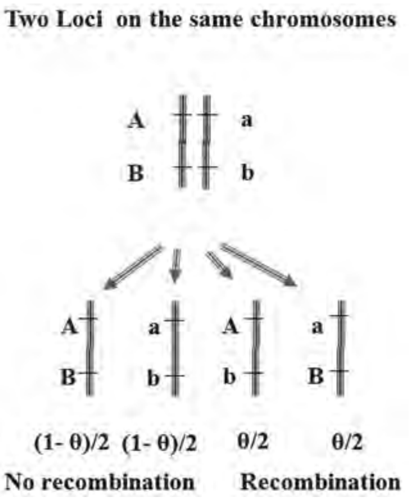


Fig. 6.1 The gamete transmission probabilities

In general, however, it will not be the case in most applications, unless data are available on several generations. Usually, only genotype data, and not haplotype data, are available, and it is not possible to assign the alleles to their original chromosomes when loci are heterozygous. For example for a heterozygous genotype Aa, it is usually not possible to determine whether the A allele is located on the chromosome inherited from their father, or from their mother. When we observe two or more heterozygote markers, *phase* is used to describe the information about chromosomal origin of the alleles. Thus we say for Fig. 6.1, the phase of the double heterozygote individual is AB/ab. The other possible phase consistent with the genotypes is Ab/aB. Phase does not imply knowledge of which parental chromosome the two alleles are on, only which alleles go together on the same chromosome.

***Remark on notation*:** We will use Aa,Bb to denote genotype data on a pair of markers, and AB/ab to denote the pair of (phased) haplotypes or the diplotype.

During the production of gametes in the meiotic cell division, the chromosomes are copied and segments of both chromosomes are exchanged during crossover events. If there is no recombination event between the 2 genetic loci, which happens with probability $(1-\theta)$, then either a gamete with the haplotype AB or a gamete with haplotype ab will be transmitted to the offspring, each with probability $\frac{1}{2}$. The conditional transmission probability that either haplotype, AB or ab, is transmitted to the offspring is then given by $(1-\theta)/2$. Using the analogous considerations for the event of a recombination between the 2 loci, one can show that either the haplotype Ab or aB is transmitted to the offspring with probability $\theta/2$. So, for the given parent, the probabilities for the joint-transmission of alleles at both loci depend on the recombination fraction and their observed distribution can be used to estimate the recombination fraction. Note that if $\theta = \frac{1}{2}$, the transmission probability of all four possible gametes is $\frac{1}{4}$.

Using analogous considerations, we can now calculate the gamete transmission probabilities for all possible configurations of parental diplotypes, which are listed in Table 6.1. A look at the diplotypes reveals that the majority of the haplotype transmission probabilities do not involve the recombination fraction parameter θ. The haplotype transmission probabilities are a function of the recombination fraction only when the parent is heterozygous for both genetic loci, i.e., the parent is doubly heterozygous. Otherwise, the haplotype transmission probabilities are just defined by the Mendelian transmission probabilities.

In terms of being able to detect linkage between the two loci, this implies that only those parental mating-types that contain at least one parent who is doubly-heterozygous at the marker locus and the disease locus can provide information about the unknown recombination fraction between the two loci. This applies to the direct-counting method that we describe here as well as to any other parametric linkage analysis method. Since the only gamete transmission probabilities which depend on the recombination fraction are those of doubly heterozygous parents, the recombination fraction is estimated by counting the number of recombinant and non-recombinant transmissions from those parents to their offspring.

Table 6.1 Transmitted haplotypes for different parental diplotypes: Transmission probabilities conditional on the parental diplotypes

Parental diplotypes	Transmitted haplotypes			
	ab	aB	Ab	AB
ab|ab	1	0	0	0
ab|aB	$\frac{1}{2}$	$\frac{1}{2}$	0	0
aB|aB	0	1	0	0
ab|Ab	$\frac{1}{2}$	0	$\frac{1}{2}$	0
ab|AB	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$
aB|Ab	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$
aB|AB	0	$\frac{1}{2}$	0	$\frac{1}{2}$
Ab|Ab	0	0	1	0
Ab|AB	0	0	$\frac{1}{2}$	$\frac{1}{2}$
AB|AB	0	0	0	1

6.2 The Direct Counting Method

We now illustrate a simple method for linkage analysis, the direct counting method. The method consists of 3-steps. In the first step, we identify all parent-offspring pairs with a doubly-heterozygous parent and their child. The same child is included in two pairs if he/she has two doubly heterozygous parents, because transmissions of the two parents are independent by Mendel's first law. Thus the effective sample size is transmissions, not offspring. Based on the observed haplotypes in the parent, one needs to infer which haplotypes are transmitted as non-recombinants and which ones as recombinants. Denoting the number of heterozygote parent-offspring pairs that yield a recombinant transmission by r, and those yielding non-recombinant transmissions by s, the total number of informative meiosis is defined by $n = r + s$. Because all transmissions from parents to offspring are independent, the likelihood is proportional to the Binomial probability of the observed data:

$$\text{Pr(r|n)} = \binom{n}{r}\theta^r (1-\theta)^{n-r}.$$

The estimate for θ can then be obtained by $\hat{\theta}$=r/n and the null-hypothesis of no linkage can by tested by χ^2 test on one degree of freedom, which is given by:

$$\chi_1{}^2 = \frac{(r-s)^2}{n}.$$

This test is often referred to as a McNemar test in the non-statistical literature, although in fact, it is simply a Pearson χ^2 test of the null hypothesis that the binomial parameter equals $1/2$.

Example: Figure 6.2 illustrates one pedigree which has data on two genetic loci; each loci has three alleles, A,B,C and 1,2,3. This pedigree and marker set up is typical of those used to create the first human linkage map showing the location of hundreds of microsatellite markers as well as other genetic landmarks. Such pedigrees had two parents, four grandparents and a large number of offspring. The grandparents were used primarily to determine the phase of the parents. Multiple offspring are desirable because transmissions to multiple offspring are independent, thus the effective sample size increases with more offspring, without requiring additional genotyping costs for new parents and grandparents. Multiallelic markers are also useful since they can increase the chance of observing heterozygotes and determining phase. In Fig. 6.2, all four grandparents are double homozygotes, allowing one to infer phase in the parents. The circled offspring are recombinants, therefore $r = 2$ and $n = 5$.

While the McNemar-test can be used here to test the null-hypothesis, likelihood-ratio tests are more commonly used in parametric linkage analysis, because they extend to handle more complicated situations. The likelihood ratio test can here be constructed by taking the natural log of the likelihood ratio,

$$\mathrm{LR}(\theta) = \frac{\theta^r(1-\theta)^{n-r}}{(\frac{1}{2})^n}, \tag{6.1}$$

comparing the likelihood evaluated under the null hypothesis $[\theta = \frac{1}{2}]$ to the maximized likelihood under alternative hypothesis (evaluated at $\hat{\theta} = \frac{r}{n}$). Since the recombination fraction is naturally restricted to the range of 0 to $\frac{1}{2}$, we set $\hat{\theta} = \frac{1}{2}$, if $\frac{r}{n} > \frac{1}{2}$. Under H_0, $P(\frac{r}{n} \leq \frac{1}{2})$ is exactly $\frac{1}{2}$ for n odd, and approaches $\frac{1}{2}$ for all n when n is large. Thus the asymptotic distribution of the log-likelihood ratio test

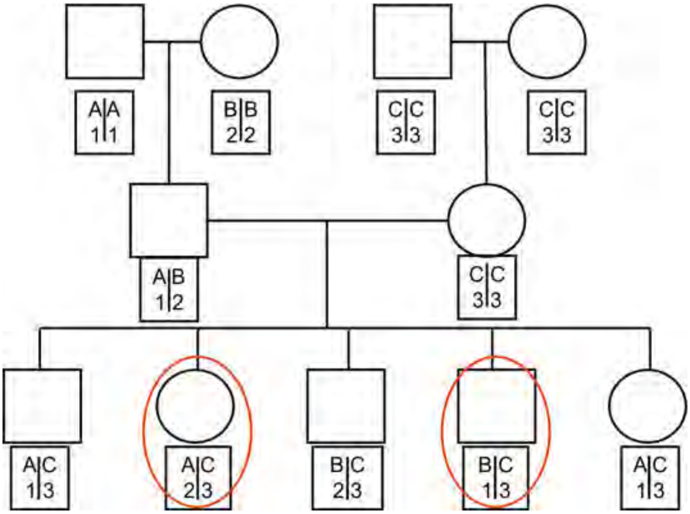


Fig. 6.2 Direct counting method. *Source*: Courtesy of Professor Peter Kraft

is ${\chi^2}_1$ with probability $\frac{1}{2}$ (for $\frac{r}{n} < \frac{1}{2}$), otherwise it equals 0 (for $\hat{\theta} = \frac{1}{2}$). This distribution is also known as χ^2 with $\frac{1}{2}$ degrees of freedom (Self and Liang 1987).

However, linkage analysis evolved far earlier than the theory of likelihood ratio tests, and as we discuss below, in general cases, it can be quite difficult to estimate θ by maximum likelihood. Instead one typically evaluates the LOD-score by

$$\text{LOD-score} = \log_{10}(\text{LR}(\theta)), \tag{6.2}$$

which is the logarithm of the likelihood ratio using base 10 and not the standard base e, as it is usually the case for likelihood ratio tests. The LOD score is a measure of support for an arbitrary value of θ in the range $(0, \frac{1}{2})$, which is maximized when θ is the maximum likelihood estimate. One reason for choosing base 10 instead of e in the logarithm is that it facilities interpretation. A lod-score of 1 says the $P(\text{data}|\theta)$ is 10 times $P(\text{data}|\theta = \frac{1}{2})$, and for a lod-score of 2, the ratio is 100, etc. Another important feature of the LOD-score method is that if data are available on several different families, the lod-score can be evaluated as a function of θ for each family separately, and these can be added over families to obtain the total evidence for each value of θ (See exercise 3 of Section 6.4). This was useful when families were collected by separate investigators at different times prior to the availability of modern computing methods; then only lod-scores for different values of θ needed to be shared to combine the total evidence. Calculating lod-scores separately for each family is also useful in the presence of genetic heterogeneity, i.e., different mutations may be responsible for the same disorder in different families.

Figure 6.3 illustrates these features of the LOD-score method. The figure shows the lod-scores as a function of the recombination fraction θ. The brown, blue, red and green curves show the lod-scores for 4 different families. The overall lod-score is then obtained by summing the lod-scores of all 4 families which is shown as the black line.

We now consider testing for linkage between a marker and a DSL, where, for simplicity, we assume that the observed trait follows an autosomal dominant pattern of inheritance, i.e., $P(\text{disease}|\text{DD or Dd}) = 1$ and $P(\text{disease}|\text{dd}) = 0$. Consider the pedigree diagrammed in Fig. 6.4. Note that the disease phenotypes are observed, but the disease genotypes given in the figure are inferred from inheritance patterns and assumptions on the prevalence functions. Because both grandfather and father are affected, and the mother is homozygous at both the disease and marker loci, we can infer that the phase of the father is DM/dm. This enables us to count recombinants $r = 1$, and non-recombinants $s = 4$, and thus $\hat{\theta} = \frac{1}{5}$; the lod score is easily computed using equations (6.1) and (6.2).

However, many complications can arise. If the grandmother's marker data is missing, we have no way of determining phase in the father. A likelihood can still be constructed for this case, by calculating r and s for each of the two possible phases, and summing over the phases. With traditional linkage analysis, the markers are

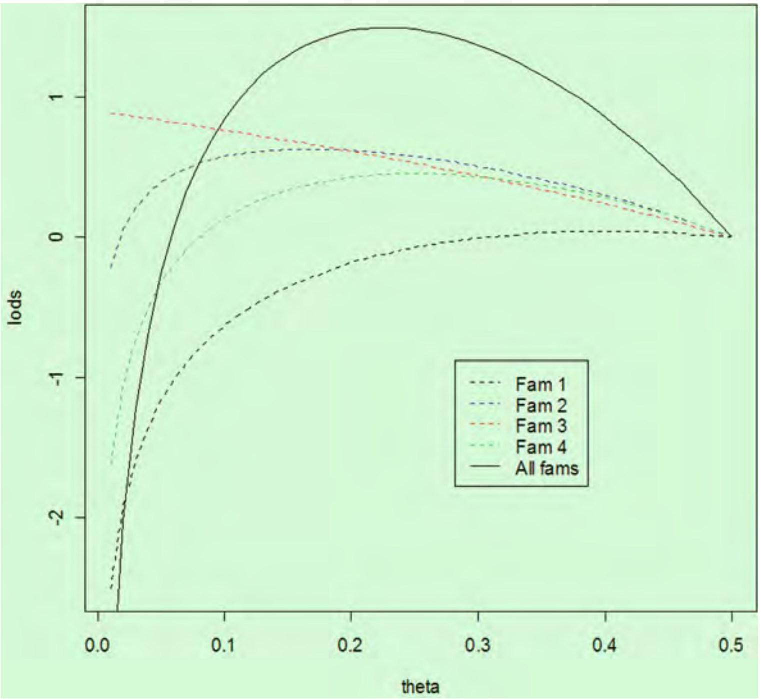


Fig. 6.3 Combining LOD scores on four different families. *Source*: Courtesy of Professor Peter Kraft

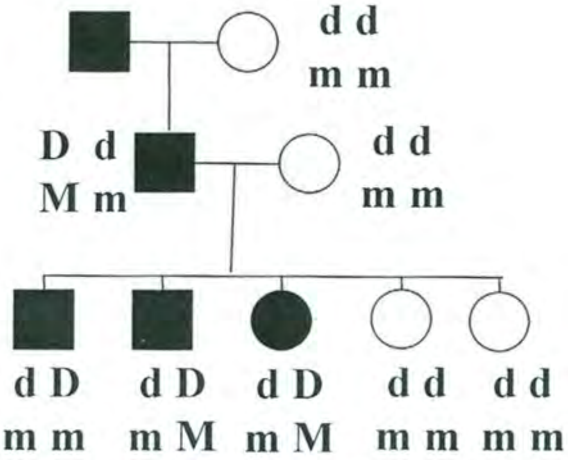


Fig. 6.4 Autosomal dominant disease pedigree with a diallelic marker. *Source*: Adapted from Thomas (2004)

sufficiently far apart so that they are in linkage equilibrium, hence both phases have probability $\frac{1}{2}$. Thus the likelihood ratio is given by

$$LR(\theta) = \frac{\frac{1}{2}\theta(1-\theta)^4 + \frac{1}{2}\theta^4(1-\theta)}{(\frac{1}{2})^5}.$$

The simple method of estimating θ using $\frac{r}{n}$ and testing $\theta = \frac{1}{2}$ using McNemar's Test is no longer applicable. In contrast, maximum likelihood estimation and the lod-score method for testing generalizes easily in this and other more complex cases.

6.3 The Interpretation of LOD Scores

When we analyze only one marker, it is possible to obtain traditional maximum likelihood and/or score tests for testing H_0: no linkage between the DSL and the marker. These tests can be used for testing at any desired α-level. However, the standard α levels of 0.05 or 0.01 are almost never used in testing for linkage. Instead, it has been customary to reject the null hypothesis if the maximum LOD score at the marker exceeds 3.0, which corresponds to an approximate α-level of 0.0001(exercise 2 of Section 6.4) (Kruglyak and Lander (1995), Altmueller et al. (2001)).

There are several arguments that can be given for this approach, most based on the idea that there is a DSL somewhere, as linkage has historically been done only for traits with clear signals from segregation analysis. The issue then is whether or not the marker is close enough to the DSL so that θ is less than $\frac{1}{2}$. With Mendelian disorders, the X-chromosome pattern of inheritance can be detected, so that we can limit our search to the 22 autosomes unless a pedigree analysis suggests an X-linked disorder. A Bayesian argument (Sham (1998)) assigns a prior distribution for θ by first assuming that $P(\theta = \frac{1}{2}) = \frac{21}{22}$ and with probability $\frac{1}{11}$, θ is Uniform(0, $\frac{1}{2}$). Some simple assumptions on n and r allow one to show that the posterior probability that $\theta < \frac{1}{2}$ is approximately 0.90 when the LOD score is 3, and 0.95 when the LOD score is 3.3.

The original derivation of LOD = 3 was given by Morton (1955), who saw linkage analysis as part of an ongoing process. Because LOD scores can be combined over data sets, we think of continuously sampling families and updating the LOD score until significance is reached, or we declare that H_0 is true. Morton used the theory of sequential testing (Wald (1947)) to fix the two types of errors (acceptance of the alternative when H_0 is true (α) and accepting H_0 when H_0 is false (β) at $\alpha = 0.001$ and $\beta = 0.01$). Then setting the prior probability of the alternative to 0.05, rejection of H_0 occurs when the LOD score is greater than 3.0. A problematic feature of Morton's approach is the need to specify a value of θ under the alternative for calculation of the LOD score. He considered several values of θ between 0.05 and 0.3, since values of above 0.3 were considered to be of little practical value. Due to the problem of needing to select a value of θ under the alternative, the criterion evolved to simply max LOD greater than 3.

Both the sequential approach and the Bayes approach were derived with a single marker (2-point analysis) in mind. As the availability of markers increased, the focus of testing shifted to a genome wide testing approach, with H_0: no marker is linked to a DSL. For a set of sparse markers, i.e., markers sufficiently far apart that data at different loci are essentially independent, we can use the Bonferroni approach to multiple testing (see Chapter 10). In this case we set the individual α-level of each marker test to be $0.05/M$, where M is the total number of markers, and 0.05 is the global α-level. With decreasing distance between the markers, the recombination events between 2 adjacent markers become less and less likely, causing the lod-score to become a smooth, continuous function for which a Bonferroni correction for multiple testing would be much too conservative.

With a dense (inter-marker $\theta < .05$) set of multiple markers, we can take a different approach. If the global null is true, i.e., H_0: no marker is linked to a DSL, then the distribution of the max LOD can be derived as a Stationary Gaussian Markov Process with an exponential decay in correlation as distance between any two loci increases. The distribution of such a process is known, and in particular, we can characterize the number of times the process crosses a boundary by a Poisson distribution, whose parameters depend upon the chosen boundary and the total length of the interval (Lander and Green (1987)). Taking into account the length of the genome in Morgans and setting the critical value for the maximum LOD score to be 3.3 leads to an overall genome wide rejection level of approximately 0.05. Because we assume dense spacing for the markers, this criterion does not depend a great deal on the actual number of markers. The reason that we can obtain a desired p-value for a whole genome linkage scan is that with a dense marker set, we know the approximate distribution of LOD score under the null; it is driven by the Markov nature of recombinations.

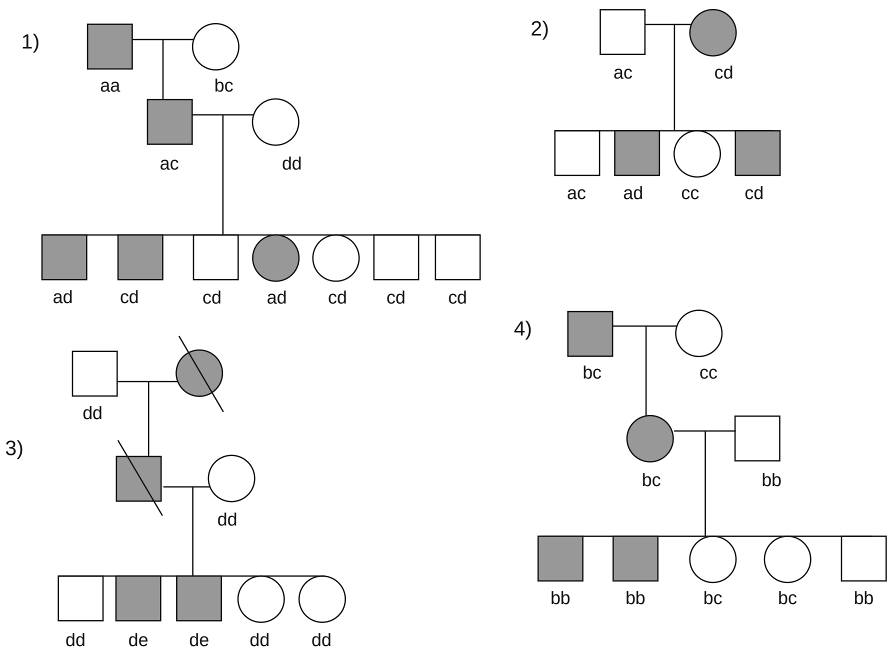
6.4 Exercises

1. The pedigrees below show data on multiplex disease families with data at one multiallelic marker (alleles are a,b,c,d,e). Assume that the disease follows an autosomal dominant model. Assume for this part that the disease is fully penetrant (P(disease|MM or Mm) $=$ 1) and there are no phenocopies (P(disease|mm) $= 0$), where M and m are the alleles at the disease locus.

   (a) Which families are phase known (the parental generation phase can be determined)?
   (b) Explain why non-diseased parents are not informative for linkage.
   (c) For pedigrees 1 and 3, determine the number of recombinants and non-recombinants. Find the ML of θ based on these 2 pedigrees.
   (d) For pedigree 4, what is the LOD score for $\theta = 0$? What would it be if the last offspring were bc rather than bb?
   (e) Write out the likelihood for pedigree 2 as a function of θ and find the ML of θ. What is the LOD at the ML? Now assume that the phase of the affected

mother is known to be Md/mc. What is the ML of θ and the LOD at 0, and LOD at $\theta = \frac{1}{2}$. Comment on the effect of not knowing phase.

(f) Suppose we assume incomplete penetrance, so that P(disease| Mm,MM) = f, $0 < f < 1$, and P(disease|mm) = 0. Consider pedigree 4. Can we assume the disease genotype is known for the unaffected grandmother? What can be inferred about the disease genotypes of the two parents in the middle of the pedigree? What unknown parameters, in addition to θ and f, do we need to specify in order to give an expression for the likelihood?



2. Recall that the standard likelihood ratio test is given by

$$LRT = 2log_e LR(\hat{\theta}).$$

and thus

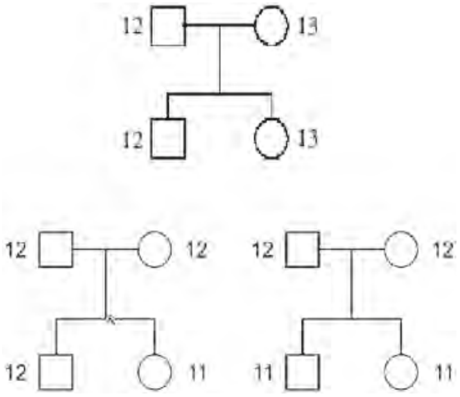
$$\max LOD = (1/2)(log_{10}\, e)LRT.$$

Using this connection between a standard likelihood ratio test and the maximized LOD score, show that a maximized LOD score of 3 corresponds approximately to an α-value of 0.0001.

3. Show that for a given value of θ, we can combine information for independent pedigrees by simply summing the LOD scores evaluated at θ. Hint, if the pedigrees are independent, then the likelihood of all the pedigrees is obtained by multiplying the likelihood of each separate pedigree.

The remaining questions use the material in Appendix A.

4. Refer to Table A.2.
   (a) Verify the sharing probabilities of 0 and 2 for a pair of siblings. Show your work.
   (b) Explain the allele sharing probabilities for monozygotic twins. What are they for dizygotic twins?
   (c) Verify the allele sharing probabilities for grandparent-grandchild pairs. Show your work.
   (d) What are the allele sharing probabilities for first cousins? Show your work.
5. For the NPL allele sharing test discussed in Appendix A.2, verify that $\text{var}(a_i) = \frac{1}{2}$.
6. Determine the number of alleles shared IBS and IBD by the sibs below.



7. Three data sets of ASP families are given. Assume that in each study the same marker has been genotyped and identity by descent can be calculated for everyone.

Study	**IBD=0**	**IBD=1**	**IBD=2**
A	14	45	41
B	8	51	41
C	10	45	51

   (a) For each study, compute the MLS score.
   (b) For each study, compute the NPL-score and the p-value.
   (c) Combine the data from the three studies and compute the scores and p-values based on both tests.
8. Verify that $P(\text{IBD} = 2 \text{ at locus2}|\text{IBD} = 2 \text{ at locus1}) = \psi^2$ under H_0 $\theta = \frac{1}{2}$.

Chapter 7
The Basics of Genetic Association Analysis

A genetic association analysis is not fundamentally different from any other statistical association analysis. The objective is to establish an association between two variables: a disease trait and a genetic marker. The disease trait can be dichotomous, a measured variable, such as lung function or a quantitative measure of obesity, or time-to-onset of a disease or disorder. The genetic marker can be a known or suspected disease-causing mutation, or a marker without any known effects on DNA coding. In the latter case, the association is created by LD between the marker and the disease allele, as discussed in Chapter 5. Another distinctive feature of genetic association analysis is that two quite different study designs can be used; one which uses only unrelated subjects and the other which uses families that have at least two members with genetic marker data. Family designs have distinct advantages and disadvantages, and are an important class of studies. This chapter deals with study designs that use unrelated subjects; Chapter 9 considers designs for association analysis which use data on families.

In experimental plant and animal genetics, design issues focus on selection and manipulation of the genotypes, but in large-scale studies with humans, it is difficult to select subjects for study on the basis of their genotypes. However human subjects are rarely selected for study randomly, and with dichotomous outcomes, most designs ascertain subjects for study on the basis of disease outcome, using either case-control or case-cohort designs. In the case-control design we select diseased cases and non-diseased controls, generally drawn from 'similar populations' in terms of age, sex, ethnicity, etc. In the case-cohort design, we draw both cases and controls from a pre-existing cohort of subjects being followed for development of disease and/or risk factors. These are both standard epidemiologic designs for studying the relationship between general risk factors and disease, with well appreciated advantages and limitations which we do not review here. The relative pros/cons of these study designs generally apply in genetic association as well, with a few exceptions. For example, possible recall bias in collection of the exposures in a case-control study can be eliminated or mediated in genetic studies provided same protocol for genotyping is followed for both cases and controls. This includes collection and processing of DNA samples, and using randomization of cases and controls to batches in the genotyping process to protect against systematic genotyping errors.

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_7,

With measured outcomes, it is also often infeasible to take ‘random samples’ from the population of interest, and convenience samples are selected. If the method of selection is related to the disease of interest, we may have selection bias, and/or an outcome variable with limited range. For example, suppose we use a cohort of subjects selected to be healthy, or employed. If the outcome of interest, obesity or FEV1 say, influences health or employment status, then the sample may be predominantly in the lower (or upper) tail of the continuous outcome, giving lower power to detect the effects of genes which influence unhealthy outcomes. Similarly, if diseased patients are used to study outcomes which measure the severity or symptoms of the disease, it will not be appropriate to extrapolate the results to non-diseased subjects. Population based cohorts involving samples of subjects from a defined geographic population regardless of health status, can be better suited, if inefficient, for this purpose. Selecting individuals from both tails of a continuous outcome distribution can be highly efficient (Lange et al. 2002) but can require screening large populations.

Like all non-randomized designs, these association designs are vulnerable to confounding if there are uncontrolled ‘environmental’ variables related both to disease and to the marker. While we may have no reason to believe that allele frequencies are associated with age and sex and other factors influencing disease, population substructure can cause spurious associations with disease and genetic markers. With population substructure, the genotype distribution varies among members of a population and cases and controls may not be balanced with regard to population substructure. This can cause inflated χ^2 statistics due to variance inflation. When the risk of disease outcome also varies with population substructure, the comparison of the two groups will be biased. We will return to these problems of bias and variance inflation in Chapter 8; in the remainder of this chapter we shall assume that no adjustment for population substructure is necessary.

Although the disease outcome can be measured in any scale, an individual’s genotype is almost always categorical. In this chapter we shall confine attention to SNPs and their 3-category genotypes, but extensions to markers with more than two alleles are straightforward. How we elect to code a genotype for analysis is a key decision which arises no matter what the disease outcome. With only two alleles, there are numerous possibilities; there are pros and cons of using two degree of freedom (DF) tests, which compare all three genotypes, or single DF tests which make some assumption (generally monotonicity) about the relationship between disease and genotype. In some cases, prior evidence, either from biological considerations, or other association studies, may suggest a particular mode of inheritance, but in testing SNPs for common disorders this will rarely be the case. The two DF tests have their proponents, but single DF tests can provide higher statistical power and the simplicity of their interpretation makes them more popular. The choices for the single DF tests are those based on the recessive, the dominant, and the additive modes of inheritance. Readers may wish to review the discussion of genetic models and modes of inheritance in Chapter 2. The most popular assumption is an additive mode, which assumes that the risk associated with the heterozygote genotype is intermediate between the two possible homozygotes. The closer the assumed genetic model is to the ‘truth’, the more powerful the test, but of course, the genetic model is rarely known in practice. In addition, an important, often overlooked fact is that the relationship between disease and disease genotype will be distorted when we test markers for disease association if the markers are not in perfect LD with the DSL. The impact on estimation and sample size calculation of testing a marker rather than a DSL will be discussed in Section 6.9. In the next sections we will focus on testing and estimation for dichotomous outcomes in order to fix ideas. Extensions of the association tests to the more general case that includes covariate adjustments and non-binary phenotypes will be considered in the context of regression models.

7.1 Testing Association with Dichotomous Disease Traits: Codominant, Recessive and Dominant Models

Because each person has two alleles, there are three possible marker genotypes, AA, Aa, and aa. The basic data can be arrayed as in Table 7.1. Here we denote the observed genotype counts by r_0, r_1, and r_2 for the cases, and s_0, s_1 and s_2 for the controls, with the subscript denoting the number of A alleles associated with the genotype and the total number of cases and controls are given by r and s, respectively.

Table 7.1 Table of observed genotype counts for cases and controls

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

***The Codominant Test*:** To test the null hypothesis of no effect of the marker on disease against a codominant alternative:

$H_0 : P(Y = 1|AA) = P(Y = 1|Aa) = P(Y = 1|aa)$
H_A: At least one inequality holds.

The standard two DF Pearson χ^2 test of independence for a 2 × 3 table is (Pearson (1909, 1910)):

$$\chi^2 = \sum (O - E)^2/E,$$

where O is the observed count in the cell, E is the expected count under independence, and summation is over all six cells of the table. The expected count is computed as the product of the corresponding row and column totals, divided by n. The test statistic has a χ^2 distribution with two degrees of freedom. This test is sometimes referred to as the codominant test, or the genotype test, or the test of homogeneity, because it does not make any assumption about the relationship between the genotype and disease. Under the null hypothesis all three genotypes are assumed to have equal disease rates; under the alternative, each genotype may have a different disease rate. The term codominant implies that all three genotypes can have different associated phenotypic risks. When we wish to assume nothing about the relationship between genotype and disease, this is the proper test, but single DF tests are generally more popular because either they correspond to simple Mendelian genetic disease models (recessive or dominant) or because they reflect a belief that there should be a monotone trend between number of alleles and disease state. In the absence of a monotone trend, the heterozygous genotype has either protective or deleterious effects that are stronger than the effects of the two homozygous genotypes. Such disease models are referred to as *heterozygote 'advantage'* or *'disadvantage'* (also known as over-dominance) in the literature.

A feature of the codominant test is that it can reject H_0 when the data support either a heterozygote 'advantage' or 'disadvantage' alternative. The concept of heterozygote advantage is a well known phenomenon in plant and animal genetics, but in humans most examples occur in the setting of two distinct endpoints. For example, the AA genotype at the hemoglobin A gene is the homozygous wild type. For malaria resistance, it is helpful have at least one S variant, i.e., to be heterozygous AS or homozygous SS (see Chapter 1). However the SS genotype predisposes to sickle cell anemia. Thus the 'fittest' individuals are homozygous AS. When testing the effects of a genotype on a single endpoint, concluding that the heterozygote risk is significantly larger/smaller than the risk associated with the two homozygotes may not be plausible in many settings; one alternative is to not reject the test if over dominance is observed, even if the p-value is less than the specified α-level. We return to the issue of testing codominance while excluding the heterozygous advantage/disadvantage when we discuss permutation tests.

***Tests for Recessive and Dominant Modes of Inheritance*:** To test recessive or dominant models of association is straightforward; we simply create the 2×2 table by combining the two appropriate columns representing either recessive or dominant genotypes. For example, for the dominant model, we have the counts shown in Table 7.2 and the null hypothesis is H_0: disease status does not depend on the presence of an A allele in the genotype. A standard χ^2 test with one degree of freedom can be applied to test the null hypothesis. A similar test can be constructed for the recessive model by combining the first and second columns of Table 7.1.

All of the tests discussed in this section, and in the next section as well, are valid under the general null hypothesis: $H_0 : P(Y = 1|AA) = P(Y = 1|Aa) = P(Y = 1|aa)$, but are designed to have optimal power under the mode of inheritance

Table 7.2 Data array for testing a dominant model

	Genotype		
	aa	Aa or AA	Total
Cases	r_0	$r_1 + r_2$	r
Controls	s_0	$s_1 + s_2$	s
Total	n_0	$n_1 + n_2$	n

specified under the alternative. Thus the dominant test will have optimal power to detect $H_A : P(Y = 1|\text{at least one A allele}) \neq P(Y = 1|aa)$, but little power to detect the alternative for the recessive mode of inheritance: $H_A : P(Y = 1|AA) \neq P(Y = 1|\text{at least one a allele})$.

7.2 The Additive Genetic Model: The Alleles Test and the Trend Test

There are two tests commonly used for testing the additive mode of inheritance: the alleles test and the trend test, also known as the Armitage trend test, or the Cochran-Armitage trend test (Armitage 1955). Both tests have the same null hypothesis, $H_0 : p_{\text{cases}} = p_{\text{controls}}$, where p_{cases} denotes the frequency of A alleles among diseased and $p_{controls}$ denotes the frequency of A alleles among non-diseased in the population. Both tests also use the same basic statistic, namely the difference in the observed frequencies of the A allele between cases and controls, $\bar{p}_{\text{cases}} - \bar{p}_{\text{controls}}$, see definitions in Box 6.1. The tests differ in how the variance of the estimated allele frequencies is calculated; the alleles test requires that HWE holds under H_0, but the trend test does not.

Despite the drawback of assuming HWE, the alleles test remains very popular for genetic association studies. To implement the test, we simply construct a single degree of freedom χ^2 test for independence in the 2 × 2 table of alleles (see Table 7.3), which cross classifies chromosomes, not subjects, according to allele. As such, the test has wide intuitive appeal and is seemingly model free. The alleles test makes the assumption that Hardy-Weinberg holds under H_0, and the α-level of the test is not maintained if HWE does not hold. We recommend the use of the Armitage test instead of the alleles test, as the latter has all of the attractive features of the alleles test, but does not require HWE for validity. We illustrate the construction of the alleles test in Box 7.1 and show that it corresponds to a test of $H_0 : p_{\text{cases}} = p_{\text{controls}}$.

Box 7.1 Calculation of the Alleles Test from a Sample of Size n Subjects

The 2 × 2 table underlying the alleles test can be obtained from Table 7.1 by counting the number of A and a alleles among the cases and the controls:

Table 7.3 Data array for the alleles test

	a	A	Total	$\bar{p}$
Cases	$r_a = 2r_0 + r_1$	$r_A = 2r_2 + r_1$	$2r$	$r_A/2r$
Controls	$s_a = 2s_0 + s_1$	$s_A = 2s_2 + s_1$	$2s$	$s_A/2s$
Total	$n_a = 2n_0 + n_1$	$n_A = 2n_2 + n_1$	$2n$	$n_A/2n$

The total number of observations in the table is chromosomes. The alleles test is the one DF χ^2 test of independence in this 2×2 table; it is an appropriate test if observations on chromosomes are independent, or equivalently, HWE holds. It can also be derived as a test of the difference in allelic frequencies: $H_0 : p_{\text{cases}} = p_{\text{controls}}$. Under H_0, $\bar{p}_{\text{cases}} - \bar{p}_{\text{controls}}$ has mean 0 and, assuming HWE, estimated variance

$$v\hat{a}r(\bar{p}_{\text{cases}} - \bar{p}_{\text{controls}}) = \bar{p}(1-\bar{p})\left(\frac{1}{2r} + \frac{1}{2s}\right) = \bar{p}(1-\bar{p})\frac{2n}{4rs}.$$

Hence

$$Z_L = 2\sqrt{rs}(\bar{p}_{\text{cases}} - \bar{p}_{\text{controls}})/\sqrt{2n\bar{p}(1-\bar{p})}$$

is approximately N(0,1) and Z_L^2 is equal to the 1 degree of freedom χ^2 test of independence for the alleles table.

The trend test is also a test of the additive mode of inheritance. It was originally introduced as a trend test in proportions:

$$p(Y = 1|X) = \beta_0 + \beta_1 X, \tag{7.1}$$

where X codes for the additive mode of inheritance and Y codes for cases and controls, i.e., $Y = 1$ for cases and $Y = 0$ for controls. We can use ordinary linear regression to estimate β_1 and test $H_0 : \beta_1 = 0$, so that the computations are simple (Rosner 1994, for example).

An alternative method of motivating the trend test is to treat X as a scaled (0, 1, 2) random variable giving the number of A alleles and compare the means of X in the two groups; this is the approach we use here, as it corresponds to our sampling design and it is easiest to see the connection with the alleles test. By definition, sample means of X in the two groups and overall are

$$\bar{X}_{\text{cases}} = \frac{2r_2 + r_1}{r} = 2\bar{p}_{\text{cases}} \tag{7.2}$$

$$\bar{X}_{\text{controls}} = \frac{2s_2 + s_1}{s} = 2\bar{p}_{\text{controls}} \tag{7.3}$$

and

$$\bar{X} = 2\bar{p}. \tag{7.4}$$

Thus testing $p_{\text{case}} = p_{\text{control}}$ is equivalent to testing that the means of X are equal in the two groups: $H_0 : E(X|\text{case}) = E(X|\text{control})$. Further, under H_0,

$$var(\bar{X}_{\text{cases}} - \bar{X}_{\text{controls}}) = \text{var}(X)\left(\frac{1}{r} + \frac{1}{s}\right).$$

$Var(X)$ can be estimated straightforwardly from the sample variance of the data, without assuming HWE, by

$$\widehat{\text{Var}}(X) = \frac{4n_2 + n_1 - n\bar{X}^2}{n},$$

and thus

$$\widehat{\text{Var}}(\bar{X}_{\text{cases}} - \bar{X}_{\text{controls}}) = \frac{4n_2 + n_1 - n\bar{X}^2}{rs}.$$

As a result, we have that the trend test is given by

$$Z_T = (\bar{X}_{\text{cases}} - \bar{X}_{\text{controls}})/\sqrt{\frac{4n_2 + n_1 - n\bar{X}^2}{rs}}.$$

Using equations (7.2), (7.3), and (7.4), we can re-express Z_T in terms of allele frequencies rather than means of X; the result is given in equation (7.6) of Box 7.2.

To compare the two tests, we write

$$Z_T{}^2 = Z_L{}^2 \frac{2\bar{p}(1-\bar{p})}{[4n_2 + n_1 - n\bar{X}^2]/n}. \tag{7.5}$$

Notice that the ratio on the right hand side of equation (7.5) is equal to the ratio of the estimated var(X) under H_0, with and without assuming HWE; the numerator in the ratio estimates the variance assuming HWE holds, and the term in the denominator is the estimated variance without the HWE assumption. We might expect Z_L to be anti-conservative in general, since it assumes the two alleles of an individual are independent. However, the variance ratio is not always bigger than 1; it depends upon how the sample deviates from HWE. In general we expect it to be bigger than one when there is population substructure, but other factors such as genotyping error can influence deviation from HWE.

Box 7.2 Summary The Trend Test

To test $H_0 : p_{\text{cases}} = p_{\text{controls}}$ without assuming HWE,

$$Z_T = 2\sqrt{rs}(\bar{p}_{\text{cases}} - \bar{p}_{\text{controls}})/\sqrt{4n_2 + n_1 - 4n\bar{p}^2} \tag{7.6}$$

is approximately N(0,1) under H_0, where $\bar{p}_{\text{cases}}$, $\bar{p}_{\text{controls}}$, and $\bar{p}$ are defined in Box 7.1 and equation 7.4. ${Z_T}^2$ is approximately χ^2 with one degree of freedom.
Note: The usual expression for the trend test is a bit different from the one given here (Sasieni 1997), but is exactly equivalent; we use this expression to emphasize the connection to the alleles test. The trend test is very similar to the multiple strata test of trend, also called the Mantel-Haenszel Extension Test (Rosner 1994, section 10.10.4), for the case where there is only one strata. The difference is a multiplicative factor of $(n-1)/n$. This accounts for a slight difference in formulas sometimes seen in the literature.

The linear model (Searle 1971) originally used to derive the trend test is not attractive for proportions from case-control or case-cohort samples. As discussed in Chapter 2, a general model for a mode of inheritance assumes

$$g[E(Y|X)] = \beta_0 + \beta_1 X \tag{7.7}$$

where $g(.)$ is a link function, typically log or logistic for binary data. However, for purposes of constructing a score test for β_1, any choice of link function gives the same test. Intuitively, this is because the variance of the score test is constructed by assuming the null is true, and in model 7.7, $g[p(Y = 1|X)]$ is constant under the null regardless of choice of $g(.)$. We leave as an exercise for the reader to derive the Armitage test of trend by using a score test with logistic model: i.e., assuming

$$\text{logit}[p(Y = 1|X)] = \beta_0 + \beta_1 X.$$

7.3 Small Sample and Permutation Tests

All of the tests discussed in 7.2 assume large samples in order to use asymptotic normality or χ^2 distributions under H_0. For some of these tests, it is also possible to obtain small-sample or exact tests based on the hypergeometric distribution. The hypergeometric distribution gives the probability of the cell counts in a two-way contingency table of counts, assuming that the row and column margins are fixed at their observed values and under the null hypothesis that the row and column variables are independent. The table probabilities given by the hypergeometric can be obtained by simple counting arguments, allowing the computation of exact tests for the tests we have discussed in previous sections. Alternatively, with case-control sampling, p-values for any of these tests and many others can also be obtained by simple permutation of case and control status (Manly 2007). The general idea of a permutation test is as follows. Under the null, we make the general assumption that there is no relationship between genotype and case-control status, hence we can draw ‘Monte Carlo samples’ by randomly assigning each subject a case or control status, while keeping the total number of cases and controls fixed. This fixes the row margins of the 2 × 3 table, and the genotype margins will also remain fixed, because each individual’s genotype does not change, just their case/control status. A large number of Monte Carlo samples, say 10,000–100,000, are drawn. The desired test is computed for each Monte Carlo sample, giving the test statistic distribution under the null. For a χ^2 test, or for ${Z_T}^2$ or ${Z_L}^2$, we then find the $(1-\alpha)100th$ percentile of this Monte Carlo derived distribution, and reject if the observed test statistic is greater than this percentile. Note that this method of assessing significance works for any test because it does not rely on any variance assumptions or HWE. Thus it can be used to give a p-value for the alleles test that does not rely on HWE.

The permutation test has the advantage of extending easily to more complicated situations. For example, we might like to test all three models, recessive, dominant and an additive test, and reject on the basis of the smallest p-value. Such a strategy clearly has inflated type one error rate, but the permutation approach affords a simple modification which preserves the α-level. The general permutation strategy is the same, but now we simulate the distribution of the maximum χ^2, computed over the three tests. All three tests are computed for each Monte Carlo sample, and the maximum is recorded. The last step of rejecting if the observed maximum χ^2 is greater than the $(1-\alpha)100th$ percentile is the same. This test is referred to as the MAX test.

Another use of the permutation test is to derive an appropriate rejection region for the codominant test where we do not reject the test if the pattern of odds-ratios suggests heterozygote advantage/disadvantage. In this case, we would compute the usual codominant test for each Monte-Carlo sample, but discard the result, if the data supports an over-dominance model.

7.4 Which Mode of Inheritance Should We Assume for Testing?

As will be discussed in Section 7.9, we can estimate power for any test under any assumed model. Since the true model is generally not known, a natural question to ask is: Which test is best under a wide range of possible models? Of course the power of a test will always be best if we choose the test for the true model, but some general principles do apply. First, the dominant test and trend test are highly correlated, and give similar results in most cases, regardless of underlying model, especially when the minor allele frequency is low. Second, the recessive test has low power unless the true model is recessive. Third, the codominant test performs almost as well as the test which uses the true model for analysis, and is only slightly out-performed by the MAX test. In practice, the tests of choice are generally the codominant or the trend test. If one is interested in recessive models, the codominant or recessive test should be used, since the additive and dominant tests have little power for a recessive model.

7.5 Estimating Effect Sizes and Confidence Intervals

Although tests of association between phenotypes and potential disease loci are not invalidated by testing markers correlated with the disease loci, estimates and confidence intervals will be most meaningful if we are testing a suspected disease locus mutation; effect estimates for a marker not known to directly influence disease may be distorted by LD between the marker and the disease. We will return to this problem in Section 7.10; for now, we assume that the marker is the true DSL.

The most popular measures of effect for dichotomous outcomes are the risk ratio and the odds ratio. The general risk ratio for an exposed and unexposed group is defined as

$$\gamma = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})} \tag{7.8}$$

In genetics, unexposed will ordinarily be the major allele homozygotes (aa). In the case of a codominant analysis, there are two possible relative risks to estimate among the three groups; typically one compares risk in the minor allele homozygotes (AA) and the heterozygotes (Aa) groups to the major allele homozygotes (aa). For the recessive model, the exposed group is the AA genotype and unexposed is the Aa and aa genotypes combined; for the dominant, the exposed is the AA and Aa genotypes combined, and the unexposed is the aa genotype.

With case-cohort and case-control sampling, the relative risks cannot be estimated. An estimable measure which approximates the risk ratio with rare disease is the odds ratio, defined as

$$\Psi = \frac{P(\text{disease}|\text{exposed})/P(\text{no disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})/P(\text{no disease}|\text{unexposed})}. \tag{7.9}$$

The odds ratio measure has the advantage that it can be validly estimated if subjects are selected on the basis of disease, genotype or randomly sampled (See Section 7.9). The relative risk measure requires either random samples or samples ascertained based on their genotype. Fortunately, the odds ratio closely approximates the relative risk when the disease prevalence is low, since in this case, the ratio of the two P('no disease') terms in 7.9 is close to one and $\gamma \approx \Psi$. Standard methods for estimating odds ratios (see Box 7.3) can be used with the codominant, recessive or dominant models. The allelic odds ratio, calculated from the counts of alleles in Table 7.3, is sometimes used to approximate a relative risk under an additive model. However the relationship of disease risk to this odds ratio is unclear, as it compares odds of disease among chromosomes with an A allele to odds of disease among chromosomes with the a allele. A better approach to estimating disease risk for the additive model is to use logistic regression, as we discuss in the next section. We present the calculation of odds ratios and corresponding confidence intervals for comparing two groups in Box 7.4.

In Section 7.7 we will show how to use different choices for a link function in a generalized linear model and how to code the genotype to estimate odds ratios and relative risks.

Box 7.3 Calculation of Odds Ratios and Confidence Intervals
Let the data array be

	E	U
Number of cases	a	b
Number of controls	c	d

Here E and U indicate Exposed and Unexposesd, which are defined by the appropriate partition of the individuals in the sample according to genotype, or a combination of genotypes; counts are counts of individuals, not alleles. Then the odds ratio, say OR, is given by

$$OR = (a/c)/(b/d) = ad/bc.$$

In large samples, log(OR) is approximately normally distributed, with mean given by the log(OR) in the population, and estimated variance

$$\text{var}[\log(OR)] \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

A $(1-\alpha)100th$ confidence interval for the population OR is found by computing the corresponding CI for the log(OR) and exponentiating the endpoints:

$$\exp^{\log(OR) \pm SE(z_{(1-\alpha/2)})},$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)100th$ percentile of the standard normal and SE is the square root of var[log(OR)].

7.6 Examples of Testing Association with Diallelic Markers

The first example we present comes from a study of TNFα-308 and acne in a Turkish Population (Baz et al. 2008). Acne is a complex, chronic inflammatory disease; although susceptibility to acne is thought to be inherited, there are limited data supporting specific mutations. Tumor necrosis factor-alpha (TNF-α) is one of the pro-inflammatory cytokines implicated in acne pathogenesis. The polymorphism at position 308, which involves substituting guanine (G) for adenine (A), has been linked to increased susceptibility to several chronic inflammatory diseases. Baz et al.

Table 7.4 Distribution of TNFα genotypes in acne patients and control subjects

	Genotype		
	GG	GA	AA
Acne patients [n (%)]	66(58.4)	43(38.1)	4(3.5)
Control subjects [n (%)]	99(86.8)	15(13.2)	0
OR	1	4.30	–[a]

[a] Odds ratio is relative to baseline GG group, and cannot be calculated for the AA genotype due to the presence of the 0 cell; technically the OR is infinitely large

(2008) carried out a case control study of 113 subjects with acne and 114 healthy control subjects; the data are shown in Table 7.4. The analysis presented here has been adapted from that in Baz et al. (2008)

The chi-square test for codominance is 24.1 with 2 DF. This is highly significant with $p < 6 \times 10^{-6}$, but the expected cell counts for the first column are both small, approximately 2. The exact test is preferable, although as a practical matter it is not likely to have much effect on the statistical significance of the observed data unless the α - level of the test is very small. Testing the recessive model is not desirable due to the low numbers of AA genotypes, and testing the dominant model will give very similar results to the codominant model (see exercise 2 of Section 7.11). Because the marker is a suspected DSL, it is reasonable to estimate odds ratios here. Note that there is an increasing trend in the estimated ORs suggesting that the A allele is more common among those with acne. The OR involving AA cannot be estimated because of the 0 observed for the control subjects. Assuming a dominant model, and constructing the 2 × 2 table for the dominant model (Table 7.2), the OR is given by

$$OR_D = (47)(99)/(15)(66) = 4.7.$$

The second example comes from a study of the D3 Dopamine Receptor and Schizophrenia. Schizophrenia is a debilitating mental illness with high heritability. The Dopamine Receptors as a group have long been implicated in causing the symptoms of schizophrenia, in part because of their important role in the central nervous system. Shaikh et al. (1996) examined a Ser-9-Gly polymorphism in the dopamine D3 receptor gene for association with schizophrenia in 133 patients and 109 controls. The data and the analysis are given in Table 7.5.

All the tests except the recessive indicate a statistically significant association $p < 0.05$) of the dopamine D3 receptor gene polymorphism Ser-9-Gly with schizophrenia. We illustrate the calculation of the two odds ratios for the codominant model, using the major alleles homozygotes (22) as baseline. Thus

$$OR_{11/22} = (7 \times 33)/(20 \times 57) = 0.203$$
$$OR_{12/22} = (69 \times 33)/(56 \times 57) = 0.713$$

The trend in the odds ratios suggests decreasing risk as the number of ‘1’ alleles increases ($OR_{22/22} = 1$).

Table 7.5 Distribution of the dopamine D3 receptor genotype in schizophrenia cases and controls

	No.(%)of alleles		No.(%)of genotypes		
	1 allele	2 allele	11	12	22
Cases	83(0.31)	183(0.69)	7(0.05)	69(0.52)	57(0.43)
Controls	96(0.44)	122(0.56)	20(0.18)	56(0.52)	33(0.30)

Alleles test: $\chi^2 = 8.46$, $p = 0.004$.
Codominant test: = $\chi^2 = 11.75$, $DF = 2$, $p = 0.003$.
Recessive Test: $\chi^2 = 3.85$, $DF = 1$, $p = 0.05$.
Trend Test: $\chi^2 = 9.49$, $DF = 1$, $p = 0.002$.
Hardy-Weinberg tests by group:
Controls $\chi^2 = 0.19$, ns,Cases $\chi^2 = 5.81$, $p = 0.02$.

Note that the heterozygotes are over-represented among the cases relative to HWE expectations (O = 69, E = $133 \times 2 \times 0.31 \times 0.69 = 56.9$); this is statistically significant at the 0.05 level $(p = 0.02)$. This is counter to what we might expect with population stratification and/or inbreeding (Chapter 3). A potential explanation could be genotyping errors. In any event, this explains why the trend test is bigger than the alleles test, since the variance of the allele frequency is decreased when the proportion of heterozygotes is increased relative to HWE.

7.7 The Regression Approach: Extensions to Covariate Adjustment and to Other Phenotypes

The simple analyses presented in the preceding sections will be adequate in most situations when the outcome is dichotomous and environmental factors have a negligible effect on disease risk, or have been carefully controlled so that they can be ignored in the analysis. While this assumption is plausible for some diseases, covariates such age, gender, smoking-behavior, etc., cannot be ignored in the analysis of many complex diseases. As we will see in the next chapter, the inclusion of covariates in the genetic association analysis is also one of the most effective ways to guard population-based designs against the confounding effects of population admixture and stratification.

In any of these situations, a regression approach using generalized linear models (GLM) is a natural extension; this approach also has the added benefit of allowing for other phenotypes, as noted in Chapter 2. For some complex diseases, the analysis of intermediate phenotypes or endophenotypes that define affection status or the severity of the disease can be a valuable alternative that can provide additional insight into the disease pathway/genetic composition of the disease. We now outline the basic concepts of the regression approach below.

Let Y_i denote the phenotype of the ith individual and let X_i denote a coding for the genotype. The regression model for a generalized linear model was given in equations (7.7) and (2.3). As discussed in Section 2.2, the phenotype can be a dichotomous disease indicator, a count or a measured outcome. The typical link functions for dichotomous outcomes are the logit and log-link. Estimation of the parameters of a genetic model using the logit link will result in the use of odds ratios to describe the genetic effect, and the use of a log-linear model results in estimated relative risks. The linear link has also been used to characterize genetic models, especially when deriving tests, but is rarely used to estimate effects with dichotomous phenotypes.

The definition of X_i depends on how we specify the genetic model; but in general, the coding is a way of transforming an individual's genotype into a quantitative variable amenable for a regression analysis. For a codominant model, X_i is a vector of two dummy variables specifying two of the three genotypes; in combination with logit or log-linear link functions it provides odds ratios and relative risk estimates for the heterozygous and homozygous minor allele genotype relative to the major allele homozygotes. The additive coding used with the log-link results in multiplicative risk models; when used with the logit link it provides an approximation to the multiplicative risk when the disease prevalence is low. These genotype codings are summarized in Table 7.6, where the aa genotype is the major allele homozygous genotype. The interpretation of β_1 also depends on the coding, although β_0 will always be the prevalence in the aa group.

The two link functions that are most commonly used are the linear for measured outcomes, i.e.,

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

and the logistic for dichotomous outcomes:

$$g[E(Y_i|X_i)] = \log\left[\frac{E(Y_i|X_i)}{1 - E(Y_i|X_i)}\right].$$

Standard likelihood ratio tests for logistic or linear regression (the latter assuming normality for the outcomes) can be used to test $H_0 : \beta_1 = 0$, or no relationship between the mean of Y and X. In the case where Y is dichotomous, the likelihood ratio tests will be approximately equivalent to the chi-square tests discussed in Sections 7.1 and 7.2 for the appropriate models – codominant, recessive and dominant and the estimated β coefficients will be equal to the log of the corresponding odds ratios. For the additive model, the trend test will be approximately the same test as the likelihood ratio test from logistic regression with additive coding for X_i. Because the regression procedures operate on variables defined for individuals, not chromosomes, there is no underlying assumption about HWE.

The estimated β_1 coefficient from fitting the logit model with additive coding for X_i estimates the log relative risk for the multiplicative model under the rare disease

Table 7.6 Coding the genotypes

Genotype	Codominant[a]	Additive	Recessive	Dominant
AA	$X' = (01)$	$X = 2$	$X = 1$	$X = 1$
Aa	$X' = (10)$	$X = 1$	$X = 0$	$X = 1$
aa	$X' = (00)$	$X = 0$	$X = 0$	$X = 0$

[a] Each column indicates the value of X for a given genotype

assumption. That is, denoting a relative risk model

$$\log E(Y_i|X_i) = \alpha_0 + \alpha_1 X_i,$$

and denoting the logit model by

$$\log\left[\frac{E(Y_i|X_i)}{1 - E(Y_i|X_i)}\right] = \beta_0 + \beta_1 X_i,$$

then under the rare disease assumption

$$\hat{\beta}_1 \approx \hat{\alpha}_1.$$

With measured outcomes, using the linear link function corresponds to a standard one-way ANOVA model with three groups for the codominant model or a standard regression model for the recessive, dominant or additive models. The effect measures for the linear models are differences in means by genotype groups. For example, for the recessive model, β_1 is the mean phenotype in the AA genotype group minus the mean phenotype in the combined Aa and aa genotype groups. For the additive model, β_1 gives the population increase in mean phenotype as one A allele is added to the genotype. Standard linear regression methods can be used for testing, estimation and computation of confidence intervals.

An added advantage of the regression approach is that it is straightforward to adjust for covariates. Letting U_i denote a vector of covariates, such as race, age, sex, etc., we can incorporate covariates into the regression by using

$$g[E(Y_i|X_i, U_i)] = \beta_0 + \beta_1 X_i + \zeta U_i,$$

where ζ is a vector of coefficients for the covariates. With a linear model, covariate adjustment is done for two primary reasons: either we wish to avoid confounding the relationship between genotype and phenotype, or we want to increase the precision of the analysis by reducing the residual variance. Setting aside the issue of confounding until Chapter 8, the use of covariates with measured outcomes can be very useful to reduce variability if there are predictors which are highly correlated with outcome. We recommend covariate adjustment with measured outcomes if covariates provide good variance reduction. However, this is a property of the linear model. With the logistic model, one cannot expect to gain precision from including covariates in the model. Nonetheless, sex, age, race and/or ethnicity are frequently controlled for in a genetic analysis by using regression, stratification, and/or design. In general, the selection of the phenotype and the corresponding covariates is a non-trivial process that requires some thought and is addressed best during the design phase of the study. For example, for the analysis of the non-binary variables, we have ignored any possible ascertainment conditions which will not be realistic in most studies, except for population-based studies without any phenotypic ascertainment conditions, e.g., Framingham Heart Study. The Framingham Heart

Study was initiated in the 1950s as a population based sample of over 5,000 adult residents in Framingham, MA, to study the development of factors influencing heart disease. Although subjects were initially free of heart disease, they were followed bi-annually for many years, and during the course of the study, their offspring and spouses and children of their offspring were also recruited into the study.

When intermediate phenotypes or endophenotypes that are correlated with affection status are analyzed in case-control studies, the phenotypic distribution of the intermediate phenotype will reflect the ascertainment conditions as well. In such situations, the ascertainment condition should be incorporated into the association analysis (Slatkin 1999; Chen et al. 2005; Huang and Lin 2007).

7.8 Association Analysis with Complex Traits: An Association Between INSIG2 and BMI

One of the first genome wide association studies led to the discovery of an association between a SNP in the region of the INSIG2 gene and body-mass index (Herbert et al. 2006). It was originally detected in the family sample of the Framingham Heart Study and, then, subsequently replicated in studies with different designs and phenotype definition, i.e., obesity as a dichotomous trait defining subjects as obese if their BMI was greater than 30. However, while several replication attempts resulted in clear confirmations of the detected association, an equal number of replications failed to confirm the signal. Here, we revisit some of these findings and examine how the differences can be caused by different phenotype choices, i.e., quantitative trait analysis versus dichotomous trait analysis.

For many complex diseases, such as obesity or asthma, we have to select the target phenotype for the analysis. This can often be the affection status definition which is typically derived based on other intermediate phenotypes. For example, obesity defined based on a cut point, usually 30, of Body Mass Index (Weight in kg)^2/(height in cm), or asthma based on lung-volume measurements such as FEV1. Instead of using affection status, one could test the intermediate phenotypes directly for association with the genotyped SNPs. The appealing property of quantitative traits in this context is that, by definition, they contain more information than do dichotomous traits, and can result in potentially more statistical power in the analysis. However, they come at the disadvantage, in that they often depend on other physical characteristics, e.g., BMI depends on gender and age, FEV1 depends height, age and gender. In the analysis, such factors should be included as covariates in the analysis.

In this section we review the analysis results for the reported association between a SNP in the INSIG2 gene region and BMI (Herbert et al. 2006; Lyon et al. 2007). In this analysis, obesity was analyzed first as a quantitative trait and then as a dichotomous trait, defining ‘obese’ as study subjects with an BMI of 30 or greater. Probands whose BMI is lower than 30 are classified as ‘unaffected’. Since the original association was reported under a recessive model, both the quantitative trait analysis and the dichotomous analysis were conducted under a recessive mode of inheritance as well.

The SNP was tested in the unrelated subjects from the Framingham Heart Study (1491 study subjects), a sample from Iceland (5187 study subjects) and the German cohort study, KORA (4082 study subjects). All 3 studies are cohort studies with no ascertainment for BMI or obesity. Consequently one might expect that, given that such a design is not enriched for 'obese' cases, the quantitative trait analysis should be more powerful than a case/control analysis with relatively few cases. However, the quantitative trait analysis many have low power if the SNP acts on the extremes of the distribution. Since all study subjects in the Framingham Heart Study were examined 6 times, with each exam being about 6 years apart from each other, the SNP was tested for association with BMI at all 6 exams.

The results for affection status are given in Table 7.7. Here, the SNP was tested for association with logistic regression analysis that is adjusted for the covariates age and gender. In the quantitative trait analysis, a linear regression model was used that described BMI as a function of the marker score, age and gender. These results are reported in Table 7.8. Comparing the quantitative trait analysis with the dichotomous trait analysis, we can see that the association analysis results for the Iceland study and the German Kora Study are consistent with regard to statistical significance at $p = 0.05$ across analysis approach. While the association between the SNP in the INSIG2 region and obesity/BMI is clearly replicated in the Iceland study, the replication fails for both phenotypes, BMI and obesity, in the German Kora Study. For the Framingham Heart Study, the results vary. While the reported association with obesity supports the original finding for the first 3 exams, the quantitative trait analyses do not. This illustrates one major issue with the phenotype selection in association analysis. The results can depend highly on the choice of the target phenotype. The quantitative trait analysis here, which might have been expected to be more powerful, does not replicate the finding as well as the dichotomous analysis. Other analyses (Heid et al. 2009) of this SNP and obesity indicate that it is better associated with more obese subjects (BMI $>$ 37.5), and in case-control or family studies specifically designed to include obese subjects. Furthermore, it is interesting to observe that the replication of the finding for the dichotomous trait seems to depend on the exam, i.e., the age of the subject. The feature that genetic association can be age-dependent has previously been observed (Lasky-Su et al. 2008a, Shi et al. 2009a,b) and highlights one potential reason for non-replications of genetic

Table 7.7 Association studies of rs7566605 CC genotype and obesity as a dichotomous trait (BMI$>$ 30)

Cohort	Odds Ratio	95% CI	*p*-value
FHS 1	1.26	0.78–2.01	0.06
FHS 2	1.52	0.95–2.43	0.08
FHS 3	1.81	1.22–2.70	0.003
FHS 4	1.18	0.80–1.74	0.39
FHS 5	1.14	0.79–1.65	0.48
FHS 6	1.12	0.79–1.59	0.51
Iceland	1.29	1.06–1.57	0.006
KORA S3	0.90	0.70–1.16	0.44

Table 7.8 Association tests of body mass index as a continuous trait under a recessive model for SNP rs7566605. BMI was log transformed and the model included adjustments for age and gender

Cohort	*p*-value
FHS1	0.270
FHS2	0.395
FHS3	0.096
FHS4	0.442
FHS5	0.514
FHS6	0.565
Iceland	0.020
KORA S3	0.81

association findings. Study heterogeneity can be introduced by many factors, including different ages at the phenotype assessment. In the design of genetic association studies that are aimed to replicate previously reported findings, it is important to consider characteristics of subjects in the original population, as well as phenotypes and study design.

7.9 Sample Size and Power Considerations for Case-Control Design

There are numerous computer packages (PBAT, PLINK, QUANTO) which allow one to calculate the power of a given test for a fixed sample size, or the sample size necessary to achieve a certain power. We recommend using one of these packages when planning a study. Our purpose here is to discuss parameters to consider when designing a study, and to provide a simple formula which can be used to compare the effects of differing assumptions. This section is organized as follows. First we present notation and a formula for power (or sample size) that can be used for most of the tests we have discussed in previous sections. Then we explain how the genetic parameters can be specified under case control sampling in order to use the formula.

In this section, we will assume that the marker is the DSL; in the next section we discuss what happens with both estimation and power and sample size calculations when the genetic marker is not the DSL. For this reason, in this section we will use the notation D and d to designate the two alleles, with D being the presumed DSL.

Notation Box

Let i index the number of copies of D (0, 1 or 2).

f_i = p(disease | i copies of the disease mutation)
p_D = frequency of the D allele
g_i = genotype frequency (not assuming HWE)
K = disease prevalence in population

	$= \sum g_i f_i$, summation over i
	Q = 1–K
Relative Risk Models:	$\gamma_i = f_i / f_0 \leftrightarrow f_i = \gamma_i f_0$
Multiplicative:	$\gamma_2 = {\gamma_1}^2$
Dominant:	$\gamma_2 = \gamma_1$
Recessive:	$\gamma_1 = 1$
Additive:	$\gamma_1 = \frac{1+\gamma_2}{2}$

Recall that the underlying genetic model specifies allele (or genotype frequencies) and penetrance functions, f_i, as defined in the notation box. Since penetrance functions are ordinarily difficult to specify, it is customary to specify effect sizes and overall disease prevalence instead. Although risk ratios cannot be estimated in case-control designs except under the rare disease assumption, for purposes of calculating power or sample size, any measure of effect size can be used. We use the relative risk because of its popularity and simplicity. The most commonly used risk ratio models are shown in the Notation Box, along with the notation for this section. The advantage of using one of the relative risk models is that, apart from the codominant model, only one parameter needs to be specified for each mode of inheritance, in addition to overall disease prevalence and allele frequency. The baseline penetrance, f_0 can be determined using the constraint imposed on the penetrance functions by the genotype frequencies and the overall prevalence.

Many of the tests that we have discussed (recessive, dominant, and the alleles test) can be framed as a comparison of two binomial proportions; this fact makes it easy to derive simple power formulas for these tests (see Box 7.4). To use these formulas in our setting, we need to translate assumptions about effect sizes, population prevalence, allele frequency and mode of inheritance into the case-control sampling framework.

Box 7.4 Power Formulas for a Case-Control Sample: Testing the Difference in the Proportion Exposed Among Cases and Controls

The approach to calculating power for the recessive, dominant or alleles test is to note that the corresponding χ^2 tests are equivalent to two sided Z-statistics, where we compare the proportions ‘exposed’ in the case and control groups. How we define ‘exposed’ depends upon the mode of inheritance being tested, but in any event, it will be determined by genotype. As before, let r, s and n denote the number of subjects in the two groups and overall, let $\bar{q}_{\text{case}}$ and $\bar{q}_{\text{control}}$ denote the proportions ‘exposed’ in each group. The null hypothesis is $H_0 : q_{\text{case}} = q_{\text{control}}$, where q_{case} and q_{control} are the proportions exposed among the diseased and non-diseased in the population. To derive power and sample size formulas, we use a large sample normal approximation to the

Binomial for the observed proportions exposed in the two groups, and make two simplifying assumptions: (1) the variance of the test statistic is approximately the same under H_0 and H_A, and (2) $(q_{\text{cases}} - q_{\text{controls}})$ is positive hence we can neglect the probability of rejecting the test statistic because it falls in the lower tail rejection region. With these simplifications, power can be calculated approximately as:

$$\text{Power} \approx 1 - \Phi\left(z_{(1-\alpha/2)} - \frac{\Delta}{\sigma_0}\right), \tag{7.10}$$

where $\sigma_0 = \sqrt{\frac{q(1-q)n}{rs}}$, $q = \frac{rq_{\text{case}}+sq_{\text{control}}}{n}$, and $\Delta = q_{\text{cases}} - q_{\text{controls}}$. Φ is the standard normal cumulative distribution function and z_t is the t-th quantile of the standard normal distribution. From this approximation, it is easy to derive sample size formulas for a fixed power, say $(1-\beta)$, assuming an equal number of cases and controls:

$$r = s = \frac{2(z_{(1-\beta)} + z_{(1-\alpha/2)})^2 q(1-q)}{\Delta^2} \tag{7.11}$$

To use the power or sample size formulas given in Box 7.4, it is necessary to say how the two probabilities, q_{case} and q_{control}, are defined in terms of the test and the genetic model, under case-control sampling. Table 7.9 gives a general expression for the genotype probabilities under case control sampling as a function of population genotype frequencies, disease prevalence and penetrance functions. This table is derived very simply by using Bayes rule and the definitions in the Notation Box. Thus $P(\text{DD genotype}|\text{case}) = f_2 g_2/K$, etc. Note that the row margins sum to one, since the probabilities are $p(\text{genotype}|\text{case-control status})$.

While Table 7.9 is very general, to be useful, we need to re-express the genotype probabilities and the penetrance functions in terms of parameters that we are able to specify, i.e., allele frequency, p_D, relative risks and K. Although it is not wise to assume HWE in constructing a test statistics, in power or sample size calculations, we usually make that simplifying assumption in power calculations in the absence of knowledge as to how the population might deviate from HWE. The penetrance functions (apart from f_0) are re-expressed in terms of the relative risks. Further, if we specify population prevalence (K), the baseline penetrance function (f_0) can be calculated as a function of the relative risks, and the allele frequency. In practice, investigators can usually approximate prevalence (K), and specify a range of relative risks and allele frequencies that they are interested in.

Table 7.9 Genotype probabilities under case-control sampling

	dd	Dd	DD
Cases	$f_0 g_0/K$	$f_1 g_1/K$	$f_2 g_2/K$
Controls	$(1-f_0)g_0/Q$	$(1-f_1)g_1/Q$	$(1-f_2)g_2/Q$

Remark: Note that if we calculate the odds ratio Ψ for any two columns, K, Q and the genotype frequencies cancel, so for comparing say, DD to dd, Ψ is a function of only the $f_i, i = 1 \ldots 3$, and

$$\Psi_2 = f_2(1 - f_0)/f_0(1 - f_2) \approx \gamma_2$$

when $(1 - f_0)/(1 - f_2)$ is close to 1 (rare disease assumption).

We are now in a position to calculate power for any two group comparison and for any assumed mode of inheritance, including the alleles test. The general approach is as follows. The first step is to define q_{case} and q_{control} in terms of the mode of inheritance we want to test. To test a recessive model, q_{case} is the proportion of DD individuals among the cases and correspondingly for q_{control}. For the dominant test, q_{case} is the proportion of individuals with at least one D allele. The alleles test is also a two-group comparison, but involving chromosomes rather than people; it is slightly more complicated as we explain below.

As an example, the genetic model for a recessive test specifies $\gamma_1 = 1$, hence assuming a recessive model,

$$q_{\text{cases}} = f_2 g_2/K = \gamma_2 f_0 p_D^2/K \tag{7.12}$$

$$q_{\text{controls}} = (1 - f_2)g_2/Q = (1 - \gamma_2 f_0)p_D{}^2/Q, \tag{7.13}$$

$$\Delta = p_D{}^2(\gamma_2 f_0/K - (1 - \gamma_2 f_0)/Q) \tag{7.14}$$

and K can be written as,

$$K = \gamma_2 f_0 p_D{}^2 + f_0(1 - p_D{}^2).$$

This allows us to express baseline penetrance as a function of K, γ_2 and allele frequency:

$$f_0 = \frac{K}{\gamma_2 p_D{}^2 + (1 - p_D{}^2)}.$$

Given p_D, γ_2, and K, we calculate f_0 and obtain values for Δ and σ_0 and compute power for a given α and sample size.

However, we can also calculate the power of the recessive test when in fact, the dominant mode of inheritance is correct. That is, we test for the recessive mode, but then assume that the true mode is dominant. In this case, we still use equations (7.12) to define the two proportions and δ, but now in calculating f_0, we assume $\gamma_1 = \gamma_2$. This gives

$$f_0{}^* = \frac{K}{\gamma_2(1 - (1 - p_D)^2) + (1 - p_D)^2}.$$

Note that $f_0{}^*$ is smaller than f_0 for a fixed K, p_D and γ_2. This results in a smaller Δ^* (see exercise 7 of Section 7.11), and a lower overall power than if the true model is recessive.

The power of the alleles test can also be derived using this simple formula, assuming HWE holds, since it can be expressed as a test of the difference in the D allele frequencies for the two groups, where the total sample size in each group is now $2r$ cases and $2s$ controls. The probabilities for the alleles test underlying the cells in Box 7.1 can be derived from the genotype table by constructing the corresponding alleles table; the two proportions tested are the D allele frequencies in each group. For an assumed mode of inheritance that matches the additive model, we can use either

$$f_1 = (f_0 + f_2)/2$$

for the additive model, or

$$f_1 = \sqrt{f_2 f_0}$$

for the multiplicative model. As above, power can also be calculated assuming the true mode of inheritance is recessive or dominant.

As noted in Section 7.2, the preferred test for the additive model is the trend test because it does not require HWE. Formulas for calculating power and sample size can be given for the trend test which follow the derivation given in Section 7.2 (Slager and Schaid 2001).

7.10 Power and Effect Estimation: Testing a Marker in LD with the DSL

Until now, we have assumed that the marker tested is the true DSL. When we relax that assumption, the estimated effect sizes as well as power calculations will be affected. We first show that the marker can be used for testing $f_0 = f_1 = f_2$, where as before, f_i = P(disease| i disease alleles). Let f_j' denote the corresponding P(disease| j marker alleles), and P_{ji} denote P(j disease alleles | i marker alleles). We use the assumption that disease depends on the DSL, and that there is LD between the DSL and the marker, but that, conditional on the genotype at the disease locus, the disease probabilities do *not* depend on the marker genotypes. In other words, only the DSL is causal, the marker does not directly contribute to disease risk. In that case the marker 'penetrance' functions are simply:

$$f_j' = \sum_{i=0,1,2} f_i P_{ij}.$$

Note that under $H_0 : f_0 = f_1 = f_2$ it follows directly that $f_0' = f_1' = f_2'$. The reverse is generally true as well except that $f_0' = f_1' = f_2'$ if the marker and the DSL are independent, i.e., P_{ji} does not depend on i. In the case of independence, f_j' reduces to K for each j. Thus a marker must be in LD with the DSL or it has no power to test for association between the disease and the DSL. Letting A and a denote the two marker alleles, and g_j' denote probability of a marker genotype with the number of A alleles equal to j, we can write the table of marker genotypes under case-control sampling as:

Table 7.10 Genotype probabilities for the marker under case-control sampling when the marker is not the DSL

	dd	Dd	DD
Cases	$f_0'g_0'/K$	$f_1'g_1'/K$	$f_2'g_2'/K$
Controls	$(1-f_0')g_0'/Q$	$(1-f_1')g_1'/Q$	$(1-f_2')g_2'/Q$

Note that expressions for K and Q are unchanged and the probability of the marker genotype with j alleles, g_j, requires specifying the marker allele frequency as well as LD between the marker and the DSL.

Effect on Estimation: In terms of estimation of odds ratios or relative risks, we have for the DD/dd comparison,

$$\Psi_2' = f_2'(1-f_0')/f_0'(1-f_2)' \approx \gamma_2',$$

again making the rare disease assumption. Since each f_j' is a weighted combination of the f_i, if we make a monotonicity assumption ($f_0 \leq f_1 \leq f_2$), it is easy to see that $f_0' \geq f_0$ and $f_2' \leq f_2$, which in turn implies

$$\gamma_2' \leq \gamma_2.$$

In words, the ‘extreme’ penetrance functions are shrunk toward their average $(f_2 + f_0)/2$ so that γ_2 estimated from data on markers, will be less than γ_2 estimated from data at the DSL (making the rare disease assumption so that the odds ratio approximates the relative risk). Similar relationships can be derived for recessive and dominant relative risks. When there is not perfect LD between the marker and the DSL, the true genetic model will be distorted when computing risk ratios at the marker. The relationship between γ_1' and γ_1 is not so clear; it will depend upon the mode of inheritance and on the pattern of LD (Lin et al. 2007). In some cases, the pattern of LD can create an apparent heterozygote advantage, or disadvantage, in testing the marker, when such a pattern is not a feature of the true disease model.

Effect on Power: While the expressions in Table 7.10 can be used to compute power when testing markers instead of the DSL, a much simpler derivation can be given for the alleles test which provides a simple approximation in other cases. We saw in Chapter 5 that the difference in marker allele frequencies can be expressed as

$$\Delta_A = \Delta_D \rho \sqrt{p_A(1-p_A)/p_D(1-p_D)} \qquad (7.15)$$

where

$$\Delta_D = (p_{D|\text{cases}} \text{ - } p_{D|\text{controls}})$$
$$\Delta_A = (p_{A|\text{cases}} \text{ - } p_{A|\text{controls}}),$$

p_D and p_A are the disease and marker allele frequency in cases and controls combined, and ρ is the correlation between the marker and the DSL.

Using the results of Box 7.4 for calculating a sample size for the alleles test assuming we test the DSL, we have

$$n_{DSL} = 2(z_{(1-\beta)} + z_{(1-\alpha/2)})^2 p_D(1 - p_D)/\Delta_D{}^2,$$

and assuming we test the marker,

$$n_{\text{marker}} = 2(z_{(1-\beta)} + z_{(1-\alpha/2)})^2 p_A(1 - p_A)/\Delta_A{}^2.$$

Using (7.15), we can rewrite n_{marker} in terms of n_{DSL} and ρ^2 as

$$n_{\text{marker}} \approx n_{DSL}/\rho^2.$$

In other words, to adjust for testing a marker in LD with the true disease locus, calculate sample size for the true DSL and adjust it by dividing by ρ^2 (Pritchard and Przeworski 2001) Because ρ^2 is less than or equal to 1, and considerably so for $\rho < 0.8$, testing a marker rather than the DSL can substantially inflate the required sample size.

7.11 Exercises

1. The data below come from the study by Knowler et al. (1988), discussed in Chapter 3, on the association between IDDM type 2 and a haplotype from the GM system human immunoglobulin G. These data include all individuals in a sample of 4,920 Native Americans of the Pima and Papago tribes. In this example, think of the GM haplotype as just an allele at a suspected DSL.

GM haplotype	# subjects	#(%) with IDDM
Present	293	23 (7.9)
Absent	4627	1343 (29.0)

   (a) What mode of inheritance can you test using these data?
   (b) Test for statistical significance of the association between the GM haplotype and IDDM and draw conclusions.

2. Refer to Table 7.4.

   (a) Test for both the additive model (using the Alleles Test) and the Dominant model, and compare results.
   (b) Give the 95% Confidence Interval for the dominant odds ratio reported in the text.
   (c) Test for the Recessive model using Fisher's Exact Test.

3. A case/control study has been conducted and a SNP genotyped. Compute the odds-ratios for the table below and the corresponding confidence intervals. Compute tests for all 3 modes of inheritance (large sample) discussed in the chapter. Discuss the results in terms of plausibility of a model.

	X=0	X=1	X=2
Cases	500	350	120
Controls	521	270	130

4. Previous studies have shown a relationship between SNP6983269 at 8q24 and the risk of colon cancer. The SNP frequency is estimated at 0.55 (the more common allele confers excess risk), and an additive model for risk, with $\gamma_1 = 1.3$, fits the data. Assume the risk of colon cancer in the population is approximately 4%. Use this information to plan a confirmation case-control study which will have power of 80%.

   (a) Assuming $r^2 = 1$ between the SNP and the DSL, what sample size is needed?
   (b) One actual follow-up study used 560 cases and 750 controls. What would the estimated power be? How much did the extra 190 controls help in terms of power, over an equal number of cases and controls? Why might investigators choose to use a few more cases or controls than needed for power?
   (c) The SNP is in a 'gene desert', i.e., there are no nearby genes. Assume that r^2 is 0.8 and then 0.5, and recalculate the required sample size.

5. Derive the trend test as a score test under a logit model for $P(Y = 1|X)$, where X uses the additive coding.
6. In the exercises below, assume r = number of cases and number of controls, and compute all variances assuming H_0 is true ($p_{\text{cases}} = p_{\text{controls}}$). For part c, it will be easiest to use the numerator of the ratio in equation (7.5) for the alleles test, and the denominator for the trend test.

   (a) Derive the variance of $(\bar{p}_{\text{cases}} - \bar{p}_{\text{controls}})$, assuming Hardy-Weinberg Equilibrium.
   (b) Derive the variance of $(\bar{p}_{\text{cases}} - \bar{p}_{\text{controls}})$ assuming inbreeding with genotype distribution given by equation (3.7) in Chapter 3. Hint, recall that $\bar{p} = \bar{X}/2$, so that $var(\bar{p}) = var(\bar{X})/4$.
   (c) Show that the expected value of the estimated variances under HWE (approximately in large sample) are $2np(1 - p)$ for both the alleles test and the trend test.
   (d) Repeat part c, now assuming the inbreeding model, and show the (approximate) expected value for the alleles test does not change, but that for the trend test is $2np(1 - p)(1 + F)$.
   (e) Comment on the appropriateness of the two tests under the inbreeding model.

7. When testing a recessive model when a dominant model is true, show that $f_0{}^*$ is less than f_0 and thus Δ is less than Δ^* and thus the power is less than when we test the true model (assume K, p_D and γ_2 are fixed).
8. Assume a monotone model ($f_0 \leq f_1 \leq f_2$) for the effect of a SNP on a disease. Find values of P_{ji} = P(j disease alleles | i marker alleles) such that the marker-disease relationship displays the heterozygote advantage/disadvantage model. Note that we do not use LD directly, since LD is defined for haplotypes, but P_{ji} depends upon genotypes.

9. In a genetic association study for late-onset Alzheimer's Disease in a Japanese population (Takei et al. 2009), a number of SNPs have been genotyped in the APOE region. The SNPs are listed here in physical order; the distance between the first and last SNP is approximately 55Kb. SNPs rs446037 through rs429358 lie in the APOE gene. SNPs rs429358 and rs7412 (not genotyped, but adjacent to rs429358) together make up alleles $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$ of the APOE gene. The results are shown in Table 7.11.
   (a) Compute the trend test for all genotyped SNPs in Table 7.11.
   (b) Compute the odds-ratios for the heterozygous genotype and the rare homozygous genotype, both relative to the common allele homozygous genotype.

Table 7.11 Genotyped SNPs in the APOE region

		Genotypes					
		Cases			Controls		
SNP ID	MAF	MM	Mm	mm	MM	Mm	mm
rs419010	0.493	112	278	150	206	348	150
rs394221	0.487	149	269	91	153	325	180
rs4803766	0.434	136	273	130	265	328	107
rs395908	0.280	228	240	70	414	261	26
rs519113	0.273	237	226	73	423	258	23
rs412776	0.257	245	230	65	440	237	19
rs3865427	0.209	299	198	43	481	206	14
rs3852860	0.321	191	268	77	359	292	36
rs3852861	0.318	196	261	80	375	294	39
rs6857	0.180	316	184	26	504	183	11
rs157580	0.467	192	262	85	162	351	187
rs157581	0.329	194	251	90	363	281	47
rs157582	0.292	212	240	80	417	250	36
rs157583	0.043	465	57	7	659	34	0
rs1160983	0.048	511	30	0	616	82	3
rs157587	0.043	467	58	6	668	36	0
rs283817	0.044	468	61	7	673	35	0
rs573199	0.038	472	52	7	675	28	0
rs1160985	0.266	323	192	28	346	300	57
rs760136	0.265	323	191	27	338	294	57
rs741780	0.269	314	193	26	340	306	57
rs394819	0.043	473	58	7	672	35	0
rs405697	0.496	98	276	167	216	348	137
rs10119	0.229	249	225	66	509	189	13
rs446037	0.044	468	60	7	674	34	1
rs434132	0.043	474	59	7	675	34	0
rs7259620	0.264	319	189	27	349	301	56
rs449647	0.049	457	70	2	658	44	1
rs769446	0.036	522	22	0	647	60	4
rs440446	0.426	125	270	129	261	319	81
rs429358	0.167	298	202	47	590	124	1
rs7256200	0.128	348	164	27	608	101	0
rs483082	0.208	274	213	51	504	172	12
rs584007	0.439	129	279	130	248	358	95
rs4420638	0.203	297	199	41	565	135	1

Chapter 8
Population Substructure in Association Studies

Genetic association studies using population-based designs have distinct features that make them an attractive approach for gene mapping. Similar to epidemiological studies, they typically use unrelated individuals. As a consequence, the study recruitment is relatively easy and the statistical analysis is straight-forward to implement using standard statistical analysis techniques. This provides population-based designs with an advantage over other designs. Since epidemiological studies have a long tradition in biomedical research and are available for many complex diseases that are expected to have a genetic component, existing epidemiological studies can be converted into genetic association studies without much effort if the DNA of the study subjects is available, e.g., blood samples, etc. The study subjects have to be genotyped at the genetic marker loci, but often no additional phenotyping or, even, recruitment of subjects is required.

In the setting where there is no epidemiological sample readily available that can be converted into a genetic association study for the phenotype of interest, a new study has to be designed, subjects have to be recruited and data has to be collected. Under these circumstances, the use of unrelated study subjects usually facilitates the recruitment process in terms of both the achievable sample size and the recruitment time. The genetic effects of single loci that contribute to complex phenotypes and/or diseases are thought to be relatively small compared to the total phenotypic variability. Consequently, studies with large sample sizes are required to have sufficient statistical power to detect the underlying genetic effects. This is especially true for scenarios in which multiple genetic loci are tested for association and the test results have to be adjusted for multiple comparisons.

Given their advantages in terms of study design, recruitment, and sample collection, population-based association studies have become extremely popular in the field of complex disease mapping. However, their popularity has also exposed and magnified their major weakness, i.e., inconsistent association findings between different studies. Many of the early association findings could not be replicated in other studies and concerns about the limitations and flaws of association mapping were articulated.

Both the advantages and the disadvantages of association mapping in unrelated study subjects originated from the same feature of the mapping technique. As a

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_8,

consequence of the recruitment of unrelated individuals, an observed genetic association can be indicative of a true genetic effect or, as it is the case in any epidemiological study, can be due to confounding, in this case, most likely population admixture or population stratification. While many complex phenotypes and diseases are well understood in terms of the knowledge about the non-genetic environmental factors that influence them, the genetic composition of the study population is usually unknown in terms of potential substructure, i.e., mixtures or distinct, genetic groups that are contained in the sample. As we have seen in Chapter 3, the presence of population substructure can lead to the number of heterozygous genotypes being reduced relative to their expected number under Hardy-Weinberg Equilibrium and the variance of the genotype distribution being inflated, if the study contains subgroups that have different allele frequencies at the locus (Wahlund Effect). If population-substructure and its effect on the genotype distribution remain unaccounted for, the variance of the association tests can be underestimated and, as a consequence, the number of false-positive findings can be notably higher than would be expected based on the specified significance level.

In the presence of population sub-structure, in which the data set consists of several sub-populations, different disease prevalences in each strata or, in the case of a quantitative trait, different phenotypic means in combination with varying allele frequencies can lead to confounding of the disease-gene relationship. Commonly used association tests may show positive results in the absence of any true association, and conversely, true effects can be obscured.

The harmful effects of population substructure on genetic association tests were understood early on and, consequently, there has been a concerted effort in the field to develop statistical approaches whose goal is to make genetic association analysis of population-based samples robust against unknown confounding due to population admixture and stratification. The presence of substructure impacts both numerator of the test statistic and its variance. As a consequence commonly used genetic association test statistics can be become biased and/or the variance can be inflated, both causing the failure to maintain the specified significance level. While tests for Hardy-Weinberg can be used for detecting population substructure, they generally have poor power, and do not offer a solution for adjusting for substructure. In the following sections we will discuss three general approaches for dealing with population substructure; all of them based on the availability of an additional set of markers which are assumed to have no functional relationship to the trait, and which are genotyped on the entire sample.

- The first method, genomic control (Devlin and Roeder 1999), addresses the effects that population-admixture has on the variance of the test statistic. Rather than relying on the theoretical variance, which may or may not be correct, the approach simply estimates empirically the variance in the χ^2 statistics computed at the null markers. A variance inflation factor is estimated by comparing the empirical variance to the variance of the χ^2 distribution. This inflation factor is used to adjust the variance at the marker loci of interest. Because genomic control makes the assumption that the population substructure at the marker loci of interest is the same as for the non-functional markers, the selection of the non-functional markers is critical for the validity of the approach.

- An alternative way to control for population substructure uses a model and data on the additional markers to infer the latent population structure and to incorporate this model into the analysis. Such approaches will be effective if the population sub-structure has effect on both the numerator of the test statistic and its variance. In order to work correctly, the model based approach requires either strong population admixture, e.g., data from a few distinct ethnicities, or a large number of null, carefully selected, ancestry informative markers (Pritchard et al. 2000). Due to the underlying model assumptions, these approaches only work well if the model is estimated correctly.
- Finally, by including null markers or linear combinations of null markers as covariates in the analysis (using linear or logistic regression), regression approaches can control for population substructure when both allele frequencies and disease prevalences vary across sub-strata. These markers, or linear combinations thereof, effectively serve as surrogates for the underlying strata (Chen et al. 2003; Zhu et al. 2002; Zhang et al. 2003; Price et al. 2006).

With each method of adjustment, it is always possible to construct scenarios in which each of a proposed method fails to adjust for the confounding correctly and provides biased test results. The topic of population admixture and stratification in population-based genetic association studies cannot be considered solved. Good epidemiological study design is still the most important and efficient way to avoid confounding in population-based designs. This is particularly true in settings where only the cases are ascertained and genotyped, and the controls are drawn from a previously genotyped sample used for a different case-control study. As we will show, no amount of adjustment can correct for poor designs where population subgroups for cases and controls do not overlap.

We remark that when dealing with case-control studies, population stratification is sometimes defined to mean that cases and controls are drawn from different subpopulations. We will continue to use the definitions given in Chapter 3, but the two definitions are practically equivalent when disease rates vary across strata.

8.1 The Impact of Population-Admixture and Stratification on Genetic Association Tests

To understand the effects that population substructure can have on the association analysis in population-based studies, we first revisit the definition of the association test statistics for designs of unrelated cases and controls. In Chapter 7, we saw that both the trend test and the alleles test can be written as tests with identical numerators, but different denominators, i.e., variances. Using the same notation as in Chapter 7, and assuming for simplicity that the number of cases and controls is equal, i.e., $r = s = n/2$, the numerator of both test statistics can be characterized as proportional to

$$T = r(\bar{X}_{\text{cases}} - \bar{X}_{\text{controls}})$$

To assess the effects of population substructure, we assume the simplest form of population admixture and stratification, i.e., the study population consists of two distinct, but unobserved, subgroups with different allele frequencies, p_1 and p_2 and different disease prevalences. Note that we assume the null hypothesis is true, so that the allele frequencies are the same for cases and controls within subpopulation, but the case-control sampling is done without regard to strata, here treated as unobserved. We let c denote the proportion of cases falling into the first subpopulation, and d denote the proportion of controls falling into the first subpopulation. If $c = d$, the design is balanced with respect to the strata, but this is unlikely to be the case if disease rates vary across strata.

We begin by looking at bias. Note that

$$\bar{X}_{\text{cases}} = c\bar{X}_{1\text{cases}} + (1 - c)\bar{X}_{2\text{cases}}$$

where $\bar{X}_{k\text{case}}$ denotes the mean for X among cases in the k^{th} strata, and similarly for

$$\bar{X}_{\text{controls}} = d\bar{X}_{1\text{controls}} + (1 - d)\bar{X}_{2\text{controls}}.$$

It follows that

$$E(\bar{X}_{\text{cases}} - \bar{X}_{\text{controls}}) = 2(p_1 - p_2)(c - d).$$

Letting K_k denote P(disease in strata k) and K denote P(disease) overall, simple algebra shows that under case-control sampling

$$E(c - d) = S_1 S_2 (K_1 - K_2)/K(1 - K),$$

where S_k denotes the proportion of subjects in strata k. We see that even when the distribution of alleles is the same among cases and controls within strata, so that the null hypothesis is true, the numerator of the test statistic has a non-zero expectation unless $p_1 = p_2$ and/or $K_1 = K_2$. The absence of variation in disease rates or the allele frequencies over strata is sufficient to eliminate this bias. This effect is referred to in the statistical literature as the *Simpson Paradox* and confounding in the epidemiological literature. This result generalizes easily when there is an arbitrary number of strata to give:

$$E(\bar{X}_{\text{cases}} - \bar{X}_{\text{controls}}) = 2\,\text{Cov}\,(p_k, K_k)/K(1 - K), \tag{8.1}$$

where p_k denotes allele frequency in the k^{th} strata, and K_k denotes the P(disease) in the kth strata (See exercise 4 of Section 8.5).

Of course, if we can identify population subgroups we can do a stratified analysis. For this reason, genetic analyses should stratify by race or ethnicity, although self-reported indicators of race of ethnicity are unlikely to completely capture genetic differences. Many researchers (e.g., Wacholder et al. 2000) argue that, controlling for race or ethnicity, any remaining biases are likely to be small, and in any event, will likely be swamped by the extraneous variability induced by admixture or stratification (Devlin et al. 2001).

Formula (8.1) indicates that with a large number of strata there is likely to be low bias unless there is a systematic monotone relationship between disease rates and allele frequencies. Although presumably rare, strong covariances have been documented in several settings (Knowler et al. 1988; Campbell et al. 2005), and are more likely with a small number of strata, or admixtures of a small number of populations.

We now consider variance inflation caused by population stratification. Recall from Chapter 3, the effect of stratification on the variance of an individual's number of alleles, X, is to inflate it to

$$\mathrm{Var}(X) = 2p(1-p)(1+F) \tag{8.2}$$

where p is the overall allele frequency, and F is the Wahlund effect. In our setting, where we assume two strata with HWE holding in each strata, $F = S_1 S_2 (p_1 - p_2)^2 / p(1-p)$, where S_i is the overall proportion of the sample in strata i, and p is the allele frequency in the overall population. In this simple case, $p_1 = p_2$ implies $F = 0$ when HWE holds within strata. More generally, F can be interpreted as the stratification factor, the inbreeding factor, or as the correlation between the two alleles transmitted by the parent. Note that $F > 0$ even if $c = d$, so that inflated variance persists even if the numerator is unbiased. The trend test assumes that individuals in the sample are independent when computing $\mathrm{var}(T)$, but with stratification, the covariance between two individuals drawn from the same subpopulation is $4Fp(1-p)$ (Devlin and Roeder 1999).

Given these expressions, var(T) is given by

$$\mathrm{Var}(T) = 4rp(1-p)[1 - F + 2rF(c-d)^2]. \tag{8.3}$$

To assess the impact of population admixture on the test statistic, we compute the ratio of its true variance given by equation (8.3), to the expectation of the variance computed in Chapter 6, assuming no substructure. For the alleles test, this latter variance is approximately $4rp(1-p)$ in large samples, and for the trend test, this is approximately $4rp(1-p)(1+F)$. (See exercise 5 of Section 6.4) Thus the two variance inflation ratios are given by:

$$\lambda_L = [1 - F + 2rF(c-d)^2]$$

and

$$\lambda_T = [1 - F + 2rF(c-d)^2]/(1+F),$$

where the subscript L denotes the alleles test ratio and the subscript T denotes the trend test ratio. Figures 8.1 and 8.2 illustrate λ_L and λ_T as a function of F, $n = 2r$, and $(c-d)$.

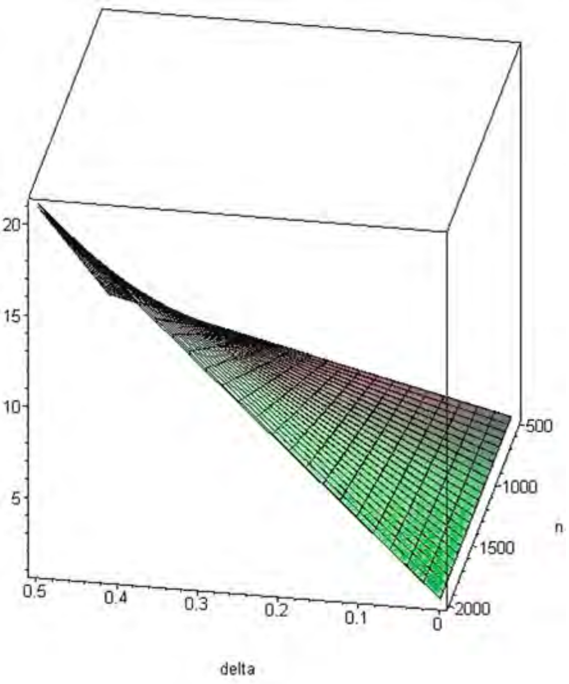


Fig. 8.1 The inflation factors, λ_L and λ_T, for $F = 0.01$ and $\delta = |c - d|$ (The upper surface corresponds to the inflation factor λ_T; it is not visible here since the two inflation factors are so close.)

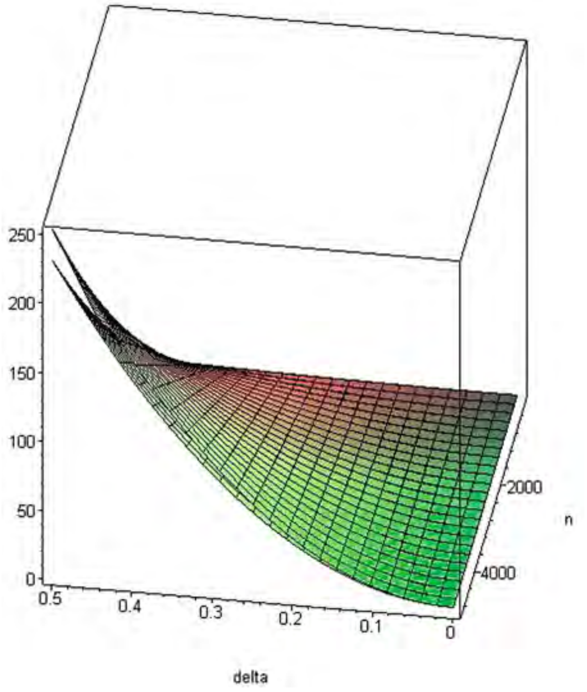


Fig. 8.2 The inflation factors, λ_L and λ_T, for $F = 0.10$ and $\delta = |c - d|$ (The upper surface corresponds to the inflation factor λ_T.)

These ratios have interesting properties that are important to keep in mind in constructing a robust test statistic:

1. If the cases and controls in the dataset are sampled with equal proportions from both subpopulations ($c = d$), but $p_1 \neq p_2$, then λ_L is $(1 - F)$. Since F can be assumed to be positive in the absence of HWE within strata, the variance of Alleles test will be slightly greater than its true variance. Consequently, although slightly too conservative, the Alleles test will be robust against stratification in this case. For the trend test, the ratio here is $(1 - F)/(1 + F)$ which also implies that the trend test will also be robust against population stratification in this scenario.

   However, if sampling proportions from the subpopulations are different, this assessment changes fundamentally. Figures 8.1 and 8.2 show the inflation factors for both the trend test and the alleles test as a function of the difference in sampling proportions in the 2 sub-populations and the sample size n. The upper surface reflects the inflation factor for the alleles test and the lower surface the inflation factor for the trend test. The impact of different sampling proportions from the sub-populations on the inflation factor can be immense. Even with small F (0.01), the true variance of both test statistics can be inflated by several magnitudes when n is large and the difference in c and d is large as well, leading to substantially increased rates of the type-1 error in the test statistic. Similar results can be obtained in cases in which the sample consists of more than 2 subpopulations (Devlin and Roeder 1999; Pritchard et al. 2000).

   In applications to genetic association studies, it is not realistic to assume that the sample consists of equal proportions from the different, unknown subpopulations, even with the best designs. Consequently, potential bias due to population admixture is an issue in virtually any genetic association study. In this context differences between test statistics, trend test or alleles test, are small. The differences between the inflation factors for the two test statistics are too negligibly small to favor one test statistic over the other. Regardless of the test statistic choice, adjustment for population admixture will be one of the most important aspects of the genetic association analysis.
2. Apart from F, the inflation factors of the 2 test statistics do not depend on the overall allele frequency p which is an important property for the construction of an universally applicable adjustment factor. Although F can vary across loci, under relatively mild conditions, F should be approximately constant across SNPs Devlin and Roeder (1999). As a consequence, one single inflation factor can be used to adjust the variance of the test statistic for all genetic markers, regardless of their allele frequencies.
3. The two inflation factors increase with the sample size of the study. Since the inflation factors are a linear function of the sample size, the problem of population admixture does not 'go away' with increasing the sample size, in contrast, it becomes even more severe. This is an important aspect to keep in mind, in particular with respect to genome wide association studies in which several thousand probands are used in the analysis to identify SNPs with relatively small genetic effect sizes.

8.2 Genomic Control Approaches

Since the inflation factors λ_T and λ_L do not depend on the allele frequency of the marker locus being tested, the inflation factor, if known, can be used to adjust the test statistic at any genetic locus across the genome. The idea of genomic control is to estimate the inflation factor by assessing the degree of population admixture that is present in the study by genotyping a set of *null loci*, i.e., a set of genetic loci that are assumed not associated with the phenotype. Under the assumption that the distribution of the selected test statistic at null loci reflects the distribution of the test statistic under the null hypothesis in the data set of interest, an estimator for the inflation factor can be constructed. Since we have to assess an empirical distribution of variances where outliers can have strong effects, a statistical approach that is robust against such effects has to be used. Instead of estimating the average variance at the null loci, the median of all test statistics at the loci is calculated and compared to its theoretical value. Thereby an estimate for the inflation factor is obtained. The observed test statistics at the marker loci of interest are then scaled by the inflation factor λ. Thereby, the effects of population admixture on the theoretical variance of the test statistic are compensated for. The scaled test statistics at the marker loci of interest should then have an asymptotic χ^2-distribution. The details of the approach are outlined in Box 8.1 and are applicable to any of the χ^2 tests discussed in Chapter 7, including the codominant test. Similarly, the genomic control approach can be extended to other phenotypes than affection status.

Box 8.1 Genomic Control Procedure

Let $\chi_1^2, \ldots, \chi_L^2$ be the χ^2-statistics at the null markers. The same type of test statistic is selected and applied to all null loci and the marker loci are tested formally for association. The inflation factor λ for the variance can then be estimated by

$$\hat{\lambda} = \frac{0.4549}{\text{median}(\chi_1^2, \ldots, \chi_L^2)}.$$

The value of 0.4549 corresponds to the median for the χ^2-distribution with 1 df. The test statistic, e.g., χ_T^2 or χ_L^2, for the marker locus of interest is then adjusted by

$$\chi_{GC}^2 = \hat{\lambda}\, \chi_L^2 \sim \chi_1^2$$

for the alleles test, and similarly for the trend test χ_T^2. For a codominant test we use the median value of a χ_2^2 distribution in the numerator of $\hat{\lambda}$.

In practice, genomic control has proven to be an effective and powerful approach to address the issue of potential confounding due to population-admixture in genetic association studies. In numerous applications, genomic control has been successfully applied and provided association results that could be robustly replicated by other studies. As with any statistical approach for population-based designs in this context, it is possible to construct scenarios in which genomic control fails to adequately adjust for population admixture, e.g., local population admixture, selection, etc. However, even in the days of genome wide association studies, genomic control is still one of the most frequently used adjustment approaches and can generally be recommended.

8.3 Modeling the Effects of Population Admixture and Stratification

An alternative, more model-based approach to compensate for the effect of population substructure is either to infer directly the population substructure based on a set of null markers (Pritchard et al. 2000) or to estimate its effect on the odds of disease (Epstein et al. 2007).

The first approach aims to identify the sub-populations in the study and, for each study subject, identify their membership in a particular subpopulation. Based on the subgroups identified, a stratified analysis is conducted. The approach requires ancestry informative markers (i.e., so called **AIM**s) that allow one to distinguish between the different subpopulations. It works particularly well if the sample consists of sub-populations that are genetically very different and this difference is captured by the AIMs. One disadvantage of such approaches is, however, that, while such marker sets exist to distinguish between different ethnicities, they are not common knowledge for most study populations likely to be present in large samples from out-bred populations. In such cases, the approach will not be able to identify subpopulations reliably. If proper attention is given to handling population admixture and stratification during the study design, study subjects who are genetically very different and subpopulations which are not apriori identified should be rare.

8.4 Regression-Based and Principal Component Approaches

A straightforward way to adjust for the effects of population-admixture and stratification is to identify markers that are informative about different strata or subgroups, and then to adjust the association analysis accordingly by including them as covariates in a logistic regression, or linear regression depending on the analyzed trait. Such approaches have been shown to work well in practice and are very flexible in terms of the statistical modeling of complex traits. The key disadvantage here, once again, is that the informative markers/null-markers need to be known and available for genotyping.

This changed fundamentally with the arrival of genome wide association studies. Having thousands of SNPs available across the entire human genome, allows identification of the informative markers for each study based on its own data. Such study-specific adjustments are highly desirable for the reasons discussed above.

In the absence of population substructure, one would expect that the genotypes between probands are not correlated. If, however, correlation is observed between the genotypes of probands this suggests that the study subjects may be *cryptically* related, i.e., they share a recent common ancestor, and that population substructure is present. Of course, with several thousand markers and a sample size of at least a few hundreds, it is not possible to examine the correlations on a proband-by-proband or marker-by-marker basis. The key idea is to construct the variance-covariance matrix of all probands and apply principal component analysis (Chen et al. 2003; Zhang et al. 2003; Zhu et al. 2002; Price et al. 2006).

By identifying the major axes of variation in the genetic markers and plotting the coordinates of the study subjects with respect to the principal components, we can examine whether the genotypes of the study subjects are correlated for the strongest components of variation in the data set. Since the principal components are a linear combination of the original genotypes, study subjects should be uncorrelated in such a plot in the absence of population substructure. However, if there are clusters in the principal component plots, this implies that the genotypes of the probands within a cluster are correlated and there is population substructure. Figure 8.3 contains a plot the first two principal components for a sample of 1000 children from the IMAGE study, a European Attention Deficit and Hyperactivity Disorder Study. Figure 8.3 clearly shows significant degree of population substructure. Note the clustering of cases in the lower tail of eigenvalue 1. Clusters of cases with no corresponding cluster of controls cannot be matched on population substructure.

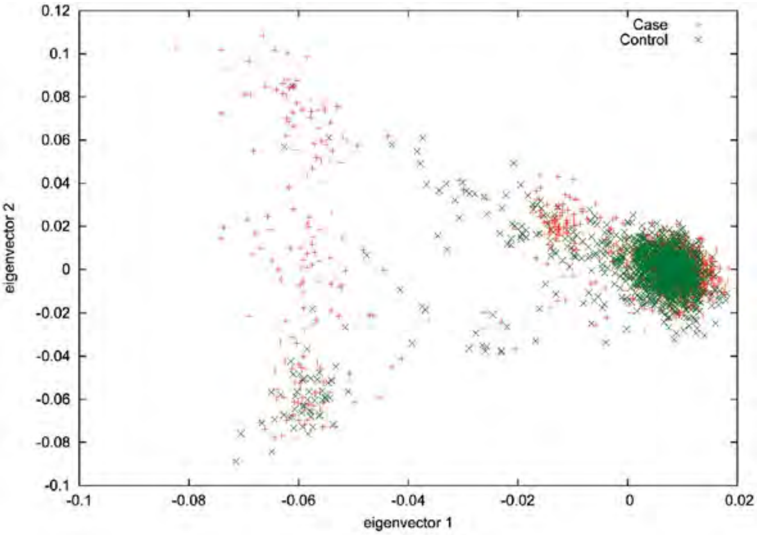


Fig. 8.3 Eigenvalue decomposition of the IMAGE sample, a study on attention deficit and hyperactivity disorder in 1,000 children sample from Belgium, Israel, France, Germany and Switzerland. Source: Courtesy of Dr. Jessica Lasky-Su

It is straightforward to adjust the analysis for such effects, either by including the significant eigenvectors of the principal component analysis in the association analysis as covariates, or by matching cases and controls based on their eigenvalue decomposition. Cases and controls that cannot be matched are removed from the analysis. The entire algorithm for principal component adjustment is outlined for the genotype adjustment in Box 8.2. With whole genome SNP coverage the vast majority of the markers should not be correlated with the phenotype under the null hypothesis, hence principal component analysis can be used to identify correlation between the phenotype data and the genetic data that is attributable to population-stratification. Including the corresponding eigenvectors in the analysis will remove such effect from the data analysis.

Box 8.2 Principal Component Adjustment of Association Studies

Notation:

- M ... Number of SNPs
- N ... Number of subjects
- $X = (z_{ij})$ an $M \times N$ matrix of standardized genotypes coded for the additive model for the i^{th} SNP in the j^{th} proband, i.e.,

$$z_{ij} = \left(X_{ij} - \bar{X}_{i.}\right) / \sqrt{\hat{p}_i(1 - \hat{p}_i)}$$

where $\hat{p}_i$ estimates the population frequency of the i^{th} SNP.

The algorithm:

- **Step 1:** Compute the Variance-Covariance matrix for the probands as $C = (X^T X)/(N - 1)$.
- **Step 2:** Compute the eigenvalue decomposition of the covariance matrix.
- **Step 3:** Select the top K eigenvalues that are statistically significant.
- **Step 4:** Include the significant eigenvectors in the linear regression of the phenotype on marker, or use the significant eigenvectors to match cases and controls, and do a matched pair analysis.

Together with the genomic control approach, the principal component approaches have became one of the standard ways to adjust for population substructure in population-based designs when thousands of markers are available. While they were less popular in the area of candidate gene studies, when only limited number of genetic markers were available for each study subject, they have become the most frequently used adjustment tool in genome wide association studies.

Although attractive in that it can theoretically handle any number of markers, it is numerically most stable to select up to 10,000 markers not in LD and with allele frequencies above about 10%. As is the case for genomic control, it is also here possible to construct scenarios in which principal component adjustment will fail to remove the effects of population admixture appropriately. However, in general, when applied carefully, they have been demonstrated to be an efficient way to control for population admixture in population-based design.

8.5 Exercises

In exercise 1 of Section 7.11, we found an association between a haplotype (GM3:5,15,14) from the human immunoglobulin G gene and IDDM in a population of Native Americans from the Pima and Papago Indian Tribes. Recall, however, from Chapter 3, that this population is admixed, as individuals have different proportions of allele frequencies depending on the number of native American ancestors. The table below gives counts of individuals with and without the GM haplotype (GM+ and GM-) and those with IDDM in parentheses for 3 of the strata:

Strata	n	# GM+ (# IDDM)	# GM- (# IDDM)
0	32	21(3)	11(2)
4	344	145(7)	199(7)
8	4264	69(12)	4195(1303)

Note that these data give different overall percentages from Table 3.2 because of age adjustment.

1. Using the data in the table above, compute the bias in the test statistics using formula (8.1). Does this suggest population stratification? Will it affect the bias, or variance, or both for the test of association?
2. Compare the relative risks computed within strata with the one computed from the data in exercise 1 of Section 7.11 and comment.

   For these next four exercises, assume the situation described in this chapter, where we have two subpopulations with allele frequencies p_1 and p_2. The proportion of cases in strata 1 is c, and the proportion of controls in strata 1 is d, K_k is disease risk, and S_k is the proportion in the k-th strata. Within strata, the allele frequencies are the same for cases and controls. In the sample, we assume an equal number of cases and controls overall ($r = s$).
3. Verify the expression given in the text for E(c–d) when there are only 2 strata.
4. Verify equation (8.1), where now we assume more than 2 strata with allele frequencies p_k and disease risk K_k.
5. Suppose you can identify the subpopulations; suggest at least one way to obtain an adjusted difference between cases and controls, or to obtain an unbiased test for the difference in mean X in cases and controls. You can consider both design and analysis strategies.
6. Show that (under H_0) the covariance between X_l and X_m for any two individuals l and m drawn from the same subpopulation is $4Fp(1-p)$, where F and p are defined for equation (8.2) in the text.

7. In a candidate gene study, 20 null markers have been genotyped. Their χ^2 statistics are listed below:
5.112124234 0.827057943 3.158134984 3.395351358 0.056900096 0.878446231
4.955161751 0.127185994 1.115390624 1.471334371 0.042577497 0.833171588
0.389633293 0.088260639 0.008057015 0.206122142 0.052385560 0.020823177
1.445823813 0.195321095
The table for the marker of interest is given by:

	X= 0	X=1	X=2
Cases	400	150	52
Controls	321	120	40

(a) Is there evidence for admixture in the data?
(b) What is the genomic control adjustment factor?
(c) Is the marker of interest associated with affection status? Use both the alleles test and the trend test, adjusted for Genomic Control.

Chapter 10
Advanced Topics

In this chapter we review specialized and advanced topics that are beyond the scope that can be covered in detail in an introductory text book. However, the topics are important research areas and the interested reader is encouraged to follow-up our brief introduction with the specialized literature.

10.1 The Multiple Testing Problem in Association Studies

In this section we consider testing a relationship between multiple genetic loci and disease phenotypes, when multiple SNPs have been genotyped; this arises in the context of a candidate gene or genomic region study, or in the GWAS setting, across the entire human genome. We assume that the nature and the locations of the DSLs are unknown. For any association study with multiple genotyped loci, but particularly for GWAS studies, the multiple testing problem is one of the major statistical hurdles that has to be addressed in the analysis of the study. In a typical GWAS analysis more than 500,000 markers are tested for genetic association, creating a substantial multiple testing problem. There are three broad categories of approaches to the multiple testing problem: (1) testing each marker separately and adjusting the significance levels of each test, (2) using permutation or re-sampling techniques, and (3) for small number of SNPs in a defined region, using haplotypes or simultaneous test strategies. The best approach will depend upon many factors, primarily the unknown nature of the true DSLs, but convenience and feasibility play a big role in what approaches are commonly used. Methods for efficient selection of SNPs will be discussed in Section 4.

10.1.1 Methods Based on P-Value Adjustment

Multiple testing is a common statistical problem, and many approaches have been suggested to handle the inflation of the error rate. An easy and popular approach to handling multiple SNP testing is to test each SNP separately and then adjust the significance-level of each test so as to preserve the overall error rate. The Bonferroni

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_10,

method is perhaps the most widely used approach. The idea can be formalized as follows: Let M denote the number of markers for testing, and for each $m = 1, \ldots, M$, we can define the null hypothesis $H_0^{(m)}$: No association between the mth SNP and the disease phenotype. In the single test setting, we set our significance level, α', to satisfy

$$\alpha' = P(\text{reject null hypothesis } H_0^{(m)} \mid H_0^{(m)} \text{ is true}). \tag{10.1}$$

But in testing multiple SNPs, our interest is in the experiment-wise α-level, or family wise error rate (FWER), defined as

$$\alpha = P(\text{rejecting at least one } H_0^{(m)} \mid H_0^{(m)} \text{ is true for all } m). \tag{10.2}$$

We denote α as the FWER. A simple method of choosing individual significance-levels to fix the FWER, say, at the desired level is to set the individual significance level $\alpha' = \alpha/M$. This is known as Bonferroni adjustment or Bonferroni-correction; it follows from Boole's inequality:

$$\begin{aligned} &P(\text{at least one rejection} | H_0^{(m)} \text{ is true} \,\forall\, m) \qquad (10.3)\\ &\le \sum_{m=1}^{M} P(H_0^{(m)} \text{ is rejected} | H_0^{(m)} \text{ is true} \,\forall\, m). \end{aligned}$$

When the significance levels of the individual tests are all equal to α', the above reduces to

$$\alpha \le M\alpha'. \tag{10.4}$$

Thus the FWER can be kept less than α if we test each individual test with significance level α/M. Notice that the Bonferroni adjustment method makes no assumption about the independence of the events. In fact, if the association tests are correlated, as we might expect them to be if there is much LD between the markers, then the Bonferroni method is very conservative, and the degree of conservatism increases as M increases. In the extreme, where rejection of one test implies rejection of all the rest, then the true FWER is α/M.

Holm (1979) has suggested a modification to the multiple-testing procedure that is uniformly more powerful than the Bonferroni procedure. The idea is as follows. Order the p-values, and compare the smallest to α/M. If that test rejects, test the next smallest against the level $\alpha/(M-1)$. Continue in this way until a test accepts, then declare all of the smaller p-values significant, and the rest not significant. It is clear that the Holm procedure is more powerful than the Bonferroni procedure because whenever the Bonferroni procedure rejects, the Holm procedure will also, but the reverse is not true. For a very large number of SNPs, $M - j$ is approximately equal to M for small to modest j, so that the Holm procedure will not be very different from the Bonferroni, but with small or modest number of markers M, it can substantially increase the overall power.

For application in situations where some prior information may be available, notice that equation (10.3) implies that we can assign each individual test different rejection levels, say α_m. To fix the overall α at a desired level, we then require that the sum of the individual α_m levels does not exceed α. Thus some tests can use rejection levels higher than α/M, and others lower, as long as (10.3) is retained. This gives rise to the weighted Bonferroni approach, whereby prior information may be used to prioritize SNPs. In the context of multiple SNPs in a gene, it may be desirable to assign higher weight (or relatively bigger α-levels) to SNPs in coding regions, than those which are not. This argument extends to GWAS as well. In GWAS, one might up-weight regions with prior information about potential DSLs, e.g., linkage studies, candidate gene studies, while regions without any known genes might be down-weighted. Various weighted Bonferroni approaches have been suggested for genome wide association scans. See, for example, Genovese et al. (2006); Roeder et al. (2007); Ionita-Laza et al. (2007). All of the multiple testing procedures can be applied either in the population based, or family setting.

A less conservative FWER testing procedure was suggested by Simes (1986). This test is also based on ordered p-values, but uses as a rejection criterion: reject H_0^i if the inequality

$$p_{(i)} \leq \frac{i}{M}\alpha$$

holds. The method is simple to apply and less conservative than the Bonferroni. The type-1 error is preserved at α provided the tests are independent.

As an alternative to the FWER approach, one can also use the *False Discovery Rate (FDR)* to adjust for the multiple statistical tests. Rather than to control for the type-1 error, FDR limits the expected number of null-hypotheses that are rejected incorrectly. Fundamentally, FWER evaluates the error rate by assuming all null hypotheses are true, while the FDR evaluates the error rate based solely on fixing the number of erroneous rejections, regardless of the number of true hypotheses. In comparison to the FWER, the FDR approach is less conservative and therefore, by definition, more powerful. Assuming that we compute M independent statistical tests, the test results can be divided as diagrammed in Table 10.1.

It follows that $M = U + V + S + T$. Then the false discovery rate is given by

$$E\left(\frac{V}{V+S}\right). \tag{10.5}$$

Table 10.1 Number of errors committed when testing M null hypotheses

	Declared non-significant	Declared significant	Total Total
true null hypotheses	U	V	M_0
non-true null hypotheses	T	S	$M - M_0$
	$M - R$	R	M

One way to control the false discovery rate is to use a modified Simes-procedure. For a pre-specified false discovery rate of α, we search for the largest i such that the inequality

$$p_{(i)} \leq \frac{i}{M}\alpha.$$

We then reject the null-hypothesis for all tests that correspond to the p-values $p_{(1)}$ to $p_{(i)}$, while the false discovery rate will be maintain at α. Similar to FWER which was originally also derived only for a set of independent test statistics, the FDR approach has been extended to incorporate correlation between the test statistics (Benjamini and Yekutieli 2001). An alternative, Bayesian interpretation of the FDR approach is discussed in Storey (2002, 2003).

Although all of these approaches provide meaningful increases in statistical power, the standard Bonferroni-correction remains the most commonly-used adjustment principle in genetic association studies. Given the many false-positive findings in the history of genetic association studies, one rather errs in being too conservative when controlling the FWER than to maximize the statistical power in order to avoid false-negative findings.

10.1.2 Permutation and Monte Carlo Tests

The terms *permutation, randomization or re-randomization tests* are often used interchangeably, and can be thought of as special case of general resampling tests. Monte Carlo refers to approximate permutation tests when the number of permutations of the data is too large to systematically evaluate them all. The idea behind these procedures is to use the observed data to simulate the distribution of the test statistics under the null. Often the justification for these test procedures is that they do not require parametric assumptions about the data. In our setting, it is often infeasible to compute a parametric distribution for the test statistic. To take a simple example, consider how we might evaluate the significance of the minimum observed p-value, or if all test statistics have the same distribution under global null, the maximum test statistic. The approach is quite straightforward in the case-control setting. Let χ^2_m denote the χ^2 statistic for testing the mth marker, and denote the ordered statistics by

$$\chi^2_{(1)} \leq \chi^2_{(2)} \leq \cdots \leq \chi^2_{(M)}. \tag{10.6}$$

Then to evaluate the appropriate p-value for the maximum χ^2, we note that if there is no association between any SNP and case-control status, the joint distribution of p-values can be computed as follows.

- Randomly assign each person in the study a case-control status, ensuring only that the total number of cases and controls is fixed.
- Calculate $\chi^2_{(M)}$ from this simulated sample.

- Repeat this procedure a large number of times. Then the randomization p-value for the observed $\chi^2_{(M)}$ is the proportion of simulated $\chi^2_{(M)}$'s which exceed the observed.

The advantage of the randomization procedure is that it is intuitive and it extends readily to calculating more complex p-values, such as the p-value of $\chi^2_{(M)} + \chi^2_{(M-1)}$, etc. It is not restricted to case-control designs or even dichotomous outcomes. In the case of measured outcomes, the traits can be randomly assigned to marker scores. The only requirement is that the trait which is randomized has the same distribution for each person under the null. In the setting where there are important covariates which predict disease traits, and may be related to genetic status as well, this assumption is questionable. Likewise, we may need to adjust samples for population substructure. Thus for some traits it may be necessary to use covariate adjustments and randomly assign adjusted traits (residuals).

Resampling tests generally outperform its competitors in terms of power Roeder et al. (2005), however major limitation of these tests is computational efficiency. If the sample is sufficiently small, all possible permutations can be enumerated, and one can get an exact p-value. Generally, however, Monte Carlo simulations will be required. Depending upon sample size and data complexity, tens of thousands of simulations will be required for p-values to stabilize. For only a modest number of SNPs and moderate sample sizes this will not be a problem, but computations can be numerically intensive for large scale studies (Sheskin 2004).

The applicability of randomization tests to family designs depends on the design. If we have only affected offspring, then reshuffling outcomes changes nothing. With discordant sib pairs, the assignment of affection status can be interchanged randomly with each pair, and inference proceeds as above. Likewise, with trios and measured outcomes, the traits can be randomly assigned to offspring. Concerns about covariate effects are less of a concern, provided the covariate is independent of the genotype at the gene, conditional on the parental genotype (Chapter 9).

10.2 Other Methods for the Analysis of Multiple SNPs, Including Haplotypes

In this section we consider the use of haplotypes and simultaneous testing that can be used with a relatively modest number of SNPs.

The HapMap project was motivated by the remarkable observation that, although the LD patterns between markers that are several mega-bases apart seem to be very random and 'erratic' markers that are relatively close to each other (10–100 kilobases) exhibit a very discrete LD-structure. For the most part, such markers are in high LD with each other and recombination events seem to mainly occur only at certain, so-called recombination hot-spots. Figure 10.1 illustrates this observation. We observe that the haplotype diversity, i.e., the number of possible combinations of marker alleles that resides on the same block of the chromosome, is very limited. For a relatively limited portion of the chromosome, the haplotype diversity is small

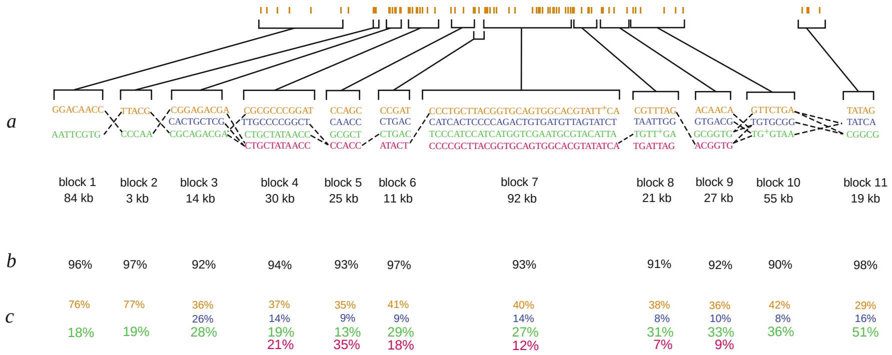


Fig. 10.1 Haplotype block structure for a region at 5.q.31. (**a**) Common haplotype patterns in each block of low diversity. Dashed lines indicate locations where more than 2% of chromosomes are observed to transition from one common haplotype to a different one. (**b**) Percentage of observed chromosomes that match one of the common haplotype patterns exactly. (**c**) Percentage of each of the common patterns among untransmitted chromosomes. *Source*: Daly et al. (2001)

compared to the possible number of distinct haplotypes, i.e., 2^M, where M is the number of SNPs across the region. Due to the reduced haplotype diversity, it is theoretically possible to distinguish the haplotypes uniquely based on only a subset of the markers that are included in the haplotype. Such SNP sets are referred to as haplotype-tagging SNPs or Tag-SNPs. For many genomic regions, the number of SNPs that is required to capture the local LD-structure perfectly is large, especially when the number of rare haplotypes in the region is high. One therefore typically defines cutoff criteria for the LD-resolution that the set of Tag-SNPs should achieve. For example, one could select Tag-SNPs that allow the unique identification of all haplotypes with a frequency of at least 5% or, alternatively, Tag-SNPs that have at least an r^2 of 80% with all SNPs that will not genotyped in the region.

Using LD information from reference populations, e.g., HAPMAP, several algorithms have been developed to identify the smallest possible sub-set of Tag-SNPs for a particular region in order to reduce the genotyping cost of the study. We refer the reader to De Bakker et al. (2006) and Wang et al. (2006) for a review and comparison of different tagging approaches. With the recent drop in genotyping costs and the arrival of genome wide SNPs chip for which such LD-consideration have been taken into account during the design of the SNP-chip when the SNPs for the chip were selected, the topic of Tag-SNPs selection is of less current interest.

In the context of association analysis, the observation of reduced haplotype diversity is remarkable in several aspects. Instead of genotyping and testing a large number of SNPs in a region, an alternative strategy is to test haplotypes. In order to infer the haplotypes, only the subset of markers that defines the haplotypes uniquely has to be genotyped, not all markers that are in the haplotype region. Although a much smaller number of SNPs is genotyped with this strategy, all genetic variation in the region is covered. Since, for most regions of the human genome, the number of observed haplotypes with frequencies of above 5% is much smaller than the number of SNPs in the haplotype block, haplotype analysis can be more efficient than a single SNP analysis in terms of the multiple testing problem. In genetic association analysis, individual SNPs or haplotypes constructed from these SNPs can be used to test for association with the phenotype of interest. In a haplotype analysis, the haplotype takes the role of the ‘allele’ that defines the genotype in the association test. One can then test either each haplotype separately and adjust the results for the number of tested haplotypes, or test all haplotypes jointly with a multivariate test in which the variance-covariance structure between the haplotypes has to be estimated. The degrees of freedom for such an overall haplotype test are given by the rank of the variance-covariance matrix for the haplotypes. As for single marker tests, the power of a haplotype association tests will be limited for small haplotype frequencies and, consequently, the exclusion of haplotypes with frequencies of less than 5% is often recommended.

Motivated by such advantages, haplotype analysis has become a standard analysis tool for genetic association studies. However, in practice, the analysis faces a major limitation: the haplotypes usually cannot be determined directly. Rather the genotypes of the markers are observed, but not the phase of the haplotype, i.e., the alleles that reside on the same chromosome. The phase of the haplotype has to be inferred based on the genotype data in order to implement an association analysis. A variety of approaches have been developed to address this problem. The key idea of such analysis strategies is to treat the unobserved phase of the haplotype as a missing data problem. Applying missing data techniques such as the EM-algorithm (Dempster et al. (1977); Slatkin and Excoffier 1996) to the marker data, the distribution of the phased haplotypes can be computed for a given set of genotypes in a study subject. Since the phase of the haplotype cannot be inferred with certainty, the additional variability has to be accounted for in the association analysis. Consequently, the haplotype uncertainty will reduce the statistical power of the genetic association test, partly diminishing the theoretical advantages of haplotype analysis.

For population-based studies, a general framework for haplotype testing, termed haplo-score has been proposed (Schaid 2001). A score-test is constructed based on a generalized linear model (equation (2.3)) which allows for binary, count or quantitative traits. The trait is modeled as a function of the possibly unknown haplotype; reducing the problem to one of ordinary generalized linear regression with missing covariates (Ibrahim 1990).

The approach provides either an overall test, testing all haplotypes simultaneously, or a haplotype-specific association test between the phenotype and the selected haplotype. For family-based designs, the FBAT-approach can be generalized to accommodate haplotype analysis as well (Horvath et al. 2004), provided that we assume that the SNPs are sufficiently close so that no recombination has occurred. The key ingredient necessary to construct family-based tests based on Mendelian transmissions from the parental generation to the offspring, requires the knowledge of the phase of the haplotype. Since it will not always be possible to reconstruct the phase of the haplotype, the extension of the FBAT-approach to haplotypes conditions on the sufficient statistic for the phase in the parental genotypes family, thereby allowing the analysis of haplotypes in family-based designs in full generality. Under the assumption of no recombination, having parents makes it easy to reconstruct phase, and family-based haplotype analyses are quite efficient, but become infeasible with missing parents and large numbers of SNPs (Rakovski et al. 2007).

Haplotype analysis has been replaced more and more by single marker tests or simultaneous tests which do not require a reconstruction of the phase. This is partly due to the fact the typical density of SNPs, as they are genotyped in association studies these days, has increased so much that they provide sufficient coverage of the common, ungenotyped variants in the region. Another contributing factor is that it is still unclear if haplotype testing is more powerful than SNP testing (Roeder et al. 2005; Rakovski et al. 2007). Clearly, power will be influenced by the nature of the DSL; if the DSL is a genotyped SNP, then SNP testing will be preferred; if it is a combination of SNPs lying on a haplotype, then haplotype testing should be more powerful. Other cases are less clear, and will depend upon the number of SNPs available and the pattern of LD across the region.

Another approach is simultaneous testing. The idea behind simultaneous testing is to use a multivariate test which tests all M null hypotheses simultaneously. The approach we describe here is sometimes also called the 'multi-marker' test, the 'regression' test, or 'locus' scoring test, the latter term used to distinguish it from tests using haplotypes. An advantage of the simultaneous method is that the haplotype-block structure does not have to be known; the correlation between the markers is estimated directly and the validity of the test does not depend on the correct specification of the haplotype-block structure.

Perhaps the most well known multivariate test is Hotelling's T^2, designed to test whether or not the means of M normally distributed variables are equal in two groups; it is simply an extension of the simple t-test in the univariate setting. To motivate the multimarker test we first give a quick review of Hotelling's T^2. Suppose we have M traits for a sample of N subjects classified into two groups. For simplicity, we consider equal allocation with n subjects per group. Our null hypothesis is H_0 : $\Delta_1 = \Delta_2 = \cdots = \Delta_M = 0$, where Δ_m is the difference in means of the mth trait between the two groups. Hotelling's T^2 statistic is given by

$$T^2 = n\mathbf{D}'\mathbf{S}^{-1}\mathbf{D}/2$$

where **D** is a vector whose mth element is the difference in the sample means of the mth variable in the two groups, and **S** is the sample variance covariance matrix of the M variables pooled over the two groups. A multiplicative factor can be applied to T^2 to give an F-distribution under the null. In large samples it is distributed approximately as a χ^2 with degrees of freedom equal to the rank of S, even without the normality assumption.

In the setting of a case-control study, the two groups being compared on the M SNPs are cases (Y=1) and controls (Y=0). Let $X_i^{(m)}$ denote the number of minor alleles for the mth SNP in each subject, for $i = 1, \ldots, M$. Using a typical logistic regression model to relate the SNP data to disease, we can write:

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \cdots + \beta_M X_i^{(M)},$$

or equivalently,

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i$$

where $\boldsymbol{\beta}$ is a vector of the M regression coefficients and $\mathbf{X}_i$ is a vector of the M genotype scores for the ith individual. In this framework, we can test $H_0 : \boldsymbol{\beta} = 0$ using a likelihood ratio or a score test. The numerator of the score test takes a particularly simple form:

$$\mathbf{U} = \bar{\mathbf{X}}_{\text{cases}} - \bar{\mathbf{X}}_{\text{controls}},$$

i.e., each element of $\mathbf{U}$ is the numerator of the Alleles Test for the corresponding SNP. Under the null hypothesis that the distribution of the SNPs are the same for both groups, the variance of $\mathbf{U}$ can be estimated by the sample variance-covariance matrix of the $X_i's$, say $\mathbf{S}$.

As with Hotelling's T^2, in large samples $n\mathbf{U}'\mathbf{S}^{-1}\mathbf{U}/2$ is approximately χ^2 when H_0 is true, with the degrees of freedom depending upon the rank of S. If the SNPs are highly correlated, the rank may be less than M; in this case it is necessary to use a generalized inverse for $\mathbf{S}$. The derivation in Clayton et al. (2004) shows that the score test approach is easily generalized to give a similar test when Y_i is measured and we assume a linear rather than a logistic model.

The multi-marker test also generalizes to the family design in a straightforward way. Letting X_{ij} denote the vector of M marker scores for the jth subject in the ith family, the natural generalization for either a dichotomous or a measured trait is

$$\mathbf{U} = \sum_{i,j} (T_{ij} - \mu)(\mathbf{X}_{ij} - E(\mathbf{X}_{ij})|S_i)$$

where the expectation of each element of $\mathbf{X}_{ij}$ is calculated conditional on the sufficient statistics for parental genotypes only at the specific marker, and T_{ij} and μ are the trait and the offset, respectively. Computing each diagonal term of var($\mathbf{U}$) conditional on the sufficient statistics for parental genotype is straightforward, but computing the covariance terms is more difficult because they depend on the joint distribution of two SNPs. When SNPs are in LD, this requires knowledge of the pairwise haplotype distributions. Rakovski et al. (2007) use an empirical variance-covariance matrix, similar to that used for adjusting the variance when testing candidate genes under a linkage peak, which does not require knowledge of the haplotype distribution.

Such multi-marker tests are often conducted in gene-based settings, i.e., all typed markers in a gene are tested for association with the phenotype of interest. While such an approach has the advantage that it is computationally simple, it becomes less appealing as the number of markers becomes large. Approaches that apply principal component analysis or canonical correlation analysis to the markers in the selected region have also been suggested to reduce the dimensionality of the multivariate score tests (Kwee et al. 2008; Gauderman et al. 2007).

10.3 Gene–Environment/Gene–Drug Interaction

These are numerous, well documented, Mendelian disorders where environmental conditions are known to influence the effect of the gene on disease, for example, PKU and diet. Ottman (1990) gives several examples where the biology of the interaction is understood. Gene-environment interactions are thought to play an especially important role as effect modifiers for many complex diseases. For example, gene-smoking interactions are believed to influence the disease risk for Lung Cancer, Asthma and Chronic Obstructive Pulmonary Disease (COPD). The spectrum of gene-environment interaction variables is large and can include variables that reflect a more general assessment of the environment that have no direct biological link with the disease phenotype. For example, the International Multicenter ADHD Genetics (IMAGE) project recruited 960 children with ADHD and their parents for a study of candidate genes for ADHD. Since Socio-Economic Status (SES) is an important predictor for ADHD and for ADHD symptom scores, possible interactions between the candidate genes and SES were of interest. Figure 10.2 illustrates a potential gene-environment interaction between Socio-Economic Status (SES) and a SNP in the BDNFa44 gene on Hyperactive-Impulsive Symptoms. While for the common homozygous and heterozygous genotype there is no effect of SES on symptom scores, SES has a strong effect on symptom scores for the rare homozygous genotype (Lasky-Su et al. 2007).

A special case of gene-environment interactions are gene-drug interactions. Today, many clinical trials collect DNA samples from the patients enrolled in the trial, to determine if the response to treatment can depend upon the genetic profile of the patient. Gene-drug interactions may enable the identification of subgroups of patients which, based on their genotypes, may respond more or less favorably to treatment. In the same way, genotypes could potentially identify patients with more or less severe side-effects to the treatment than in the general population. There are

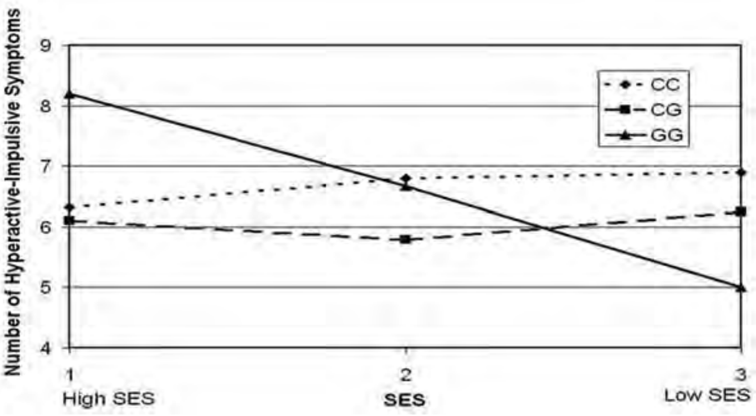


Fig. 10.2 Gene-environment interaction for attention deficit hyperactivity disorder. *Source*: Courtesy of Dr. Jessica Lasky-Su

major ongoing efforts to use genotype profiles to improve the safety and efficacy of a drug (Giacomini et al. 2007).

While gene-environment interactions can be studied in the context of linkage analysis and, even, in aggregation and segregation analyses, the recent focus of gene-environment interaction analysis has been in association studies. With population-based studies, using the regression approach to the analysis of association allows the gene-environment interaction to be integrated into the association analysis straight-forwardly by including an additional term in the regression model, e.g.,

$$g(E(Y)) = \beta_0 + \beta_1 \times X + \beta_2 \times E + \beta_3 \times X \times E, \tag{10.7}$$

where Y denotes the phenotype of interest, X the coded marker score and E the environmental exposure variable. The function g is the link-function that is selected based on the trait type of the phenotype. The parameter β_1 corresponds to the genetic main effect, β_2 is the environmental main effect and β_3 the gene-environment interaction. Under the null-hypothesis that $\beta_3 = 0$, the effects of the gene and the environment are additive on the scale specified by the link-function g (e.g., linear or log-linear). We remark that equation (10.7) is strictly speaking only valid if X models the true disease variant at the locus. Valid tests can still be constructed when X codes for a SNP in LD with the true DSL, but as in the case of main effects, it should not be used for estimation unless X codes for the true DSL.

As with any statistical interaction analysis, it is important to note that this approach tests only for a statistical interaction, but not for a necessarily biological interaction which is here understood to mean the environmental variable directly alters the biological action of the gene. While a biological interaction implies the presence of a statistical interaction, the opposite is not necessarily true (Cordell 2002).

Furthermore, the definition of a statistical interaction is scale dependent. To illustrate this concept, we assume that a dichotomous phenotype is given and that the genotype variable X and the environmental variable E are also dichotomous. Let $R_{X,E}$ denote the relative genetic risk given X and E, i.e., $R_{X=x,E=e} = P(Y = 1\,|X = x, E = e)/P(Y = 1\,|X = 0, E = 0)$. Under the null-hypothesis, the relative risk on a linear scale, i.e., $g(E(Y)) = E(Y)$, is given by

$$R_{X=1,E=1} = R_{X=1,E=0} + R_{X=0,E=1} - 1 \tag{10.8}$$

and on a log-linear or multiplicative scale, i.e., $g(E(Y)) = \log(E(Y))$ is given by

$$\log(R_{X=1,E=1}) = \log(R_{X=1,E=0}) + \log(R_{X=0,E=1}). \tag{10.9}$$

Departure from the additive model on each scale implies a statistical interaction. It is easy to see that presence of a statistical interaction on the linear scale can be consistent with no interaction in the multiplicative scale and vice versa.

While gene-environment interactions in population-based designs can be analyzed in a relatively straight-forward manner by including the interaction into the corresponding regression model, family designs present additional difficulties. One might think of extending the FBAT statistic by using as a trait $Y \times E$, giving the numerator of the FBAT statistic as

$$(Y \times E - \text{offset})(X - E(X\,|P\,)), \tag{10.10}$$

where P denotes the parental genotypes. However, such an approach will not generally be valid. The reason is the conditioning on the parental information in combination with the assumption of Mendelian transmission from the parents to the offspring. In the presence of a main genetic effect ($\beta_1 \neq 0$), the transmission of the alleles from the parents to the offspring will also depend on the phenotype Y even under H_0. Under these circumstances, it difficult to estimate the transmission probabilities without making model assumptions that would give up the robustness properties of the FBAT approach. Several ways around this have been suggested. One way is to construct an overall test statistic that tests the joint-null hypothesis of no main effect ($\beta_1 = 0$) and no interaction ($\beta_3 = 0$). It is relatively easy to extend the general FBAT approach to handle this, e.g., FBAT-GEE (Lunetta et al. 2000; Lange et al. 2003a). However, this has the drawback that if the null-hypothesis is rejected, one cannot conclusively infer the presence of an interaction; rejection can be driven by a main effect alone.

For simple cases, e.g., affected offspring in nuclear families, by assuming a log-linear model for the relative risk, one can show that $E(X|Y = 1, P, E)$ is independent of E, provided X and E are conditionally independent given the sufficient statistics for parental genotype, P (Umbach and Weinberg 2000). Hoffmann et al. (2009) extended this result to conditioning on the sufficient statistics for parental genotypes. In this case, a simple interaction test statistic is given by

$$\Sigma(Z - E(Z|Y = 1, S))(X - E(X|Y = 1, S)), \tag{10.11}$$

where summation is over all affected offspring. Both expectations in equation (10.11) and the conditional variance of X can be evaluated empirically, thereby circumventing the need to specify a genetic main effect model or to estimate environmental effects (Lake and Laird 2004; Hoffmann et al. 2009).

While this maintains the robustness of family-based association tests, it is not feasible for quantitative phenotypes and complex exposure variables. For such application, extensions using causal inference methodology have been suggested (Vansteelandt et al. 2008). In general, the development of statistical approaches for the gene-environment analysis in family-based designs is active field of research and many questions, e.g., optimal designs, still need to be addressed.

***Compositional Epistasis and Compositional Gene-Environment Interactions*:** In the previous section, it was noted that interaction was scale dependent; interaction could be present on one scale but absent on another. For the model in (equation 10.7), if $g(E(Y)) = E(Y)$ and β_3 is non-zero then one would say that a statistical interaction is present on the additive scale; if $g(E(Y)) = \log(E(Y))$ and β_3 is non-zero then one would say that a statistical interaction is present on the log-linear or multiplicative scale. Similar concepts apply also when considering gene-gene interactions. It was also noted in the previous section that a statistical interaction need not necessarily imply an interaction in a biological sense.

A more biologically oriented form of interaction between two genes was described by Bateson (1909) as 'epistasis' in which the effect of one genetic factor would be masked unless another genetic factor was also present. Currently 'epistasis' is often used synonymously with 'gene-gene interaction' and thus some authors have proposed using the term 'compositional epistasis' (Phillips 2008) to refer to epistasis in Bateson's sense of the term. Suppose that two genetic factors, X_1 and X_2, are binary indicators for genotypes at loci A and B respectively so that $X_1 = 1$ denotes a high risk variant at locus A and $X_2 = 1$ denotes a high risk variant at locus B. If Y is a dichotomous trait then compositional epistasis (epistasis in the sense of masking) would be present if for some individuals, their outcome under different values of X_1 and X_2 would be that given in Table 10.2 so that for individuals with the low risk variant ($X_1 = 0$) at locus A, X_2 has no effect on the outcome.

Such examples of compositional epistasis do not in general correspond to interaction terms in a statistical model. There are, however, relations between a statistical model such as (equation 10.7), with X and E replaced by X_1 and X_2, and compositional epistasis that can sometimes be used to empirically test for compositional epistasis. If there is no confounding of the effects of X_1 and X_2 on the outcome Y so that the probabilities of the outcome conditional on X_1 and X_2 reflect the true causal effects of these genes then if $g(E(Y)) = E(Y)$ and $\beta_3 > 2\beta_0$ it can be shown that there must be some individuals in the population who have a response pattern like that given in Table 10.2, i.e., compositional epistasis must be present (VanderWeele 2010). Under some additional assumptions, weaker tests can be used. If it can be assumed that for at least one of the two genetic factors, X_1 or X_2, its effect is never preventive for any individual (i.e., a change from 0 to 1 would never change the outcome from 1 to 0 for any individual) then $g(E(Y)) = E(Y)$ and $\beta_3 > \beta_0$ will imply individuals with a response pattern like Table 10.2. If for both of the genetic factors, X_1 and X_2, their effects are never preventive for any individual then $g(E(Y)) = E(Y)$ and $\beta_3 > 0$ will imply individuals with a response pattern like Table 10.2.

If a log-linear model with $g(E(Y)) = \log(E(Y))$ is used then $\beta_3 > \log(3)$ will imply compositional epistasis (i.e., individuals with response pattern as in Table 10.2) provided that both main effects β_1 and β_2 are non-negative. If for at

Table 10.2 Example of compositional epistasis: The outcome Y for a particular individual under different combinations of X_1 and X_2

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0	0
$X_1 = 1$	0	1

least one of the two genetic factors, X_1 or X_2, its effect is never preventive for any individual then $\beta_3 > \log(2)$ will imply compositional epistasis provided that both main effects β_1 and β_2 are non-negative. If for both factors, X_1 and X_2, their effects are never preventive for any individual then $\beta_3 > 0$ will again imply compositional epistasis. Similar results hold if the genetic factors have more than two levels (VanderWeele 2010). These tests for linear or log-linear models could also be applied to cases of gene-environment interaction in order to test for 'compositional gene-environment interaction.'

Note, however, that these tests for compositional epistasis, although somewhat more biologically motivated, still do not necessarily imply physical molecular interaction between various proteins, which Phillips (2008) referred to as 'functional epistasis'.

10.4 Exercises

1. Show that the presence of a statistical interaction on the additive scale is consistent with no statistical interaction on the multiplicative scale and vice versa.
2. Compute the Bonferroni correction for the SNPs listed in Table 7.11 and decide which SNPs achieve overall-significance
3. Using the FDR, for which of the SNPs listed in Table 7.11 do you reject the null-hypothesis?

Chapter 11
Genome Wide Association Studies

11.1 Introduction

The key requirement for genetic association, linkage disequilibrium (LD), is a short distance property that extends only for a limited physical distance across the human genome. As we showed in Chapter 7, if there is low LD between the genotyped marker and the DSL, there will be low power to detect association between the disease and the DSL. In the early years of association testing, the strategy was mainly used to test specific regions, e.g., genes which were selected on the basis of function relative to the biology of the disease, or on the basis of linkage analysis. By restricting testing to a small enough region, markers can be selected for testing which should be in LD with the DSL anywhere in the region. In particular, SNPs in the coding region of a gene are often chosen as markers. With Genome Wide Association Studies (GWAS) the idea is instead to cover the entire genome with a sufficiently dense set of SNPs that all untyped polymorphsims (including DSLs) are in reasonably high LD with a tested SNP. For this reason, GWAS studies are sometimes called ‘unbiased’ because every region of the genome is searched, not just those meeting determined selection criteria.

Whether testing one or several genes, or the entire genome, the selection of the SNPs that are genotyped in a genetic association study is a crucial decision that defines the likelihood for the success of the study. In principle, the SNPs that will be genotyped in a GWAS, should be selected so that they are sufficiently correlated, i.e., in strong linkage disequilibrium, with the SNPs that will not genotyped in the same region. Thereby, the presence of a DSL in the region, independently of whether the DSL was genotyped or not, can be detected with an association test.

In the early 2000s, when genetic association studies became one of the most popular mapping tools, the local LD structure was unknown for most parts of the human genome and researchers had to assess the local LD structure in the genomic region of interest prior to the actual association study. For each study, it was necessary to genotype a small subset of subjects at a large number of SNPs in the genomic region in order to study the LD patterns and to define the suitable set of tag SNPs for genotyping. This was a labor-intensive process with a high degree of redundancy as the same genomic area was often studied by more than just one group.

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2_11,

This led to the HapMap project (International HapMap Consortium, The 2003, 2005, 2007), a concerted effort to centralize this work and to provide the scientific community with a comprehensive LD-catalog for the entire human genome across for four different ethnicities (Section 5.1). Several genome-centers and companies across the world genotyped millions of SNPs in the four different HapMap populations. The observed LD patterns were cataloged and made accessible to the scientific community via the Internet. This large-scale effort went side-by-side with major improvements in the genotyping technology. The genotyping expenses dropped to a fraction of the original costs, and *SNP-chip*s were developed which made it possible to genotype several hundred thousand SNPs across the human genome. For the current SNP-chip generation, the LD-information of the HapMap project has been incorporated so that the chips provide adequate coverage of the entire human genome for most ethnicities, i.e., SNPs that are not included on the SNP-chip are in strong LD with at least one genotyped SNP ($r^2 > 0.7$ or 0.8). Recent trends in the SNP-chip development show a trend to more *gene-centric* approaches, i.e., saturating known genes with densely spaced SNPs.

With the arrival of SNP-chips, it became possible to search for disease loci across the entire human genome, using the concept of indirect association. A study in which hundreds of thousands of SNPs are genotyped across the genome and tested for association with the phenotype of interest is referred to as a *Genome Wide Association Study*. For numerous complex diseases, they have led to the discovery of novel associations between genetic loci and disease phenotypes that can be reproduced and replicated robustly in other populations and studies (Manolio et al. 2008, Fig. 11.1). While this technological development offers great scientific potential to understand the genetic architecture of complex diseases, it creates substantial statistical challenges. We will discuss two of the main challenges in the next two sections: the development of statistical filters to ensure sufficient genotyping quality and the handling of the multiple testing problem.

11.2 Quality–Control for the Genotype Data

As in any statistical analysis, the quality of the data is one of the deciding factors that defines the validity of the findings and the conclusion. Quality control and plausibility checks of the data are compulsory parts of any statistical analysis. While selection and assessment of disease outcomes are crucial in genetic association studies, the issues are largely generic to any epidemiology study and here we focus on genotyping quality. The process of genotyping is technically complex and the genotyping quality generally depends on multiple factors that are not always under the control of the investigators. Such factors include the quality of the DNA in the sample, depending in turn on the type of sample (blood vs. buccal swab), the handling and storage of the sample, and the genotyping platform, etc. After the genotyping process is completed and all platform specific error checks have been performed, additional statistical quality control steps are needed to ensure the validity of the data.

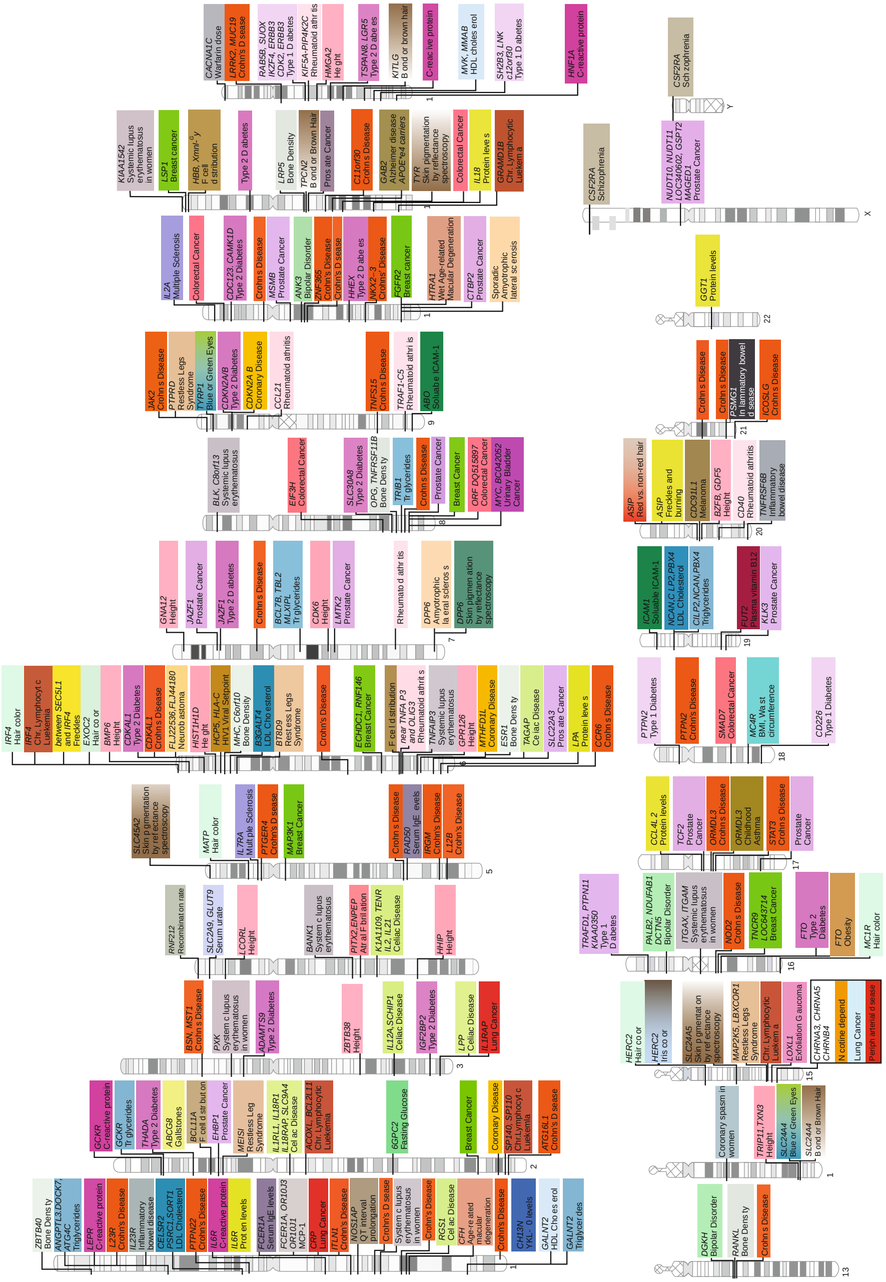
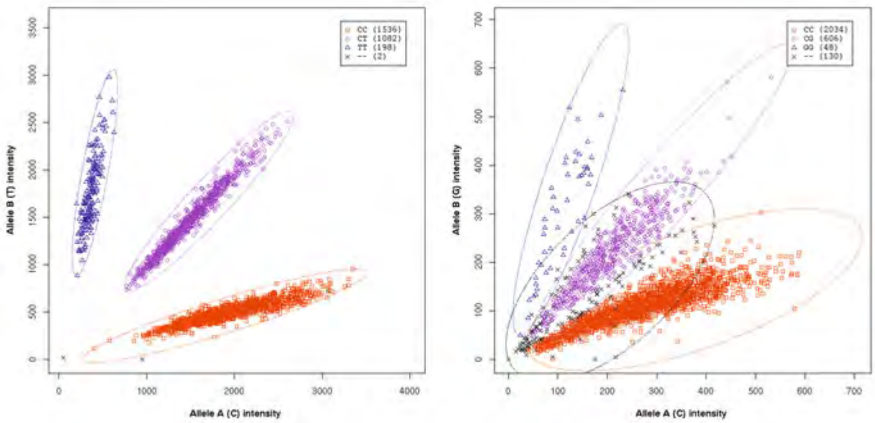


Fig. 11.1 Overview of GWAS findings for chromosomes 1–5. *Source*: Manolio et al. (2008)

In population-based studies, genotyping error that occurs completely at random for each genotype, i.e., independently of the study subject's disease status or genotype, will result in reduced statistical power in the analysis, but generally does not bias the α-level of the test (Fardo et al. 2009a; Gordon and Ott 2001). However, in the case of family-based designs, the same completely at random genotyping errors can lead to an increase of the overall α-level, leading to false-positive results (Mitchell et al. 2003, Gordon et al. 2000, 2001). The situation is worse in the case of non-random genotyping errors, e.g., errors whose probability distribution depend on the true genotype or phenotype. Errors depending on phenotype can occur when the cases and controls are genotyped in different groups, e.g., different laboratories, different genotyping technicians, etc. Most of such error sources can be avoided by a careful planning of the study, using balanced designs with respect to the most important factors that influence genotyping quality.

In the very early days of genotyping, genotype determinations (also known as genotype calls) were made separately for each person by eye, based on electrophoresis gel data. Today, the genotyping process has been highly automated. *Calling Algorithm*s (Rabbee and Speed 2006; Teo et al. 2007) are computerized statistical algorithms used to make genotype calls based on intensity data of the sample for the two alleles of each SNP (see Fig. 11.2). During the genotyping process, each person's DNA segment is greatly amplified, so that the two alleles can be identified by their intensity. A person with genotype AA should have zero intensity for the a allele, and vice versa for the aa genotype.

There are now various genotyping platforms which are able to provide reproducible and high quality data with relatively low genotyping errors or missing



(a) SNP with good genotyping quality

(b) SNP with poor genotyping quality

Fig. 11.2 Intensity plots for 2 SNPs: For the SNP on the left, the clustering genotype calling algorithm works well. The clusters which correspond to the three different genotypes separate well and can be clearly identified. The genotyping quality for the second SNP (**b**) is much poorer. Here the cluster overlap makes it impossible to identify the genotype for a relatively high number of study subjects. *Source*: Courtesy of Dr. Jessica Lasky-Su and Dr. Ross Lazarus

genotypes. For GWAS studies, we require genotyping platforms that can process and call hundreds of thousands of genotypes in thousands of individuals. Because of the large amount of data processing, individual genotype calls cannot be inspected manually for accuracy. Even if the average genotype error per SNPs is small, e.g., 1%, given the large number of SNPs that are genotyped in a GWAS, most GWAS will contain SNPs and study subjects with substantial amount of genotyping error. Including such SNPs and subjects in the analysis can lead to reduced overall statistical power, or systematic bias (highly inflated alpha levels in some cases).

As a result, substantial additional error checking is routinely done after the genotyping process is complete, before proceeding with the data analysis (Laurie et al. 2010; Laird and Lange 2009). This error checking protocol takes the form of applying a series of quality-control filters. Such QC-protocols can often contain filters that are specific to the genotyping platforms used, e.g., quality scores, and are not discussed in this book. Some of quality control filters are based on statistical concepts that we have already discussed in this book and that are universally applicable to any genotype data, such as candidate gene studies as well as a GWAS. Some are only possible with marker sets covering the whole genome. The following statistical QC-filters are commonly implemented in quality-control protocols and ensure that genotyping quality is sufficient for the statistical analysis.

Filter 1 *Filtering for minor allele frequency and for missing rate* For most SNP chips, the genotyping quality depends upon the minor allele frequencies. SNP-chips determine the genotypes based on the intensities for each allele and divide the observed data points into three clusters which correspond to the three possible genotypes. Figure 11.2 illustrates this concept. Establishing the genotype clusters becomes very difficult and error-prone for small minor allele frequencies, since the cluster for the rare homozygous genotype is difficult to identify. In applications, it is standard practice to exclude SNPs with a minor allele frequency of less than 5% from the main analysis. Given ever-improving genotyping technology, this cut-off is likely to become smaller over the next couple of years, but is unlikely to disappear. Similar to this, an excess of missing genotypes either for a specific SNP or for a study subject can indicate a problem with the genotyping process and is indicative of an increased rate of genotyping error. Currently, subjects for which more than 2% of the SNPs are missing are considered as problematic. Such subjects are then also excluded from the analysis.

Filter 2 *Filtering for SNPs whose genotype frequencies are in violation of the Hardy–Weinberg assumption.*

In a GWA for a complex disease/phenotype, the genetic effect sizes of the true DSLs are believed to be small and, unless samples from different ethnicities have been included in the sample, this will also be true for the effects of population-admixture and stratification at a single SNP level. It is therefore plausible to assume that the most likely reason for a SNP's departure from Hardy-Weinberg equilibrium is caused by genotyping error. In applications, it is standard procedure to test all SNPs for departures from Hardy-Weinberg equilibrium and remove those SNPs whose p-values for the Hardy-Weinberg test are smaller than 10^{-5}. Fardo et al. (2009a) give a practical example for the effectiveness of the Hardy-Weinberg filter. A SNP that was genotyped as part of a GWAS for Alzheimer's disease tested highly significant for departure from Hardy-Weinberg equilibrium and was removed from the subsequent analysis. This SNP was then re-genotyped by sequencing. The genotypes for the same SNP obtained from sequencing do not show a departure from Hardy-Weinberg equilibrium anymore, suggesting that the first genotypes were contaminated with strong genotyping error. The data are shown in Table 11.1.

Filter 3 *In family samples, filtering for Mendelian errors/Mendelian Inconsistencies* In family-based samples with observed parental genotypes, Mendelian errors, i.e., transmission patterns that are not possible under Mendel's law, can be used as another measure of the genotyping quality. The Mendelian errors in the sample are identified and then both SNPs and study families whose Mendelian Errors exceed a certain threshold are excluded from the analysis. A typical cutoff here is 5 errors, but often this parameter is sample-dependent and varies across studies. We note that finding Mendelian errors does not remove all genotyping errors; only an offspring genotype inconsistent with the parental genotypes will be detected. Genotyping errors can also be consistent with Mendelian transmissions.

Filter 4 *Filtering for subjects with excessive transmission patterns* For family-based association studies, proband-specific tests can be constructed that compare the transmission rates at a genome wide level within a single proband (Fardo et al. 2009b). Since the null-hypothesis of no genetic association will be true for most SNPs, the transmission probability at the heterozygous genotype should be the same for both alleles. However, in the presence of genotyping error, it can be shown that the common allele tends to be transmitted more often than the rare allele. Thus transmission rates differ from 50/50. Based on this observation, a within-proband TDT/FBAT can be constructed that detects genotyping error and has a χ^2-distribution under the null-hypothesis of no genotyping error. This test can be utilized to identify subjects with particular poor genotyping quality.

Table 11.1 Genotyping error detected by departure from Hardy-Weinberg equilibrium. The genotyping results that were obtained from an Affymetrix 5.0 SNP-chip are denoted here as 'Observed genotypes.' The genotypes obtained from sequencing are referred to as 'True genotype'

	Observed genotypes			
True genotype	aa	Aa	AA	Total
aa	48	6	0	54
Aa	175	198	1	374
AA	0	2	1009	1011
Total	223	206	1010	1439

Filter 5 *Relatedness checks of study subjects*

Using genome wide data one can assess, for any pair of study subjects, how many alleles the two study subjects have in common and compare it with reference data sets on related individuals, e.g., siblings, offspring-parent pairs, etc. This enables identification of closely-related individuals. The general rule-of-thumb is then to remove one of the study subjects in order to ensure the assumption that all study subjects are unrelated (Purcell et al. 2007), as cryptic relatedness can invalidate the usual variance formulas for commonly used test statistics.

After the QC-filtering process is completed, the quality of the genotype data can be assessed by overall criteria. Simple association tests are usually computed, e.g., logistic or linear regression with or without principal component adjustments for population admixture. Since the majority of the SNPs will be under the null-hypothesis of no genetic association, one expects to see a p-value distribution as it would be expected under the null-hypothesis. Q-Q plots can then be used to assess the validity of this assumption (Fig. 11.3). If such plots exhibit any systematic patterns that one would not expect to be observe under the null-hypothesis, this suggest that the data still contain a substantial/detectible amount of genotyping error. In such situation it is recommended to repeat the data filtering process with more stringent filter criteria. A compromise has to be reached then between rescuing as much genetic data as possible for the analysis and achieving acceptable genotyping error rates in the analyzed data. If, as in Fig. 11.3, only a handful of SNPs appear to be outliers, one option is to redo the genotyping of these SNPs on a more accurate platform.

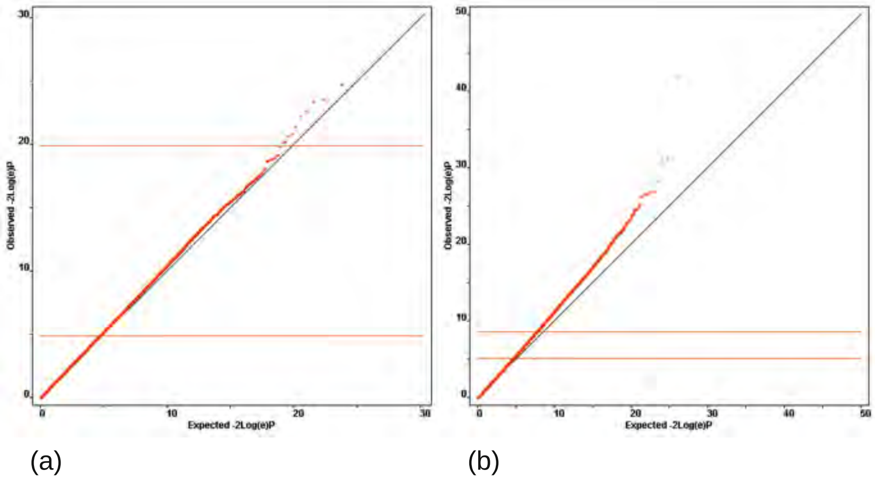


Fig. 11.3 Q-Q plots for two genetic association studies. *Source*: Courtesy of Dr. Jessica Lasky-Su and Dr. Ross Lazarus (**a**) QQ-plot with only a few points deviating from the diagonal, suggesting good genotyping quality. (**b**) QQ-plot with many points deviating from the diagonal, suggesting poor genotyping quality

11.3 Multi-Stage Designs

So far we have assumed that, prior to the analysis of a GWAS, all study subjects have been genotyped at a genome wide level, using one of the available SNP chips, and that each marker is tested for genetic association, using genotype data on all study subjects. A SNP is then declared as genome wide significant if the p-value for the corresponding association test is significant after adjusting for multiple comparisons, based on one of the procedures outlined in Chapter 10, e.g., Bonferroni, FDR correction, or permutation testing. Such designs and analysis plans are often referred to as *one-stage designs* or *single-stage designs*.

Although SNP chips are now commonly used research tools, genotyping based on genome wide SNP chips is still a relatively expensive procedure. This is especially true compared to targeted genotyping of a much smaller numbers of SNPs. In order to minimize the genotyping expenses, multi-stage designs have been proposed. The first step of a multi-stage design with K steps is to divide the study sample into K independent subsets and assign each subset to one of the K stages. Then the first subset of study subjects is genotyped on genome wide SNP chips and the most promising SNPs in terms of p-values for the genetic association tests are selected to be genotyped in the second stage of the design, i.e., SNPs whose p-values are smaller than the cut-off value α_1 for the first stage. Since the cut-off value for the first stage, α_1, is typically in the range of 0.01 to 0.10, the number of SNPs that have to genotyped in the second stage of the study is much smaller. In the kth stage, the SNPs that were identified in the previous stage of the design are genotyped in the kth subset of the study sample and SNPs whose p-values for the association test are smaller than the cut-off value, α_k, are pushed through to the next stage. This process is continued until the final stage of multi-stage design is reached. In the final stage, SNPs whose p-values for the association test are smaller than α_K are declared as genome wide significant. Since only the study subjects in the first stage are genotyped at a genome wide level, the genotyping cost for a *multi-stage design* are usually lower than for a single-stage design with the same number of study subjects (Thomas et al. 2009; Skol et al. 2006). However, many details, such as size of the subsets and choice of $\alpha_k, k = 1, \ldots, K$ need to be considered.

While assuming the total sample size is constant, one-stage designs are obviously more powerful than a multi-stage design because more data are available about each SNP. The multi-stage design has the advantage that the most expensive genotyping, the genome wide SNP-chips, are only applied to a subset of the study. Multi-stage designs therefore tend to be more cost-efficient than one-stage designs. However, given that the SNP-chip prices are falling more rapidly than the prices for the selective ‘genotyping’ that is used in the subsequent stages, the cost advantage of multi-stage designs may soon become irrelevant.

Another aspect that can favor multi-stage designs is that the genotyping and the analysis of each stage is spread over a larger time-window. In situations where the study subjects still need to be recruited and are not available at the beginning of the study, or, where there are funding constraints in terms of the annual budget, multi-stage designs can be excellent choices that are worth consideration.

So far our discussion of multi-stage design has focused on the issues related to the study design. We turn now to the analysis of multi-stage designs. The standard analysis approach for multi-stage designs is to compute the association test statistics within each subset of subjects for all markers genotyped at the kth stage. Similar to one-stage designs, one has to control for the FWER α. The 'typical' FWER α-level for a one-stage GWAS is 10^{-7}, using a Bonferroni correction for an overall α-level of 5% and 500,000 tested SNPs.

One way to define genome wide significance in a multi-stage design to declare significance for all SNPs whose association p-values are smaller than the pre-specified cut-off levels $\alpha_k, k = 1, .., K$ in each stage of the design (Satagopan and Elston (2003), Satagopan et al. (2004a,b), Thomas et al. (2004); Wang et al. 2006). To see how the cut-off values $\alpha_k, k = 1, .., K$ have to be specified in order to maintain the overall type-1 error of α, let's look at the special case of a 2-stage design in which M SNPs are genotyped in the first stage. Under the null-hypothesis of no genetic association for any of the M markers, the cut-off level α_1 can be thought of as the probability that, for any given marker, the association test rejects the null hypothesis in the first stage, given that the null hypothesis is true. The equivalent statement is true for the cutoff-level α_2 of the second stage. Setting the family wise error rate for the GWAS to α and assuming that there is no LD between the markers, we can write

$$\begin{aligned}
1-\alpha &= P(H_0 \text{ is not rejected for any of the } M \text{ SNPs at stage 2})\\
&= \sum_{i=0}^{M} P\begin{pmatrix}\text{for } i \text{ SNPs} : H_0 \text{ is rejected in stage 1,}\\ \text{but not rejected in stage 2}\end{pmatrix}\\
&= \sum_{i=0}^{M} P(\text{for } i \text{ SNPs} : H_0 \text{ is rejected in stage 1}) \times\\
&\quad \times P(\text{for } i \text{ SNPs} : H_0 \text{ is not rejected in stage 2})\\
&= \sum_{i=0}^{M} \left[\binom{M}{i}\alpha_1^i(1-\alpha_1)^{m-i}\right] \times (1-\alpha_2)^i\\
&= \sum_{i=0}^{M} \left[\binom{M}{i}(\alpha_1(1-\alpha_2))^i\,(1-\alpha_1)^{m-i}\right]\\
&= [\alpha_1(1-\alpha_2)+(1-\alpha_1)]^M = (1-\alpha_1\alpha_2)^M
\end{aligned} \tag{11.1}$$

Using a second order Taylor-expansion, the expression $(1-\alpha)^{\frac{1}{M}}$ can be approximated by $1-\frac{\alpha}{M}$. Applying the approximation to equation (11.1), we obtain the following relationship between the overall significance level α and the two cut-off values α_1 and α_2:

$$\frac{\alpha}{M} = \alpha_1 \times \alpha_2. \tag{11.2}$$

For multi-stage designs with K stages, one can show that $\frac{\alpha}{M} = \prod_{k=1}^{K} \alpha_i$, again assuming SNPs are independent. Equation (11.2) requires that the cut-off levels α_1 and α_2, which can also be interpreted as the individual significance levels of each stage, have to multiply to the overall α-level of the GWA that is corrected by Bonferroni-correction for the number of SNPs that have been genotyped in the first stage, i.e., $\frac{\alpha}{M}$. The individual cut-off values for each stage should be specified so that the overall power of the design will be maximized. Thus, for example, when the same number of study subjects are available in each stage of the design, one can show that the most powerful way to define the cut-off values α_1 and α_2 will be $\alpha_1 = \alpha_2 = \sqrt{\alpha/M}$. Additional insight into multi-stage designs can be obtained by solving equation (11.2) for α_2,

$$\alpha_2 = \frac{\alpha}{M \times \alpha_1}. \tag{11.3}$$

Note that the one-stage design uses α/M as the marker cutpoint for genome wide significance, and that under H_0, the expected number of SNPs genotyped in stage two is $M_1 = \alpha_1 \times M$. Hence equation (11.3) implies that $\alpha_2 = \alpha/M_1$, or the standard Bonferoni correction for m_1 tests. Since this can be seen as an independent validation of the SNPs that have been selected in the first stage of study, the last stage of a multi-stage design is sometimes referred to as *'replication'* or *'replication stage'* and it raises the question of the most powerful analysis strategy for multi-stage designs. Thus one strategy would be to analyze in stages according to the multi-stage design and declare SNPs as genome wide significant whose p-value for the final stage is smaller then the cut-off value α_K.

A second possibility is to consider a joint-analysis for the SNPs that are genotyped in all stages of the study. This has the advantage that all data are used in the final stage, but the stages are no longer independent. In the *joint analysis*, the association tests of all stages are combined, e.g., using Fisher's method or the Liptak-method (Lipták 1959; Rice 1990), and adjusted for multiple comparisons for all markers that were genotyped at the genome wide level in the first stage, M. Assuming that the Z-score for mth marker in the k stage of the design is given by Z_{mk} and the corresponding p-value by p_{mk}, Fisher's method for combing p-values can be used, i.e., the combined test statistic for the mth marker is given by

$$Z_m^2 = -2 \sum_{k=1}^{K} \log_e(p_{mk}), \tag{11.4}$$

where Z_m^2 has a χ^2-distribution with $2K$ degrees of freedom. Alternatively, in the Liptak-approach, the combined test statistic for the m th marker is constructed as the weighted sum of the Z-scores for each stage, i.e.,

$$Z_m = \left(\frac{1}{\sqrt{\sum_k n_k}} \sum_k \sqrt{n_k}\, Z_{mk} \right) \quad \sim \quad N(0, 1), \tag{11.5}$$

where n_k denotes the sample size that is available at the kth stage. This method of combining test statistics as weighted Z-scores is often referred to as the Liptak method (Lipták 1959).

In the the joint analysis, the p-values of the combined test statistics have to be smaller than the overall type-1 error that is adjusted for all SNPs that were genotyped in the first stage of the study. In other words, we use the same Bonferroni-correction that we would use for a one-stage design, α/M (Skol et al. 2006).

To address the question whether a staged or ‘replication' analysis is more powerful than a joint analysis, extensive power studies have been conducted. There results suggest that, unless there is substantial heterogeneity between the stages of the study, the joint-analysis is always more powerful (Skol et al. 2006).

11.4 Testing Strategies for Family-Based Studies

In family-based association studies, the information about the genetic association between a marker locus and a phenotype can be decomposed into two statistically independent component. This feature of the data allows the construction of staged testing strategies for GWAS that maximize the power of family-based studies, while maintaining their original robustness against population sub-structure (Van Steen et al. 2005; Zheng et al. 2007; Feng et al. 2007). Such testing strategies in family-based studies are similar to a 2-stage design, but have the distinct difference that they apply the 2-steps of the testing strategy to the 2 independent components of the same data set.

The first component, the so called ‘between'-family component, contains information about the SNP-trait association at a population level, which is assessed based on the proband's phenotype, Y, and the parental genotypes, P (Lange et al. 2003a,b; Van Steen et al. 2005; Laird and Lange 2006). For example, when a quantitative trait is analyzed, the offspring phenotype and parental genotypes can be used to construct estimates for the genetic effect size. The second component, the so-called ‘within'-family component of the data characterizes the SNP-trait association at the family level, i.e., the allele transmissions from the parents to their offspring (Rabinowitz and Laird 2000; Laird et al. 2000b; Spielman and Ewens 1998). The within-family component corresponds to the FBAT/TDT statistic that we discussed previously. Family-based association tests such as the TDT or FBAT are conditional tests that treat the offspring genotype, X, as random, conditioning upon the offspring phenotype, Y, and the parental genotypes P. It is important to note that, when parental information is missing and additional siblings have been genotyped, the approach can easily be adopted by conditioning on the sufficient statistic S instead of the parental genotypes P. The evidence for SNP-trait association is evaluated by comparing the observed offspring genotype with the expected offspring genotype, which is computed by conditioning upon the parental genotypes, assuming Mendelian transmissions. Since the offspring genotype is the only random component of the FBAT/TDT statistic, the implication is that other information in the

FBAT/TDT statistic (i.e., the offspring phenotype and parental genotypes) can be used to assess the evidence for association without biasing the significance level of the FBAT/TDT statistic.

Based on the two information sources about association in family-based designs, the density of the joint distribution for X, Y, and P can then be partitioned into two statistically independent components (Laird and Lange 2006),

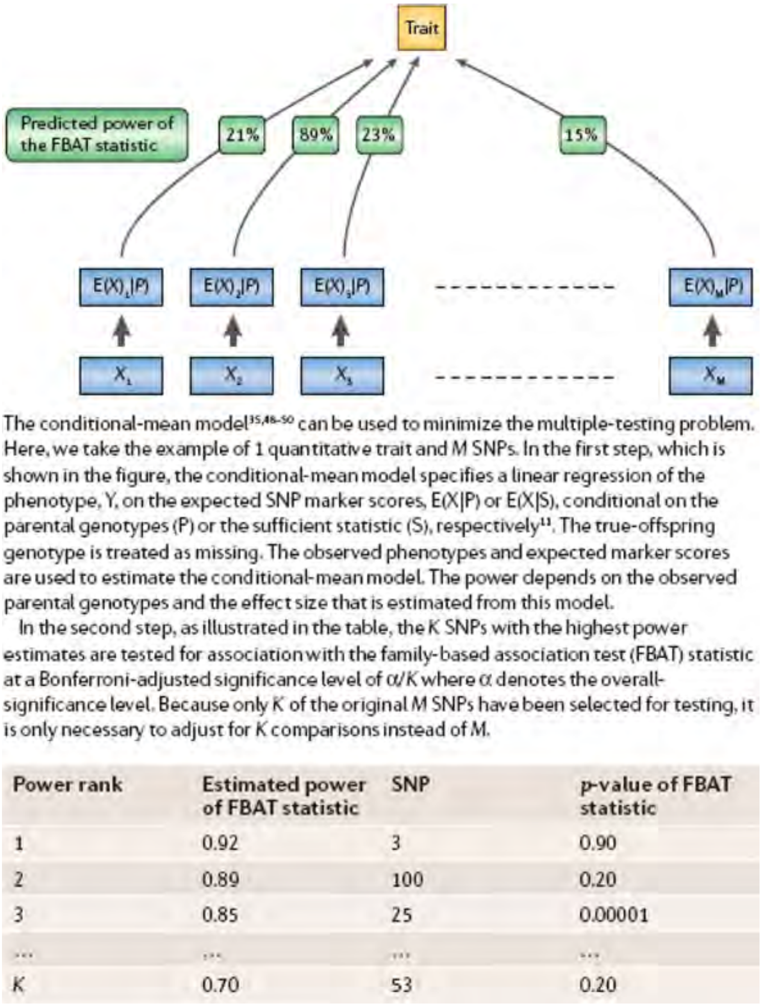
$$p(X, Y, P) = p(X\,|Y, P\,) \times p(Y, P) \tag{11.6}$$

The density $p(Y, P)$ is the basis of the first step of the testing strategy, often referred to as the screening test, and the density $p(X\,|Y, P\,)$ is the basis of the family-based association FBAT/TDT that is applied in the second step, the testing step. The likelihood decomposition (11.6) implies that the two steps, the screening step and the testing step, of the testing strategy are independent. The 'evidence of association' (i.e., the genetic effect size estimate) for each marker from the screening step can be utilized to prioritize SNPs in the second stage without having to adjust the overall significance level for the estimation of the genetic effect size in the first stage. There are various ways in which the information from the screening-step step can inform the application of the FBAT/TDT statistic in the second step. One approach selects a very small number of SNPs based on screening step, typically less than 100, and just test those SNPs for association in the testing step. Since less than 100 SNPs are tested for association and the testing step is independent of the screening step, the adjustment for multiple comparison is much less stringent compared to an analysis that tests all genotype SNPs, e.g., 500,000 and more. Consequently, the power gains over analysis approaches that test all genotyped SNPs can be substantial. This concept is outlined in Fig. 11.4. The algorithm has been used for family-based GWAS for obesity, Alzheimer's disease and Attention Deficit and Hyperactivity Disorder and led to the discovery of novel genetic associations (Herbert et al. 2006; Bertram et al. 2008; Lasky-Su et al. 2008b).

Since the original approach, several extensions and power improvements have been suggested, for measured, time-to-onset and dichotomous phenotypes (Won et al. 2009; Lasky-Su et al. 2010; Ionita-Laza et al. 2007, 2008).

11.5 Replication, Non-replications and Meta-analysis

As in any association study, the most important step after the discovery of a novel association between a SNP and a trait is to validate or replicate the association in an independent studies (Chanock et al. 2007; Pearson and Manolio 2008). The same SNP is genotyped in independent studies in which the same or a related phenotype is available and tested for association. There is general agreement in the field to consider a replication attempt in another study a success if all of the following conditions are met:



The conditional-mean model[35,48–50] can be used to minimize the multiple-testing problem. Here, we take the example of 1 quantitative trait and M SNPs. In the first step, which is shown in the figure, the conditional-mean model specifies a linear regression of the phenotype, Y, on the expected SNP marker scores, E(X|P) or E(X|S), conditional on the parental genotypes (P) or the sufficient statistic (S), respectively[11]. The true-offspring genotype is treated as missing. The observed phenotypes and expected marker scores are used to estimate the conditional-mean model. The power depends on the observed parental genotypes and the effect size that is estimated from this model.

In the second step, as illustrated in the table, the *K* SNPs with the highest power estimates are tested for association with the family-based association test (FBAT) statistic at a Bonferroni-adjusted significance level of α/K where α denotes the overall-significance level. Because only *K* of the original *M* SNPs have been selected for testing, it is only necessary to adjust for *K* comparisons instead of *M*.

Power rank	Estimated power of FBAT statistic	SNP	*p*-value of FBAT statistic
1	0.92	3	0.90
2	0.89	100	0.20
3	0.85	25	0.00001
...	...	...	...
K	0.70	53	0.20

Fig. 11.4 Two-stage testing strategy for family-based association studies

- **The association test for the replication sample is significant at a nominal α-level of 5%:** The association test for exactly the same marker/SNP and a phenotype that is comparable to the phenotype originally used for the discovery of the association has to be significant at a nominal 5% level. Since the study design can vary in terms of the phenotype (quantitative vs dichotomous, population-based vs family-based, etc.) the association test that is used in the replication study does not have to be necessarily the same the test statistic with which the association was originally discovered.
- **The association test for the replication sample has to be based on the same mode of inheritance:** To ensure the consistency and the robustness of the finding,

the association test statistic that is used for the replication sample should be based on the same genetic model as the model that is applied in the original report of the association.

- **The association signal has to have the same direction:** The direction of the genetic effect has to be consistent with the direction of the initial finding. For example, if, in the report of the original association, the minor allele was associated with an increase in disease risk, the replication study has to show the same pattern.
- **Sufficient Sample Size for the Replication Sample:** The sample size in the replication sample should be sufficiently large that it provides adequate statistical power to detect the initial association finding, i.e., at least 80–90% power. Since studies whose sample sizes are too small cannot reliably confirm true genetic association, their inclusion in replication attempts can lead to false negative findings and the dismissal of true findings.

While these criteria for the definition of a successful replication are relatively strict, they are designed to minimize the risk of a false positive findings. It is important to note that non-replications do not necessarily have to mean that the initial association finding is a false-positive. Beside of the possibility of a false positive result, there are numerous reasons for replication failures. The marker that was tested originally to identify the association is unlikely to be the true DSL, but is more likely to be in LD with it. Although the findings of the HapMap project suggest that the local LD-structure is relatively stable across similar populations, such findings are based on healthy subjects. In affected study subjects, the local LD pattern between the identified SNP and the true DSL does not necessarily have to follow the those of healthy individuals. It has been shown that relatively small changes in the LD-structure can lead to the so-called ‘flip-flop’ effect, i.e., the direction of the association is reversed (Lin et al. 2007; Clarke and Cardon 2009). The flip-flop effect will still lead to a significant replication test statistic, but the observed effect size direction in the replication sample will be in the opposite direction of the original finding. Another source for failing to replicate a true positive findings can be genetic effects which vary according to covariates (e.g., age) or phenotype definition. For example, for the phenotype Body Mass Index (BMI) age-dependent genetic effects have been reported for a number of genes, e.g., INSIG2, FTO and ROBO1 (Lasky-Su et al. 2008a). While some associations with BMI can be easily be detected during childhood and up to an age of 20 years, such genetic effects can become less detectible for older age. Table 7.7 also indicates that variation in phenotype definition may lead to non-replication. A reason for replication failures can also be that the replication study varies from the original study in terms of important predictor variables for the analyzed phenotype. Such confounding factors can be body characteristics such as age, but also environmental factors, as well as study design (Heid et al. 2009).

Another reason, of course, for a non-replication can also be that the replication study was underpowered or that replication simply failed just by chance. Unless the power is extremely high, an association test remains a statistical experiment that can result in false negative findings as well. For all these reason, it is obvious that numerous replication studies are needed to either conclusively confirm an identified association or to reject it. It is common practice to replicate GWAS findings in several independent studies of different design type to ensure the robustness of the finding and in studies with different ethnicities to generalize the finding as much as possible.

The results of the replication studies are likely to vary, so they are often combined in a meta-analysis to reach an overall conclusion. Since the genetic effect is likely to be the strongest in the study that identified it, it has become standard practice to exclude the original study from the meta-analysis. In the genetics literature, there are currently two approaches that are favored for the meta-analysis. One way is to combine the p-values of the individual studies, using Fisher's methods of combining p-values. It is important to note that one-sided p-values that take the original effect size direction into account should be used here in order to fulfill the conditions for a successful replication. Since this approach does not take the potentially different sample sizes of the replication studies into account, an alternative way to conduct the meta analysis is to use Z-scores that are weighted by the sample size of each study (equation (11.5)). Based on the overall results of the meta-analysis, the identified association is then either confirmed or rejected. It is also recommended to test for heterogeneity between the replication studies. Approaches that have been developed for the meta-analysis of clinical trials can be utilized for this (DerSimonian and Laird 1986; Emerson et al. 1996; Heid et al. 2009).

While meta-analysis is an important tool to combine replication studies that have been motivated by the findings of a single GWAS, it is also used to combine the results of multiple GWAS for the same phenotype. For many complex diseases/phenotypes, multiple GWAS studies may be available and can be combined to maximize the statistical power to identify novel associations. The major issue that one faces with the meta-analysis of several GWAS studies is that the different studies are not necessarily genotyped on the same SNP-chip. Since the overlap of SNPs between the different GWA chips is marginal, especially for SNP chips from different companies, not all SNPs will have been genotyped in all GWA studies. This problem can be addressed by imputing the un-genotyped SNPs based on the genotyped SNPs in the same area and LD information from the HapMap, assuming that the LD patterns in HapMap are transferrable to the GWA of interest. Several approaches have been proposed and successfully implemented (Lin and Huang 2007; Marchini et al. 2007; Browning and Browning 2009).

11.6 Exercises

1. For Table 11.1, compute the HW tests for both genotyping platforms.
2. For a multi-stage design, show that when the same number of study subjects are available in each stage of the design, the most powerful way to define the cut-off values α_1 and α_2 will be $\alpha_1 = \alpha_2 = \sqrt{\alpha/M}$ (Section 11.3 on page 184).
3. For a multi-stage GWAS design with K stages, show that $\frac{\alpha}{M} = \prod_{k=1}^{K} \alpha_i$.

Appendix A
Basic Concepts of Linkage Analysis (Continued from Chapter 6)

A.1 General Issues with Parametric Linkage Analysis

While the direct counting-method described in the linkage chapter illustrates the key concepts of parametric linkage analysis, i.e., the use of doubly heterozygous parents to test and estimate the recombination fraction, it makes various simplifying assumptions that will not be realistic in most applications. The first one is that we are actually able to observe or infer the haplotypes in the parents and in the offspring. When only marker genotype data and disease traits are available, we will not always be able to reconstruct the haplotypes with certainty, even in multigenerational pedigrees due to missing or non-informative parental or grandparental genotypes. This is usually addressed in the construction of the likelihood function by introducing the unobserved haplotypes as unknown variables, and summing over the possible unobserved phases. With linkage analysis, an underlying assumption is that two loci, either observed markers or DSL, may be linked, but they are far enough apart so that there is no LD between them. The absence of LD implies that the probability of each phase (or each possible pair of haplotypes) is $\frac{1}{2}$. Knowing the probability of phase allows us to sum over all possible combinations of haplotype-genotype configurations; this can be a computationally intensive problem when dealing with complex pedigrees. Since only a small number of these configurations are possible given the observed marker data, sophisticated algorithms have been developed to reduce the computational burden of this summation (Elston and Stewart 1971) and make it feasible in applications.

Another simplification required to extend the direct counting method to disease loci is to assume that the phenotype can be used to infer the genotype at the DSL; this requires a fully penetrant disease without any phenocopies. Even in this case, as was the case with segregation analysis, inferring genotypes from observed phenotypes requires additional assumptions, such as rare disease allele frequencies. When we cannot assume simple Mendelian penetrance functions, then the parameters of the genetic model must be specified along with allele frequencies at the DSL. Hence the term parametric linkage analysis. In the general parametric likelihood-approach one uses a formal genetic model to relate the genotype at the DSL to the trait, and integrates over the possible genotypes at the DSL; it is straightforward to include

N.M. Laird, C. Lange, *The Fundamentals of Modern Statistical Genetics*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-7338-2,

penetrance probabilities in the likelihood model if the parameters are known. Using the likelihood function to estimate these parameters during linkage analysis turns out to be too computationally intensive to be feasible in real applications. Typically one uses estimates for these parameters in the likelihood function that were obtained by a prior segregation analysis. While it is theoretically possible to estimate the nuisance parameters together with the recombination fraction during the likelihood maximization, treating the penetrance probabilities as unknown in the model makes the numerical optimization of the likelihood function much more complex and is often avoided for that reason. In parametric linkage analysis, it is standard practice to use instead the estimates for the penetrance probabilities obtained from a segregation analysis in the same or a different study (Chapter 4).

Another problem which arises in linkage analysis, especially with late onset diseases is missing parental and/or grandparental data. In this case, assumptions need to be made about allele frequencies of the markers in order to assign probabilities to the unobserved haplotypes (using HWE) of the parents. Allele frequencies for the markers are estimated and treated as known in the likelihood.

The example in Table A.1 illustrates this issue. Table A.1 shows two-point LOD scores for data collected on seven markers in the vicinity of the APP gene on chromosome 21, including a polymorphism in the APP gene, APP(*Eco*R1) (the eighth marker, APP(*Bcl*1) was genotyped later). The pedigree structure was shown in Fig. 5.2. Two-point analysis means that each LOD-score is calculated using only the data on phenotype (here assuming a dominant model, consistent with segregation analysis), and one marker. Genotype data were not available for any parent or grandparent, requiring the specification of marker allele frequencies. The results for the first seven markers were used to exclude all but the APP gene as the possible location of the DSL (see Fig. 5.2) because of inferred recombinations between the DSL and other markers. Sequencing the APP gene revealed the presence of a base pair substitution (C to A) at exon 17 in APP(*Bcl1*), that was present in affected individuals. In this case, previous studies of other early onset AD families had found recombinants between the DSL and the APP mutation, suggesting genetic heterogeneity. This motivated a study of linkage in only one large family. Indeed, this rare mutation explains very few cases of AD, early onset or otherwise.

Table A.1 Two-point linkage analyses (lod scores) between Alzheimer's disease and polymorphic DNA markers on the long arm of chromosome 21 Goate et al. (1991)

	Recombination fraction(θ)				
Locus	0.00	0.01	0.05	0.10	0.20
D21S16	1.36	1.35	1.28	1.18	0.92
D21S13	2.21	2.17	2.06	1.89	1.44
D21S1	2.65	2.61	2.45	2.22	1.67
APP(*Eco*R1)	2.93	2.88	2.82	2.36	1.72
D21S17	0.45	0.51	0.61	0.63	0.54
D21S156	−6.31	−4.29	−2.73	−1.76	−0.77
D21S167	−8.02	−3.44	−1.88	−1.17	−0.49
APP(*Bcl*1)	3.37	3.31	3.07	2.76	2.09

As can be seen from the examples, assuming a simple Mendelian model and having data on three or more generations is very helpful simplifying the likelihood and in resolving phase, but the computational complexities can be vastly increased if haplotypes cannot be inferred or if genotype data are missing for the previous generations. In addition, there can be considerable sensitivity of the results to misspecification of the genetic model.

A.2 Non-parametric Linkage Analysis

In contrast to the parametric approach, non-parametric linkage analysis does not depend on any statistical model specification for a functional relationship between the phenotype and and the disease locus. Thus the recombination fraction θ, the mode of inheritance, or the penetrance probabilities do not have to be specified. While the parametric approach is optimal in terms of statistical power, the correct specification of the statistical model for the likelihood function is required for its validity (and optimal power), and differing assumptions can have a strong influence on the results of a parametric linkage analysis. The idea of non-parametric linkage analysis is to avoid the need for any model building specifications and to obtain analysis results that are less susceptible to model assumptions.

Non-parametric linkage analysis is based on a simple and intuitive principle: If a marker locus is in linkage with a disease locus, affected relatives should share more genetic information in the area of the DSL (i.e., marker alleles) than expected just by chance. This is because we assume them to be sharing at least one allele at the DSL. Under the null-hypothesis of no linkage between the marker and underlying DSL, the amount of genetic information that the two affected relatives share at the marker locus is determined by Mendel's law because their affection status is irrelevant. The key concept of non-parametric linkage analysis is to construct a test, either a score or a likelihood-ratio, that compares the observed genetic 'similarity' at the marker among two affected relatives with the genetic similarity that is expected under Mendel's Law of random transmissions.

To construct such linkage test, we have to formalize the idea of genetic similarity by defining the amount of genetic material that offspring or related individuals within one pedigree share at a specific marker. There are two ways commonly used to define genetic similarity:

1. Two alleles at the same genetic marker locus are called identical by state (IBS) if their DNA sequence is physically identical, i.e., both alleles are A or a, for example.
2. Two alleles at the same genetic marker locus are called identical by descent (IBD) if they are copies of the identical allele carried by a recent common ancestor. If the case of siblings, this means that the allele shared IBD is from the same parental chromosome, assuming no inbreeding.

Both definitions are illustrated in Fig. A.1 for a pair of siblings. Since we can observe only two alleles in a person at any genetic locus, two individuals can share

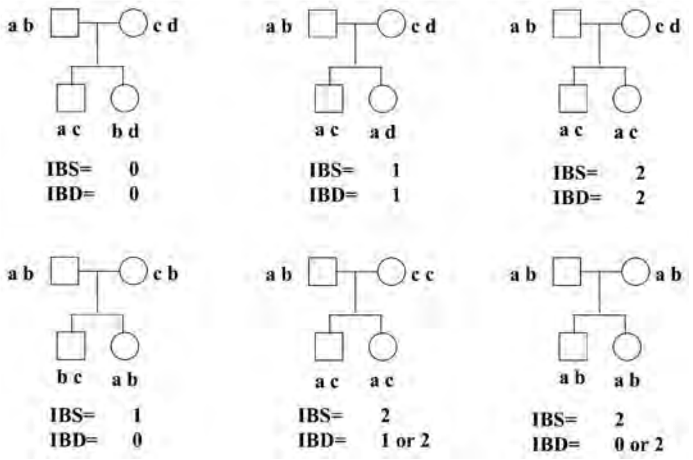


Fig. A.1 Illustration of allele sharing identical by state (IBS) and identical by descent (IBD)

only 0, 1 or 2 alleles by definition. It is clear that identical by descent always implies identical by state, since it is the much stronger genetic concept. The simple examples in Fig. A.1 also show it is always possible to observe IBS, but it will not always be possible to determine IBD status uniquely when the original chromosomes of the alleles cannot be identified with absolute certainty. Further, as we will discuss, excess sharing of marker alleles IBD among diseased relatives implies linkage with the disease locus (see Figs. A.3 and A.4), while this is not necessarily the case for allele sharing IBS. Note also that the frequency of IBS can depend strongly on allele frequency, but not IBD. In addition, as we will show, the joint distribution of IBD at two markers can be easily determined as a function of the recombination parameter between them, but this does not hold for IBS.

In order to derive a non-parametric test for linkage, the first step is to derive the distribution of IBD under the null hypothesis of no linkage. Under the null hypothesis of no linkage of a marker to the DSL, Mendel's laws apply to transmissions of marker alleles to offspring regardless of affection status. In Section 4.2, we showed that the probability of sharing 2 alleles for siblings is $\frac{1}{4}$. It can likewise be shown that the probability of sharing 0 is also $\frac{1}{4}$, and for sharing 1 it is $\frac{1}{2}$ (See exercise 4 of Section 6.4). In a similar way, we can calculate the sharing probabilities for more distant relatives (Table A.2).

The probabilities in Table A.2 hold when there is no selection of individuals with regard to disease traits. They also hold if we select diseased relatives, but the DSL is not linked with the marker ($\theta = \frac{1}{2}$). What happens under the alternative hypothesis ($\theta < \frac{1}{2}$)? Under the alternative, the IBD allele sharing at the marker locus should depart from what is expected under the null, in the direction of increased sharing. In some circumstances (for example parents are ab,cd or ab,bc), it is possible to count the number of offspring pairs sharing 0,1, or 2 alleles IBD at the marker. In order to derive the basic non-parametric tests, we assume for now that we observe

Table A.2 Distribution of IBD for different pedigree relationships

	Probability of Sharing Two Alleles IBD		
Type of relative pair	p_0	p_1	p_2
Monozygotic Twins	0	0	1
Full Sibs	1/4	1/2	1/4
Parent-Offspring	0	1	0
First Cousins	3/4	1/4	0
Double First Cousins	13/16	1/8	1/16
Grandparent-Grandchild	1/2	1/2	0

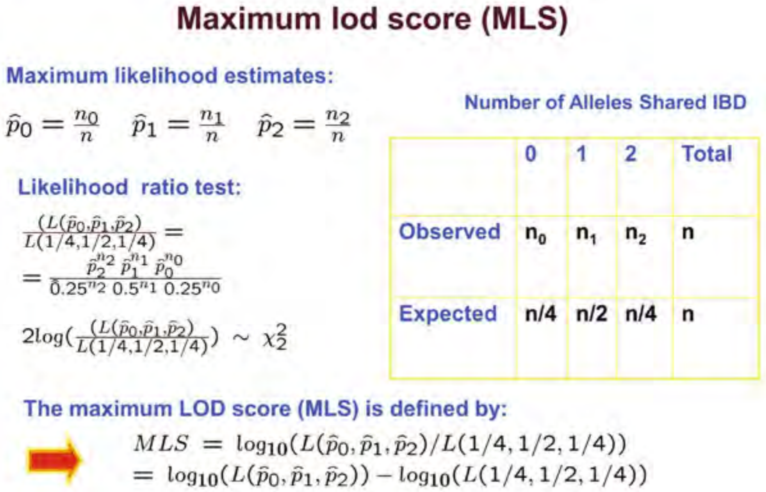
n affected sib pairs with perfect marker information, i.e., IBD can be inferred. The number of sib-pairs who share 0 alleles is denoted by n_0, the number of sib-pairs who share 1 allele by n_1 and the number of sib-pairs who share 2 alleles by n_2, with the probability of sharing k alleles IBD given by p_k, $k = 1, 2, 3$. Then the likelihood function of the data is multinomial, and given by:

$$L(p_0, p_1, p_2) = {p_0}^{n_0} {p_1}^{n_1} {p_2}^{n_2}.$$

The maximum likelihood estimates for the sharing probabilities are given by

$$\hat{p}_0 = \frac{n_0}{n}, \ \hat{p}_1 = \frac{n_1}{n}, \ \hat{p}_2 = \frac{n_2}{n}.$$

The calculation of the likelihood ratio test, and the maximized LOD score are given in Fig. A.2. In analogy to the parametric approach the logarithm taken with base 10 gives the maximized lod-score (MLS).



Maximum lod score (MLS)

Maximum likelihood estimates:

$$\hat{p}_0 = \frac{n_0}{n} \quad \hat{p}_1 = \frac{n_1}{n} \quad \hat{p}_2 = \frac{n_2}{n}$$

Likelihood ratio test:

$$\frac{(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)}{L(1/4,1/2,1/4)} = \\ = \frac{\hat{p}_2^{n_2}\ \hat{p}_1^{n_1}\ \hat{p}_0^{n_0}}{0.25^{n_2}\ 0.5^{n_1}\ 0.25^{n_0}}$$

$$2\log(\frac{(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)}{L(1/4,1/2,1/4)}) \sim \chi^2_2$$

Number of Alleles Shared IBD

	0	1	2	Total
Observed	n_0	n_1	n_2	n
Expected	n/4	n/2	n/4	n

The maximum LOD score (MLS) is defined by:

$$\begin{aligned} MLS &= \log_{10}(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)/L(1/4,1/2,1/4)) \\ &= \log_{10}(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)) - \log_{10}(L(1/4,1/2,1/4)) \end{aligned}$$

Fig. A.2 MLS computation for n affected sib-pairs

However, similar to the parametric approach, use of the likelihood ratio test also has its caveats in this setting. The problem is caused by the maximum likelihood estimates for the sharing probabilities. It can be shown that the parameter estimates have to be restricted to a range (the so-called Holmans triangle, Holmans 1993) that corresponds to plausible disease models under the alternative hypothesis. For example, given the construction of the sharing approach, one always has to expect an excess of sharing rather than a lack of sharing for a linked marker under the alternative hypothesis. When the maximization of the likelihood is restricted with respect of the sharing probabilities to the range that is defined by the plausible disease models, the asymptotic distribution of the likelihood ratio test will be a mixture of ${\chi^2}_1$ and ${\chi^2}_2$ and estimation of the parameters is only possible using numerical optimization routines.

These issues can be circumvented by the construction of a one-sided, non-parametric score test (NPL). The idea is to compare the observed number of shared alleles with the expected number of shared alleles. Defining the number of alleles that are shared IBD by the ith affected sib-pair as a_i, for $i = 1, \ldots n$, it follows that

$$E(a_i|H_0) = 0\frac{1}{4} + \frac{1}{2} + 2\frac{1}{4} = 1$$

$$\mathrm{Var}(a_i|H_0) = \frac{1}{2} \quad \text{(exercise 5 of Section 6.4).}$$

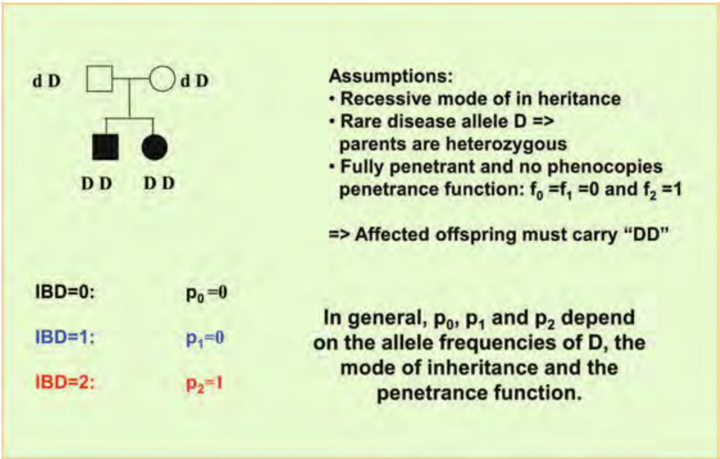
Note that the concept of IBD does not depend upon the particular genotype of the parent, so its distribution is the same for every pair of siblings.

The score test for the entire sample is then defined by

$$Z = \left(\sum_{i=1,\ldots n} a_i - n \right) / \sqrt{n/2} = (n_2 - n_0)/\sqrt{n/2},$$

where a_i are the number of pairs sharing i alleles IBD. The advantage of the score test is that, by adding up the contributions of each family and obtaining $n_2 - n_0$ in the numerator of the score test, the same restrictions required for the Holmans triangle can therefore be easily enforced in the NPL score by using a one-sided test statistic, requiring $n_2 > n_0$.

The only simplistic assumption that we used for the derivation of the non-parametric approaches was that the IBD status could always be inferred. In practice, of course, this will not be the case. To avoid losing these families in the analysis, one has options. Instead of IBD, we can compare the observed number of alleles shared IBS with its expected number which will be computed here based on the Mendelian transmission and the allele frequencies. An advantage of this case is that no knowledge of parental genotypes is necessary. However, allele frequencies have to be known and can vary between populations; this is a clear disadvantage for the use of the IBS concept. Alternatively, one can compute the MLS or the NPL score based on the ‘observed’ likelihood in which the IBD is missing for some individuals, and we sum over likelihood functions for all genotype data that is consistent with



Assumptions:

- Recessive mode of in heritance
- Rare disease allele D => parents are heterozygous
- Fully penetrant and no phenocopies penetrance function: $f_0 = f_1 = 0$ and $f_2 = 1$

=> Affected offspring must carry “DD”

IBD=0: $p_0 = 0$

IBD=1: $p_1 = 0$

IBD=2: $p_2 = 1$

In general, p_0, p_1 and p_2 depend on the allele frequencies of D, the mode of inheritance and the penetrance function.

Fig. A.3 At the disease locus: Derivation of the IBD probabilities p_K for two affected siblings under the alternative hypothesis

the observed phenotypic and genetic data. If parental genotypes are observed, the likelihood calculations are simple and allele frequencies are not required. However, without parental genotypes, allele frequencies must be supplied as in the case of using IBD (Risch 1990b).

For a fully penetrant recessive disease, we can also estimate the sharing probabilities directly at the DSL and at a marker locus that has a recombination fraction θ with the DSL. The calculations for the sharing probabilities at the DSL are shown in Fig. A.3 for a hypothetical sibpair; to calculate the IBD probabilities at a nearby marker, we need to specify θ between the DSL and the marker.

Figure A.4 shows the decay in expected sharing probabilities at nearby loci as a function of θ. At the DSL, affected offspring will share 2 alleles identical by descent with 100% probability and then, with an increasing recombination fraction θ, the probability of sharing 2 alleles IBD drops slowly until it has reached 25%, its value under the alternative hypothesis. While this departure from the sharing probabilities during the null-hypothesis can certainly not be expected to be that strong in scenarios with more realistic assumption about the penetrance probabilities, it still will be present and the example highlights the basic idea of non-parametric linkage analysis. The calculations underlying Fig. A.4 are based on the Markov nature of recombinations, as we now discuss.

A.3 Multipoint Linkage Analysis

Figure A.4 shows the predicted IBD sharing probabilities as we move away from the probabilities observed at a marker (in this case, the DSL). The basis for these calculations is the fact IBD sharing will remain constant for a pair of offspring, from one

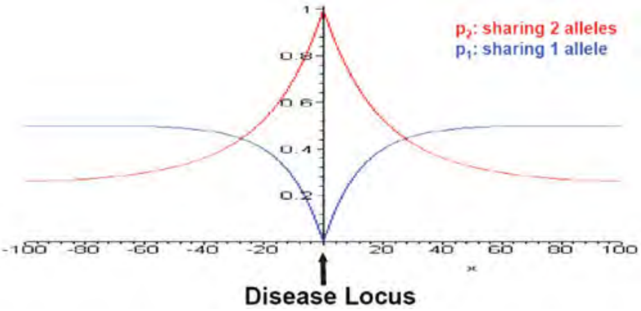


Fig. A.4 At a marker locus: IBD probabilities p_k for two affected siblings under the alternative hypothesis

locus to the next, unless a recombination occurs in the formation of gametes between the two loci. Consider the parent in Fig. 6.1, and IBD sharing for this single parent at the A locus. Suppose the two offspring inherit A and a from this parent, then their IBD status is 0 at the A locus. If there is no recombination in the transmission for either offspring, with probability $(1-\theta)^2$, or if there is a recombination for both offspring (θ^2), the IBD status will remain zero at locus B. Conversely, if there is a recombination in one offspring but not the other ($2\theta(1-\theta)$), IBD status will be 1 at the B locus. It follows that the probability sharing IBD from this parent moves from 0 at locus A to 1 at locus B with probability ($2\theta(1-\theta)$). Since recombinations for different parents are independent, by looking at all possible IBD values at locus A and considering all possible transmissions from both parents to both offspring, we can (exercise 8 of Section 6.4) derive the transition probabilities given in Table A.3, where

$$\psi = \theta^2 + (1-\theta)^2.$$

When there are three markers, we need only compute P(IBD at marker 3| IBD at marker 2) since recombinations are independent from one locus to the next unless distances between markers are very small. Thus IBD status only depends upon the most recently observed marker.

One use of the transition probabilities is to compute the marker sharing probabilities under the alternative, given a set of assumed sharing probabilities at the DSL. This will enable sample size and/or power calculations for different assumptions on θ and the disease model.

Table A.3 IBD transition probabilities: P(IBD at locus 2|IBD at locus 1)

	IBD at Locus 1		
IBD at Locus 2	0	1	2
0	ψ^2	$\psi(1-\psi)$	$(1-\psi)^2$
1	$2\psi(1-\psi)$	$1-2\psi(1-\psi)$	$2\psi(1-\psi)$
2	$(1-\psi)^2$	$\psi(1-\psi)$	ψ^2

The use of transition probabilities also leads us directly to the concept of multipoint analysis. With multipoint analysis, the data on all markers are used to calculate sharing for each pair at every marker, and at every point in between. A genetic map which gives θ between every successive pair of markers is necessary to implement a multipoint analysis. There are two main advantages of multipoint analysis. First, if we have incomplete information at a marker (say the data for a family was not observed at a marker, or the parents were both homozygotes so that nothing could be inferred about sharing), the transition probabilities allow us to predict their sharing based on the data from neighboring markers, and the known recombination fraction between the two loci. Secondly, we can predict sharing, and thus calculate a maximized LOD score for each point on the chromosome, not just at observed loci. Thus with a set of markers spanning the entire linkage map, we can test the null hypothesis H_0: no linkage of the DSL to any loci in the genome. The test is based on the maximized LOD score (maximizing over the entire genome), using simulation or theory based on Gaussian Markov processes to calculate p-values.

Computation of the maximized LOD score, whether parametric or non-parametric, is in practice complicated by the failure to observe IBD status (or phase in the parametric setting) for each individual, and by the need to use data on flanking markers on both sides of a given location. The Lander-Green algorithm is widely used for computing multi-point likelihoods with a large number of markers but pedigrees of small size, e.g., nuclear families (Lander and Green 1987).

Appendix B
A Class of Score Tests for Family Designs

The simple TDT test is a score test, based on the likelihood of the offspring genotypes, conditioned on the offspring trait and the parental genotypes (Schaid and Sommer 1996). Here we develop this approach in a general setting. For the ith family, let Y_i denote the vector of offspring traits, X_i denote the vector of offspring genotypes and $f(Y_i, |X_i, P_i, \gamma)$ denote the probability density of the traits, conditioned on the offspring genotype, the parental genotype, and unknown parameters, γ. In genetic terminology, $f(\bullet)$ is the penetrance function and specifies the genetic disease model. Because the distribution of X_i is completely determined by P_i, $f(Y_i, |X_i, P_i, \gamma)$ is generally assumed not to depend directly on the parental genotypes when offspring genotypes are in the model and henceforth we omit them. In addition, we ordinarily assume each offspring's phenotype depends only on their own genotype, and not the genotype of another sibling. In general, $f(Y_i, |X_i, \gamma)$ does not factor into separate contributions for each offspring when there is correlation due to shared environmental or unobserved genetic factors. However, assuming phenotypic independence leads to a particularly simple form for the test statistic. While we may improve power by modeling the phenotypic correlation, as we show below, assuming independence does not invalidate the a-level of the test. Henceforth, we assume that

$$f(Y_i, |X_i, \gamma) = \prod f(Y_{ij}|X_{ij}, \gamma) \tag{B.1}$$

where the product is over all j in the ith family.

The vector γ will ordinarily contain both association parameters, say β, and nuisance parameters, say α, which will describe other aspects of the trait distribution. For example, if the trait is normal, we typically assume that the genotype affects the mean, but not the variance. In particular, we parameterize so that $f(Y_{ij}, |X_{ij}, \gamma) = f(Y_{ij}|X_{ij}, \beta, \alpha)$, and under the null, $\beta = 0$, so that $f(Y_{ij}|X_{ij}, \beta = 0, \alpha) = f(Y|\alpha)$, i.e., the distribution of the trait does not depend on the marker genotypes of the offspring under the null. Further, let $f(X_{ij}|P_i)$ be the probability density of the offspring genotype conditioned on parental genotype. Note that the latter is completely known and determined by Mendel's laws (in the case of a single offspring or where $\theta = \frac{1}{2}$ is included in H_0), whereas the penetrance function reflects our alternative hypothesis, and is hypothesized rather than known.

The conditional likelihood for the offspring genotype (X_{ij}) given parental genotypes (P_i) and the offspring trait (Y_{ij}) is given by:

$$f(X_{ij}|Y_{ij},\gamma) = \frac{f(Y_{ij}|X_{ij},\gamma)\,f(X_{ij}|P_i)}{\sum_{j=1,\ldots,n_i} f(Y_{ij}|X_{ij},\gamma)\,f(X_{ij}|P_i)}, \tag{B.2}$$

where summation is over all values of all X_{ij} compatible with P_i. To construct the total data log-likelihood we take the log of each $f(X_{ij}|Y_{ij},\gamma)$ and over all i and j to obtain

$$L = \sum_{j=1,\ldots n_i, i=1,\ldots,n} \ln f(X_{ij}|Y_{ij},\gamma).$$

To obtain the score statistic we simply take the derivative of L with respect to β, and evaluate at $\beta = 0$. Now we assume that $f(Y_{ij}|X_{ij})$ takes an exponential family form with a GLM for the mean of Y_{ij} as given in equation (7.7). Thus the score is simply expressed as

$$U = \sum_{j=1,\ldots n_i, i=1,\ldots,n} [(Y_{ij} - E(Y_{ij})][X_{ij} - E(X_{ij}|P_i)], \tag{B.3}$$

where $E(Y_{ij})$ is calculated under $H_0 : \beta = 0$. Assuming $\theta = \frac{1}{2}$ under H_0, the variance of U is computed quite simply as

$$\mathrm{var}(U) = \sum_{\substack{i=1,\ldots,n,\\ j=1,\ldots,n_i}} {T_{ij}}^2 \mathrm{var}(X_{ij}|P_i)$$

where both $E(X_{ij}|P_i)$ and $\mathrm{var}(X_{ij}|P_i)$ are computed using Mendel's Law.

Properties of the Score Test

Note that under H_0, $f(Y_{ij}|X_{ij},\gamma)$ does not depend on X_{ij} under the null, and hence cancels out of the likelihood in equation (B.2). The same is true whether or not we assume phenotypic independence as in equation (B.1). All we require is that $f(Y_i,|X_i,P_i,\gamma)$ evaluated at $\beta = 0$ does not depend upon X_i . Because we condition on the traits, the distribution of the score test depends only on the distribution of X_{ij} under the null and thus under H_0 the test retains robustness to assumptions concerning $f(Y_i,|X_i,P_i,\gamma)$. The important feature of conditioning on P_i is that any nuisance parameters which govern the distribution of the parental genotypes, such as allele frequencies and random mating assumptions, and Hardy-Weinberg Equilibrium are not needed. The distribution depends upon the joint distribution of the X_{ij} given P_i. $E(X_{ij}|P_i)$ is given by Mendel's first law, but the variance of U needs to be adjusted if linkage is present and we have multiple offspring per family, as discussed in Section 9.2. A complication arises in the case of nuisance parameters α. Under H_0, $f(Y_{ij}|\alpha)$ drops out of the conditional likelihood, so that α cannot be estimated by ML. Notice that for the general FBAT statistic, we only need to specify $E(Y_{ij})$ under H_0. If the data are sampled randomly from the population, this can be estimated by the sample mean of the trait. But for dichotomous traits, this is more of a problem. See Section 9.3.

Missing Parents

In the case of missing parents, we replace $f(X_i|P_i)$ by $f(X_i|S_i)$, where X_i is the vector of offspring genotypes in family i and S_i is the sufficient statistic for parental genotypes; $S_i = P_i$ if no parental genotypes are missing. The sufficient statistic, S_i, has the property that

$$f(X_i, P_i, \Omega) = f(X_i|S_i)\, f(S_i|P_i, \Omega),$$

where Ω contains all of the parameters and assumptions governing the distribution of P_i, e.g., allele frequencies, random mating, HWE, etc. Note that $f(X_i|P_i)$ does not depend upon Ω, only on Mendel's laws. By replacing $f(X_i|P_i)$ in the likelihood with $f(X_i|S_i)$ we see that the score statistic remains the same except we replace $E(X_{ij}|P_i)$ with $E(X_{ij}|S_i)$ and $\mathrm{var}(U)$ with

$$\mathrm{var}(U) = \sum_{i=1,\ldots,n}\ \sum_{j=1,\ldots,n_i,\, j'<j} T_{ij}{}^2\mathrm{var}(X_{ij}|S_i) + 2T_{ij}T_{ij'}\mathrm{cov}(X_{ij}, X_{ij'}|S_i).$$

In general, additional siblings are required to obtain non-degenerate distributions for $f(X_i|S_i)$. The derivation of these distributions for autosomal markers are given in Rabinowitz and Laird (2000), and implemented in the FBAT and PBAT software packages. The X-chromosome is a straightforward extension and has been integrated into both packages.

Appendix C
The TDT Tests for Both Linkage and Association (LD)

In this appendix we show why the TDT is a test for both linkage and association. The proof does not require HWE, or any formal genetic model; the proof is similar to that of originally given by Ott (1989). We assume only the existence of a disease allele, D, with the property that $f(\mathrm{D}) \neq f(\mathrm{d})$, where $f(\mathrm{D})$ is the probability of disease in offspring with a D allele and $f(\mathrm{d})$ is the probability of disease given any other allele at the DSL. The function $f(\mathrm{D})$ depends on many factors, including the mode of inheritance, and assumptions about the distribution of disease genotypes among the parents. It is not necessary to assume anything about these quantities here, just that D is an allele (or collection of alleles) at a locus which directly influences risk of disease differently from allele d and that the marker lies on autosomal chromosomes. Further we assume that the marker has no direct effect on disease, conditional on disease allele status. The argument can be extended to handle the X-chromosome.

Let θ denote the recombination fraction between the marker and the DSL, with $\theta = \frac{1}{2}$ meaning no linkage, and $\theta = 0$ meaning the two loci are the same. Let δ denote the LD between the two loci, with $\delta = 0$ implying no LD and let τ denote the transmission probability, i.e.,

$$\tau = \mathrm{P(Aa\ parent\ transmits\ A|Y{=}1}, \theta,\ \delta),$$

where $Y = 1$ denotes an affected offspring. Thus $(1 - \tau)$ is the corresponding probability that the a is transmitted. We will show that $\tau = \frac{1}{2}$, and thus H_0 for the TDT is satisfied, if either $\theta = \frac{1}{2}$, or $\delta = 0$, or both. As a result, $\tau \neq \frac{1}{2}$ unless both $\theta < \frac{1}{2}$ and $\delta \neq 0$.

Consider the diagram in Table C.1 which shows the three possible unphased genotypes involving the marker and the DSL for individuals who are Aa at the marker. The bottom row of the table shows the possible phases for the double heterozygote.

It is clear that the parents who are homozygous at the disease allele are not informative under the alternative. The same disease allele is transmitted with both A and a, thus disease status of the offspring is unaffected by transmission at the marker, hence

Table C.1 Passive unphased genotypes and corresponding possible haplotypes for an Aa parent

Possible parental unphased genotypes		
Aa DD	Aa Dd	Aa dd
Possible Pairs of Parental Haplotypes		
$A \mid a$ $D \mid D$	$A \mid a$ $A \mid a$ $D \mid d$ $d \mid D$	$A \mid a$ $d \mid d$

$$\begin{aligned}\tau &= P(\text{Aa parent transmits A} \mid Y = 1, \theta, \delta, \text{DD or dd at DSL})\\ &= P(\text{Aa parent transmits a} \mid Y = 1, \theta, \delta, \text{DD or dd at DSL})\\ &= \frac{1}{2}.\end{aligned}$$

This is the same as for a conventional linkage study where double heterozygous parents are required to show linkage. Our objective here is to determine conditions on θ and δ such that

$$\tau = \text{P(parent transmits A} \mid Y = 1, \text{DH}, \theta, \delta) \neq \frac{1}{2}$$

or equivalently,

$$\begin{aligned}&P(\text{parent transmits A}, Y = 1 \mid \text{DH}, \theta, \delta)\\ &\neq P(\text{parent transmits a}, Y = 1 \mid \text{DH}, \theta, \delta),\end{aligned}$$

where DH means the parent is a heterozygote at both loci.

Let π be the p(phase=AD/ad). Conditioning on being a double heterozygote parent, it is straightforward to see that $\delta = 0$ implies $\pi = \frac{1}{2}$. disease allele will depend upon phase, and in general, the probability of phase depends on haplotype frequencies at the two loci, δ, plus assumptions about random mating, etc. For a pair of haplotypes we consider both possible phases and find:

$$P(\text{parent transmits A }, Y = 1 \mid \text{ DH}) = f(\text{d}) + \theta Z + \pi Z(1 - 2\theta)$$

and

$$P(\text{parent transmits a }, Y = 1 \mid \text{ DH}) = f(\text{D}) - \theta Z - \pi Z(1 - 2\theta),$$

where $Z = f(\text{D}) - f(\text{d})$. Suppose $\theta = \frac{1}{2}$, then both of the above transmission probabilities are $[f(\text{D}) + f(\text{d})]/2$. Likewise, if $\delta = 0$, then $\pi = 1/2$, and both transmission-probabilities again equal $[f(\text{D}) + f(\text{d})]/2$. Thus $\tau = \frac{1}{2}$ if $\theta = \frac{1}{2}$ or δ = 0 or both, and $\tau \neq \frac{1}{2}$ only if both linkage and LD are present. In fact, we can show that

$$\text{P (parent transmits A} \mid Y = 1, \text{DH parent}) = \frac{1}{2} + \left(\pi - \frac{1}{2}\right)(1 - 2\theta)\Delta$$

where $\Delta = [f(\mathrm{D}) - f(\mathrm{d})]/[f(\mathrm{D}) + f(\mathrm{d})]$. In practice, for non-negligible δ and $\theta \approx 0$, we have

$$\mathrm{P}\,(\text{parent transmits A} \mid Y = 1, \text{DH parent}) = \frac{1}{2} + \left(\pi - \frac{1}{2}\right)\Delta\,.$$

Bibliography

Abecasis G, Cardon L, Cookson W (2000) A general test of association for quantitative traits in nuclear families. American Journal of Human Genetics 66:279–292

Ahrengot V, Eldon K (1952) Distribution of abo-mn and rh types among Eskimos in South-west Greenland. Nature 169:1065

Allison D (1997) Transmission-disequilibrium tests for quantitive traits. American Journal of Human Genetics 60:676–690

Altmueller J, Palmer L, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. The American Journal of Human Genetics 69(5):936–950

Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11(3): 375–386

Bateson W (1909) Mendel's Principles of Heredity. Cambridge University Press, London

Baz K, Emin Erdal M, Yazici A, Söylemez F, Güvenç U, Tasdelen B, Ikizoglu G (2008) Association between tumor necrosis factor-alpha gene promoter polymorphism at position -308 and acne in Turkish patients. Archives for Dermatolological Research 300(7):371–376

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57:289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29:1165–1188

Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan M, Schjeide B, Hooli B, DiVito J, Ionita I, et al (2008) Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. The American Journal of Human Genetics 83(5):623–632

Browning B, Browning S (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics 84(2):210–223

Burnham K, Anderson D (2004) Understanding AIC and BIC in model selection. Sociological Methods & Research 33(2):161–304

Campbell C, Ogburn E, Lunetta K, Lyon H, Freedman M, Groop L, Altshuler D, Ardlie K, Hirschhorn J (2005) Demonstrating stratification in a European American population. Nature Genetics 37(8):868–872

Chanock S, Manolio T, Boehnke M, Boerwinkle E, Hunter D, Thomas G, Hirschhorn J, Abecasis G, Altshuler D, Bailey-Wilson J, et al (2007) Replicating genotype? Phenotype associations. Nature 447(7145):655–660

Chen HS, Zhu X, Zhao H, Zhang S (2003) Qualitative semi-parametric test to detect genetic association in case–control design under structured population. Annals of Human Genetics 67:250–264

Chen Z, Zheng G, Ghosh K, Li Z (2005) Linkage disequilibrium mapping of quantitative-trait loci by selecive genotyping. American Journal of Human Genetics 77:661–669

Christensen K, Arnbjerg J, Andresen E (1985) Polymorphism of serum albumin in dog breeds and its relation to weight and leg length. Hereditas 102(2):219–223

Clarke G, Cardon L (2009) Aspects of observing and claiming allele flips in association studies. Genetic Epidemiology 34(3):266–274

Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. The American Journal of Human Genetics 65(4):1170–1177

Clayton D, Chapman J, Cooper J (2004) Use of unphased multilocus genotype data in indirect association studies. Genetic Epidemiology 27(4):415–428

Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42(2):393–399

Coon K, Myers A, Craig D, Webster J, Pearson J, Lince D, Zismann V, Beach T, Leung D, Bryden L, et al (2007) A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. The Journal of Clinical Psychiatry 68(4):613

Cordell H (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics 11(20):2463–2468

Daly M, Rioux J, Schaffner S, Hudson T, Lander E (2001) High-resolution haplotype structure in the human genome. Nature Genetics 29(2):229–232

De Bakker P, Burtt N, Graham R, Guiducci C, Yelensky R, Drake J, Bersaglieri T, Penney K, Butler J, Young S, et al (2006) Transferability of tag SNPs in genetic association studies in multiple populations. Nature Genetics 38(11):1298–1303

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society 39:1–38

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Controlled Clinical Trials 7(3):177–188

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55(4):997–1004

Devlin B, Roeder K, Bacanu S (2001) Unbiased methods for population-based association studies. Genetic Epidemiology 21(4):273–284

Dudbridge F (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Human Heredity 66(2):87–98. Epub 2008 Mar 31

Edwards J (1963) The genetic basis of common disease. The American Journal of Medicine 34(5):627–638

Elston R (1998) Methods of linkage analysis – and the assumptions underlying them. The American Journal of Human Genetics 63(4):931–934

Elston R, Stewart J (1971) A general model for the genetic analysis of pedigree data. Human Heredity 21(6):523–542

Emerson J, Hoaglin D, Mosteller F (1996) Simple robust procedures for combining risk differences in sets of $2x2$ tables. Statistics in Medicine 15(14):1465

Epstein M, Allen A, Satten G (2007) A simple and improved correction for population stratification in case–control studies. The American Journal of Human Genetics 80(5):921–930

Ewens W, Li M, Spielman R (2008) A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. PLoS Genetics 4(9):e1000,180

Fabricius-Hansen V (1939) Blood groups and MN-types of Eskimos in East Greenland. The Journal of Immunology 36:523–530

Falconer D, Mackay T (1996) Heritability. Introduction to Quantitative Genetics pp 160–183. 4th edn. Benjamin Cummings.

Fardo D, Becker K, Bertram L, Tanzi R, Lange C (2009a) Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy–Weinberg equilibrium. European Journal of Human Genetics 17(12):1676–1682

Fardo D, Ionita-Laza I, Lange C (2009b) On quality control measures in genome-wide association studies: a test to assess the genotyping quality of individual probands in family-based association studies and an application to the hapmap data. PLoS Geneties 7:e1000,572. Epub 2009 Jul 24

Feng T, Zhang S, Sha Q (2007) Two-stage association tests for genome-wide association studies based on family data with arbitrary family structure. European Journal of Human Genetics 15:1169–1175

Fisher R (1936) Has Mendel's work been rediscovered? Annals of Science 1:115–137

Fisher R (1965) Introductory notes on Mendel's paper. In: Bennett JH (ed) Experiments in Plant Hybridization, Oxford and Boyd, London

Foulkes A (2009) Applied Statistical Genetics with R: for Population-Based Association Studies. Use R Series. Springer, New York

Gauderman J (2003) Candidate gene association analysis for a quantitative trait, using parent–offspring trios. Genetic Epidemiology 25:327–338

Gauderman W, Murcray C, Gilliland F, Conti D (2007) Testing association between disease and multiple SNPs in a candidate gene. Genetic Epidemiology 31:383–395

Genovese C, Roeder K, Wasserman L (2006) False discovery control with p-value weighting. Biometrika 93(3):509

Giacomini K, Brett C, Altman R, Benowitz N, Dolan M, Flockhart D, Johnson J, Hayes D, Klein T, Krauss R, et al (2007) The pharmacogenetics research network: from SNP discovery to clinical drug response. Clinical Pharmacology & Therapeutics 81(3):328–345

Goate A, Chartier-Harlin M, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L, et al (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349:704–706

Gordon D, Ott J (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pacific Symposium on Biocomputing, vol 2001, pp 18–29

Gordon D, Leal S, Heath S, Ott J (2000) An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. Pacific Symposium on Biocomputing, 663:74

Gordon D, Heath S, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. The American Journal of Human Genetics 69(2):371–380

Guo S (1998) Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. American Journal of Human Genetics 63(1):252–258

Heid I, Huth C, Loos R, Kronenberg F, Adamkova V, Anand S, Ardlie K, Biebermann H, Bjerregaard P, Boeing H, Bouchard C, et al (2009) Meta-analysis of the INSIG2 association with obesity including 74,345 individuals: does heterogeneity of estimates relate to study design? PLoS Genetics 5(10):e1000,694, DOI 10.1371/journal.pgen.1000694

Herbert A, Gerry N, McQueen M, Heid I, Pfeufer A, Illig T, Wichmann EH, Meitinger T, Hunter D, Hu F, Colditz G, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn J, Laird N, Lenburg M, Lange C, Christman M (2006) Genetic variation near insig2 is a common determinant of obesity in Western Europeans and African Americans. Science 312(5771):279–283

Hoffmann T, Lange C, Vansteelandt S, Laird N (2009) Gene–environment interaction tests for dichotomous traits in trios and sibships. Genetic Epidemiology 33(8):691–699

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2):65–70

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. American Journal of Human Genetics 52(2):362–374

Horvath S, Xu X, Laird N (2001) The family based association test method: strategies for studying general genotype–phenotype associations. European Journal of Human Genetics 9(4):301–306

Horvath S, Xu X, Lake S, Silverman E, Weiss S, Laird N (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genetic Epidemiology 26:61–69

Howie B, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6):e1000,529

Huang B, Lin D (2007) Efficient association mapping of quantitative trait loci with selective genotyping. American Journal of Human Genetics 80:567–576

Ibrahim J (1990) Incomplete data in generalized linear models. Journal of the American Statistical Association 85:765–769

International HapMap Consortium, The (2003) The international hapmap project. Nature 426(6968):789–796

International HapMap Consortium, The (2005) A haplotype map of the human genome. Nature 427:1299–1320

International HapMap Consortium, The (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Ionita-Laza I, McQueen M, Laird N, Lange C (2007) Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. The American Journal of Human Genetics 81(3):607–614

Ionita-Laza I, Perry G, Raby B, Klanderman B, Lee C, Laird N, Weiss S, Lange C (2008) On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. Genetic Epidemiology 32(3):273

Javaras K, Laird N, Hudson J, Ripley B (2010) Estimating disease prevalence using relatives of case and control probands. Biometrics 66(1):214–221, Epub 2009 May 18

Knapp M (1999a) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. American Journal of Human Genetics 64:861–870

Knapp M (1999b) Using exact p values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib transmission/disequilibrium test. American Journal of Human Genetics 65(4):1208–1210

Knowler W, Williams R, Pettitt D, Steinberg A (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. American Journal of Human Genetics 43(4):520–526

Korn J, Kuruvilla F, McCarroll S, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins P, Darvishi K, et al (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics 40(10):1253–1260

Kruglyak L (2008) The road to genome-wide association studies. Nature Reviews Genetics 9(4):314–318

Kruglyak L, Lander E (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. American Journal of Human Genetics 57(2):439

Kruglyak L, Daly M, Reeve-Daly M, Lander E (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. American Journal of Human Genetics 58(6):1347

Kwee L, Liu D, Lin X, Ghosh D, Epstein M (2008) A powerful and flexible multilocus association test for quantitative traits. The American Journal of Human Genetics 82:386–397

Laird N, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nature Review Genetics 7(5):385–394

Laird N, Lange C (2009) The role of family-based designs in genome wide association studies. Statistical Science 24(4):388–397

Laird N, Blacker D, Wilcox M (1998) The sib transmission/disequilibrium test is a mantel-haenszel test. American Journal of Human Genetics 63:1915

Laird N, Fitzmaurice G, Schwartz A (2000a) The analysis of case–control data: epidemiologic studies of familial aggregation. Handbook of Environmental and Public Health Statistics 18:465–482

Laird N, Horvath S, Xu X (2000b) Implementing a unified approach to family-based tests of association. Genetic Epidemiology 19:S36

Lake S, Laird N (2004) Tests of gene–environment interaction for case–parent triads with general environmental exposures. Annals of Human Genetics 68(1):55–64

Lander E, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences of the United States of America 84:2363–2367

Lange C, Laird N (2002) Power calculations for a general class of family-based association tests: dichotomous traits. American Journal of Human Genetics 71(3):575–584

Lange C, DeMeo D, Laird N (2002) Power and design considerations for a general class of family-based association tests: quantitative traits. American Journal of Human Genetics 71:1330–1341

Lange C, Silverman E, Xu X, Weiss S, Laird N (2003a) A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics 4:195–206

Lange C, DeMeo D, Silverman E, Weiss S, Laird N (2003b) Using the noninformative families in family-based association tests: a powerful new testing strategy. American Journal of Human Genetics 79:801–811

Lange K (2002) Mathematical and Statistical Methods for Genetic Analysis, 2nd edn. Springer, New York, NY

Lasky-Su J, Faraone S, Lange C, Tsuang M, Doyle A, Smoller J, Laird N, Biederman J (2007) A study of how socioeconomic status moderates the relationship between SNPs encompassing BDNF and ADHD symptom counts in ADHD families. Behavior Genetics 37(3):487–497

Lasky-Su J, Lyon H, Emilsson V, Heid I, Molony C, Raby B, Lazarus R, Klanderman B, Soto-Quiros M, Avila L, et al (2008a) On the replication of genetic associations: timing can be everything! The American Journal of Human Genetics 82(4):849–858

Lasky-Su J, Neale B, Franke B, Anney R, Zhou K, Maller J, Vasquez A, Chen W, Asherson P, Buitelaar J, et al (2008b) Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics 147B(8):1345–1354

Lasky-Su J, Won S, Mick E, Anney R, Franke B, Neale B, Biederman J, Smalley S, Loo S, Todorov A, et al (2010) On genome-wide association studies for family-based designs: an integrative analysis approach combining ascertained family samples with unselected controls. The American Journal of Human Genetics 86(4):573–580

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenburg HJ, et al (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiology. Wiley Online Library

Li B, Leal S (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics 83(3):311–321

Lin D, Huang B (2007) The use of inferred haplotypes in downstream analyses. The American Journal of Human Genetics 80(3):577–579

Lin P, Vance J, Pericak-Vance M, Martin E (2007) No gene is an island: the flip-flop phenomenon. The American Journal of Human Genetics 80(3):531–538

Lipták T (1959) On the combination of independent tests. Magyar Tudományos Akadémia Matematikai Kutató Intezetenek Kozlemenyei 3:1971–1977

Lu A, Cantor R (2007) Weighted variance FBAT: a powerful method for including covariates in FBAT analyses. Genetic Epidemiology 31(4):327–337

Lunetta K, Faraone S, Biederman J, Laird N (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. American Journal of Human Genetics 66:605–614

Lyon H, Emilsson V, Hinney A, Heid I, Lasky-Su J, Zhu X, Thorleifsson G, Gunnarsdottir S, Walters G, Thorsteinsdottir U, Kong A, Gulcher J, Nguyen T, Scherag A, Pfeufer A, Meitinger T, Bronner G, Rief W, Soto-Quiros M, Avila L, Klanderman B, Raby B, Silverman E, Weiss S, Laird N, Ding X, Groop L, Tuomi T, Isomaa B, Bengtsson K, Butler J, Cooper R, Fox C, O'Donnell C, Vollmert C, Celedon J, Wichmann H, Hebebrand J, Stefansson K, Lange C, Hirschhorn J (2007) The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. PLoS Genetics 3(4):e61

Madsen B, Browning S (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics 5(2): e1000384

Mange E, Mange A (1999) Basic Human Genetics, 2nd edn. Sinaeur Associates, Sungerland, MA

Manly B (2007) Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall/CRC, Boca Raton, FL

Manolio T, Brooks L, Collins F (2008) A HapMap harvest of insights into the genetics of common disease. The Journal of Clinical Investigation 118(5):1590

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature 39:906–913

Martin E, Monks S, Warren L, Kaplan N (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. American Journal of Human Genetics 67:146–154

Martin E, Bass M, Kaplan N (2001) Correcting for a potential bias in the pedigree disequilibrium test. The American Journal of Human Genetics 68(4):1065–1067

Mitchell A, Cutler D, Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. The American Journal of Human Genetics 72(3):598–610

Monks S, Kaplan N (2000) Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. American Journal of Human Genetics 66:576–592

Morton N (1955) Sequential tests for the detection of linkage. American Journal of Human Genetics 7(3):277–318

Neel J, Valentine W (1947) Further studies on the genetics of thalassemia. Genetics 32(1):38–63

Newcombe H (1964) Tests for polygenic inheritance. In: Second International Conference on Congenital Malformations. International Medical Congress, New York, NY, p 348

Ott J (1979) Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. The American Journal of Human Genetics 31(2):161

Ott J (1989) Statistical properties of the haplotype relative risk. Genetic Epidemiology 6:127–130

Ott J (1999) Analysis of Human Genetic Linkage. The Johns Hopkins University Press, Baltimore, MD

Ottman R (1990) An epidemiologic approach to gene–environment interaction. Genetic Epidemiology 7(3):177–185

Pearson K (1909) On a new method of determining correlation between a measured character a, and a character b, of which only the percentage of cases wherein b exceeds (or falls short of) a given intensity is recorded for each grade of a. Biometrika 7(1–2):96

Pearson K (1910) On a new method of determining correlation, when one variable is given by alternative and the other by multiple categories. Biometrika 7(3):248

Pearson T, Manolio T (2008) How to interpret a genome-wide association study. JAMA 299(11):1335

Phillips P (2008) Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. Nature Review Genetics 9(11):855–867

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38:904–909

Pritchard J, Przeworski M (2001) Linkage disequilibrium in humans: models and data. American Journal of Human Genetics 69(1):1–14

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81(3):559–575

Rabbee N, Speed T (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22(1):7

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Human Heredity 47(6):342–350

Rabinowitz D, Laird N (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Human Heredity 50(4):211–223

Rakovski C, Xu X, Lazarus R, Blacker D, Laird N (2007) A new multimarker test for family-based association studies. Genetic Epidemiology 31:9–17

Rice W (1990) A consensus combined P-value test and the family-wide significance of component tests. Biometrics 46(2):303–308

Risch N (1990a) Linkage strategies for genetically complex traits. I. Multilocus models. American Journal of Human Genetics 46(2):222–228

Risch N (1990b) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. The American Journal of Human Genetics 46(2):242–253

Roeder K, Bacanu S, Sonpar V, Zhang X, Devlin B (2005) Analysis of single-locus tests to detect gene/disease associations. Genetic Epidemiology 28(3):207–219

Roeder K, Devlin B, Wasserman L (2007) Improving power in genome-wide association studies: weights tip the scale. Genetic Epidemiology 31(7):741–747

Rosner B (1994) Fundamentals of Biostatistics, 4th edn. Duxbury Press, Belmont MA

Sasieni P (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261

Satagopan J, Elston R (2003) Optimal two-stage genotyping in population-based association studies. Genetic Epidemiology 25:149–157

Satagopan J, Venkatraman E, Begg C (2004a) Two-stage designs for gene-disease association studies with sample size contraints. Biometrics 60:589–597

Satagopan J, Verbel D, Venkatraman E, Offit K, Begg C (2004b) Two-stage designs for gene-disease association studies. Biometrics 58:163–170

Schaid D (1998) Transmission disequilibrium, family controls, and great expectations. The American Journal of Human Genetics 63(4):935–941

Schaid D (2001) Evaluating associations of haplotypes with traits. Genetic Epidemiology 27: 348–364

Schaid D, Sommer S (1996) General score tests for association of genetic markers with disease using cases and their parents. Genetic Epidemiology 13:423–449

Searle S (1971) Linear Models. Wiley, New York, NY

Self S, Liang KL (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82:605–610

Shaikh S, Collier D, Sham P, Ball D, Aitchison K, Vallada H, Smith I, Gill M, Kerwin R (1996) Allelic association between a ser-9-gly polymorphism in the dopamine d3 receptor gene and schizophrenia. Human Genetics 97(6):714–719

Sham P (1998) Statistics in Human Genetics. Oxford University Press, New York, NY

Sheskin D (2004) Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, Boca Raton, FL

Shi G, Gu C, Kraja A, Arnett D, Myers R, Pankow J, Hunt S, Rao D (2009a) Genetic effect on blood pressure is modulated by age: the Hypertension genetic epidemiology network study. Hypertension 53(1):35

Shi G, Rice T, Gu C, Rao D (2009b) Application of three-level linear mixed-effects model incorporating gene-age interactions for association analysis of longitudinal family data. BMC Proceedings 3(7):S89

Simes R (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Skol A, Scott L, Abecasis G, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nature 38:209–213

Slager S, Schaid D (2001) Case–control studies of genetic markers: power and sample size approximations for armitage's test for trend. Human Heredity 52(3):149–153

Slatkin M (1999) Disequilibrium mapping of a quatitative-trait locus in an expanding population. American Journal of Human Genetics 64:1765–1773

Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. Heredity 76:377–383

Spielman R, Ewens W (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. American Journal of Human Genetics 62:450–458

Spielman R, McGinnis R, Ewens W (1993) Transmisson test for linkage disequilibrium: the insulin gene region and insulin-dependent Diabetes Mellitus (IDDM). American Journal of Human Genetics 52:506–516

Storey J (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society 64:479–498

Storey J (2003) The positive false discovery rate: a bayesian interpretation and the q-value. Annals of Statistics 31(6):2013–2035

Takei N, Miyashita A, Tsukie T, Arai H, Asada T, Imagawa M, Shoji M, Higuchi S, Urakami K, Kimura H, et al (2009) Genetic association study on in and around the APOE in late-onset Alzheimer disease in Japanese. Genomics 93(5):441–448

Taliaferro W, Huck J (1923) The inheritance of sickle-cell anaemia in man. Genetics 8(6):594–598

Teo Y, Inouye M, Small K, Gwilliam R, Deloukas P, Kwiatkowski D, Clark T (2007) A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 23(20):2741

Thomas D (2004) Statistical Methods in Genetic Epidemiology. Wiley, New York, NY

Thomas D, Xie R, Gebregziabher M (2004) Two-stage sampling designs for gene association studies. Genetic Epidemiology 27:401–414

Thomas D, Casey G, Conti D, Haile R, Lewinger J, Stram D (2009) Methodological issues in multistage genome-wide association studies. Statistical Science 24:414–429

Umbach D, Weinberg C (2000) The use of case-parent triads to study joint effects of genotype and exposure. American Journal of Human Genetics 66(1):251–261

Van Steen K, McQueen M, Herbert A, Raby B, Lyon H, DeMeo D, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman E, Laird N, ST Weiss, Lange C (2005) Genomic screening and replication using the same data set in family-based association testing. Nature Genetics 37:683–691

VanderWeele T (2010) Epistatic interactions. Statistical Applications in Genetics and Molecular Biology 9(1):1–22

Vansteelandt S, DeMeo D, Su J, Smoller J, Murphy A, McQueen M, Schneiter K, Celedon J, Weiss S, Silverman E, Lange C (2008) Testing and estimating gene–environment interactions in family-based association studies. Biometrics 64(2):458–467

Wacholder S, Rothman N, Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. Journal of the National Cancer Institute 92(14):1151–1158

Wald A (1947) Sequential Analysis. Wiley, New York, NY

Wang H, Thomas D, Pe’er I, Stram D (2006) Optimal two-stage genotyping designs for genome-wide association scans. Genetic Epidemiology 30(4):356

Weedon M, Lango H, Lindgren C, Wallace C, Evans D, Mangino M, Freathy R, Perry J, Stevens S, Hall A, et al (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics 40(5):575

Weinberg C (1999) Allowing for missing parents in genetic studies of case-parent triads. American Journal of Human Genetics 64:1186–1193

Won S, Wilk J, Mathias R, O’Donnell C, Silverman E, Barnes K, O’Connor G, Weiss S, Lange C (2009) On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association Studies. PLoS Genetics 5(11):e1000,741. Epub 2009 Nov 26

Xu X, Rakovski C, Xu X, Laird N (2006) An efficient family-based association test using multiple markers. Genetic Epidemiology 30(7):620–626

Yang J, Benyamin B, McEvoy B, Gordon S, Henders A, Nyholt D, Madden P, Heath A, Martin N, Montgomery G, Goddard M, Visscher P (2010) Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. Epub 2010 20 June, doi:10.1038/ng.608.

Yang M (2000) Introduction to Statistical Methods in Modern Genetics, Asian Mathematics Series, vol 3. CRC Press, Boca Raton, FL

Zhang S, Zhu X, Zhao H (2003) On a semi-parametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genetic Epidemiology 24:44–56

Zheng G, Song K, Elston R (2007) Adaptive two-stage analysis of Genetic association in case–control designs. Human Heredity 63(3–4):175–186

Zhu X, Zhang S, Zhao H, Cooper R (2002) Association mapping using a mixture model for complex traits. Genetic Epidemiology 23:181–196

Index