

## Genetic Association Studies

Recent technological advancements allowing for large-scale sequencing efforts present an exciting opportunity to uncover the genetic underpinnings of complex diseases. In an attempt to characterize these genetic contributors to disease, investigators have embarked in multitude on what are commonly referred to as *population-based genetic association studies*. These studies generally aim to relate genetic sequence information derived from unrelated individuals to a measure of disease progression or disease status. The field of genomics spans a wide array of research areas that involve the many stages of processing from genetic sequence information to protein products and ultimately the expression of a trait. The breadth of genomic investigations also includes studies of multiple organisms, ranging from bacteria to viruses to parasites to humans. In this chapter, two settings are described in which population-based genetic association studies have marked potential for uncovering disease etiology while elucidating new approaches for targeted, individualized therapeutic interventions: (1) complex disease association studies in humans; and (2) studies involving the Human Immunodeficiency Virus (HIV).

In both settings, interest lies in characterizing associations between multiple genetic polymorphisms and a measured trait. In addition, these settings share the essential need to account appropriately for patient-level covariates as potential confounders or modifiers of disease progression to make clinically meaningful conclusions. While these two settings are not comprehensive, together they provide a launching point for discussion of quantitative methods that address the challenges inherent in many genetic investigations. This chapter begins by describing types of population-based studies, which represent one class of investigations within the larger field of genomics research. Also discussed are the fundamental features of data arising from these investigations as well as the analytical challenges inherent in this endeavor.

## 1.1 Overview of population-based investigations

Population-based genetic association studies can be divided roughly into four categories of studies: candidate polymorphism, candidate gene, fine mapping and whole or partial genome-wide scans. In the following paragraphs, each of these types of studies is described briefly, followed by a discussion of how population-based genetic investigations fit within a larger context of genomic-based studies. Further discussions of population-based and family-based designs can be found in Thomas (2004) and Balding (2006).

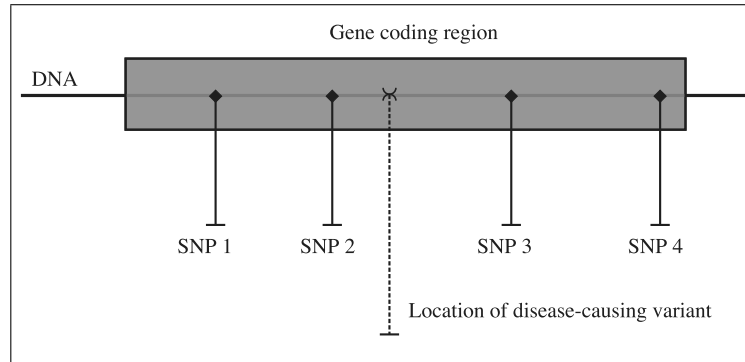
### 1.1.1 Types of investigations

#### *Candidate polymorphism studies*

Investigations of genotype–trait associations for which there is an *a priori* hypothesis about functionality are called *candidate polymorphism* studies. Here the term *polymorphism* is defined simply as a genetic variant at a single location within a gene. Technically, a variation must be present in at least 1% of a population to be classified as a polymorphism. Such a variable site is commonly referred to as a single-nucleotide polymorphism (SNP). Candidate polymorphism studies typically rely on prior scientific evidence suggesting that the set of polymorphisms under investigation is relevant to the disease trait. The aim is to test for the presence of association, and the primary hypothesis is that the variable site under investigation is *functional*. That is, the goal of candidate polymorphism studies is to determine whether a given SNP or set of SNPs influences the disease trait directly.

#### *Candidate gene studies*

*Candidate gene* studies generally involve multiple SNPs within a single gene. The choice of SNPs depends on defined linkage disequilibrium (LD) blocks and is discussed further in Section 3.1. The underlying premise of these studies is that the SNPs under investigation capture information about the underlying genetic variability of the gene under consideration, though the SNPs may not serve as the true disease-causing variants. That is, the SNPs that are being studied are not necessarily functional. Consider for example a setting in which we want to investigate the association between a gene and disease. A gene comprises a region of deoxyribonucleic acid (DNA), representing a portion of the human genome. This is illustrated by the shaded rectangle in Figure 1.1. In a simple model, we might assume that a mutation at a single site within this region results in disease. In general, the precise location of this disease-causing variant is not known. Instead, investigators measure multiple SNPs that are presumed “close” to this site on the genome. The term “close” can be thought of as physical distance, though precise methods for choosing appropriate SNPs are described in more detail in Section 3.1.



**Fig. 1.1.** Marker SNPs

These proximate SNPs are commonly referred to as *markers* since the observed genotype at these locations tends to be associated with the genotype at the true disease-causing locus. The idea underlying this phenomenon is that, over evolutionary time (that is, over many generations of reproduction), the disease allele was inherited alongside variants at these marker loci. This occurs when the probability of a recombination event in the DNA region between the disease locus and the marker locus is small. Thus, capturing variability in these loci will tend to capture variability in the true disease locus. Further discussion of recombination is provided in Section 1.3.1.

#### *Fine mapping studies*

The aim of *fine mapping* studies tends to differ from those of candidate gene and candidate polymorphism approaches. Fine mapping studies set out to identify, with a high level of precision, the *location* of a disease-causing variant. That is, these studies aim to determine precisely where on the genome the mutation that causes the disease is positioned. Knowledge about this location can obviate the need for investigations based on marker loci, thus reducing the error and variability in associated tests. Within the context of mapping studies, the term *quantitative trait loci (QTL)* is used to refer to a chromosomal position that underlies a trait. Methods for mapping and characterizing QTLs based on controlled experiments of inbred mouse lines are described in Chapter 15 of Lynch and Walsh (1998). Mapping studies are not a focal point of this text; however, we note that in some contexts the term “mapping” is used more loosely to refer to association, the topic of this text, in both family- and population-based studies. For comprehensive and advanced coverage of gene mapping methods, the reader is referred to Siegmund and Yakir (2007).

#### *Genome-wide association studies (GWAS)*

Similar to candidate gene approaches, studies involving whole and partial genome-wide scans, termed *genome-wide association studies (GWAS)*, aim

to identify associations between SNPs and a trait. GWAS, however, tend to be less hypothesis driven and involve the characterization of a much larger number of SNPs. Partial scans generally involve between 100Kb and 500Kb segments of DNA, while whole-genome scans range from 500Kb to 1000Kb regions. While the underlying goal of candidate gene studies and GWAS can be similar, the data preprocessing is generally more extensive and the computational burden greater in the context of GWAS, requiring the application of software packages designed specifically to address the high-dimensional nature of the data, as described in Section 3.3. While GWAS have gained in popularity in recent years due to the advent and widespread availability of “SNP chips”, they do not obviate the need for candidate gene studies. Candidate gene studies serve to validate findings from GWAS as well as further explore the biological and clinical interactions between genes and more traditional risk factors for complex diseases, such as age, gender, and other patient-level clinical and demographic characteristics. Importantly, the fundamental statistical concepts and methods described throughout this text are broadly relevant to both candidate gene studies and GWAS.

### 1.1.2 Genotype versus gene expression

The term “association” study has come to refer to studies that consider the relationship between genetic *sequence* information and a phenotype. Gene expression studies, based on microarray technology, on the other hand, aim to characterize associations among gene *products*, such as ribonucleic acid (RNA) or proteins, and disease outcomes. While the scientific findings from these investigations will likely lend support to one another, it is important to recognize that the two types of studies focus on different aspects of the cell life cycle. In the context of association studies, the raw genetic information as characterized by the DNA sequence is the primary predictor variable under investigation, and the aim is to understand how polymorphisms in the sequences explain the variability in a disease trait. Gene expression studies, on the other hand, focus on the extent to which a DNA sequence coding for a specific gene is transcribed into RNA (transcriptomics) and then translated into a protein product (proteomics). The former arises from gene chip technology and is commonly referred to as expression data, while the latter is an output of mass spectrometry. Since transcription and translation depend on many internal and external regulation factors, the expression of a gene sequence represents a different phenomenon than the sequence itself.

A fundamental unit of analysis in population association studies is the *genotype*. As described in Section 1.2, genotype is a categorical variable that takes on values from a predefined set of discrete characters. For example, in humans, most SNPs are *biallelic*, indicating there are two possible bases at the corresponding site within a gene (e.g., *A* and *a*). Furthermore, since humans are *diploid*, each individual will carry two bases, corresponding to each of two homologous chromosomes. As a result, the possible genotype values

in the population are  $AA$ ,  $Aa$  and  $aa$ . In studies of gene expression, on the other hand, the basic unit used in analysis is the gene product, which is typically a real-valued positive number. Notably, investigators may subsequently dichotomize this variable, though this additional level of data processing will depend on the scientific questions under consideration and prior knowledge.

In both settings, a measure of disease status or disease progress, referred to as the *trait* in this text, is also collected for analysis. Notably, in population association studies, we generally treat the genotype as the *predictor* variable and the trait as the *dependent* variable. In gene expression studies, this may or may not be the case. Consider for example the setting in which investigators aim to uncover the association between breast cancer and gene expression. In this case, the expression of a gene, as measured by how much RNA is produced, may serve as the main dependent variable, with cancer status as the potential predictor. The alternative formulation is also tenable. In this text, since emphasis is on population-based association studies, it is always assumed that genotype precedes the trait in the causal chain.

While careful consideration must be given to the several notable differences in the form as well as the interpretation of the data, many of the statistical methods described herein are equally applicable to gene expression studies. In the context of genotype data, we might for example test the null hypothesis that cholesterol level is the same for individuals with genotype  $AA$  and genotype  $aa$ . In the expression setting, the null hypothesis may instead be framed as the gene expression level is the same for individuals with cardiovascular disease and those without cardiovascular disease. In both cases, a two-sample test for equality of means or medians (e.g., the two-sample  $t$ -test or Wilcoxon rank sum test) could be performed and similar approaches to account for multiple testing employed. Notably, preprocessing of gene expression data prior to formal statistical analysis also has its unique challenges. Several seminal texts provide discussion of statistical methods for the analysis of gene expression data. See for example Speed (2003), Parmigiani *et al.* (2003), McLachlan *et al.* (2004), Gentleman *et al.* (2005) and Ewens and Grant (2006).

Finally, we distinguish between genetic association studies and the rapidly growing field of research in epigenetics. The term *epigenetics* is used to describe heritable features that control the functioning of genes within an individual cell but do not constitute a physical change in the corresponding DNA sequence. The *epigenome*, defined literally as “above-the-genome”, also referred to as the *epigenetic code*, includes information on methylation and histone patterns, called *epigenetic tags*, and plays an essential role in controlling the expression of genes. These tags can inhibit and silence genes, leading to common complex diseases such as cancer. In this text, we consider traditional epidemiological risk factors, such as smoking status and diet, that may play a role in defining an individual’s epigenetic makeup; however, we do not address directly the challenges of epigenetic data. For a further discussion of the role of epigenetics in the link between environmental exposures and disease phenotypes, see Jirtle and Skinner (2007).

### 1.1.3 Population-versus family-based investigations

The term “population”-based is used to refer to investigations involving unrelated individuals and distinguished from family-based studies. The latter, as the name implies, involves data collected on multiple individuals within the same family unit. The statistical considerations for family-based studies differ from those of population-based investigations in two primary regards. First, individuals within the same family are likely to be more similar to one another than are individuals from different families. This phenomenon is referred to in statistics as clustering and implies a within-family correlation. The idea is that there is something unmeasurable (latent), such as diet or underlying biological makeup, that makes people from the same family more alike than people across families. As a result, the trait under investigation is more highly correlated among individuals within the same family. Accounting for the potential within-cluster correlation in the statistical analysis of family-based data is essential to making valid inference in these settings.

In population-based studies, a fundamental assumption is that individuals are unrelated; however, other forms of clustering may exist. For example, individuals may have been recruited across multiple hospitals so that patients from the same hospital are more similar than those across hospitals. This within-cluster correlation can arise particularly if the catchment areas for the hospitals include different socioeconomic statuses or if the standards for patient care are remarkably different. Alternatively, we may have repeated measurements of a trait on the same individual. This is another common situation in which the assumption of independence is violated. In all of these cases, analytical methods for correlated data are again warranted and are essential for correctly estimating variance components. In this text, attention is restricted primarily to methods for independent observations, though consideration is given to clustered data methods in Section 4.4.2. Tests for relatedness are also described in Section 3.3. In-depth and comprehensive coverage of correlated data methods can be found in Diggle *et al.* (1994), Vonesh and Chinchilli (1997), Verbeke and Molenberghs (2000), Pinheiro and Bates (2000), McCulloch and Searle (2001), Fitzmaurice *et al.* (2004) and Demidenko (2004).

A second remarkable difference between population- and family-based studies involves what is termed *allelic phase* and is defined as the alignment of nucleotides on a single homolog. Allelic phase is typically unobservable in population-based association studies but can often be determined in the context of family studies. This concept is described in greater detail in Section 1.2 and Chapter 5. As a result of these differences in the data structure, the methods for analysis of family-based association studies tend to differ from those developed in the context of population-based studies. Though some of the methods described herein, particularly adjustments for multiplicity, are applicable to family-based studies, this text focuses on methods specifically relevant to population association studies, including inferring haplotypic phase (Chapter 5). Elaboration on the specific statistical considerations and methods for

family-based studies can be found in Khoury *et al.* (1993), Liu (1998), Lynch and Walsh (1998), Thomas (2004), Siegmund and Yakir (2007) and Ziegler and Koenig (2007).

### 1.1.4 Association versus population genetics

Finally, we distinguish between population-based association studies (the topic of this text) and *population genetic* investigations. Population genetics refers generally to the study of changes in the genetic composition of a population that occur over time and under evolutionary pressures. This includes, for example, the study of natural selection and genetic drift. In this text, we instead focus on estimation and inference regarding the association between genetic polymorphisms and a trait. Statistical methods relevant to population genetics are described in a number of texts, including Weir (1996), Gillespie (1998) and Ewens and Grant (2006).

## 1.2 Data components and terminology

Data arising from population-based genetic association studies are generally comprised of three components: (1) the *genotype* of the organism under investigation; (2) a single trait or multiple *traits* (also referred to as *phenotypes*) that are associated with disease progression or disease status; and (3) patient-specific *covariates*, including treatment history and additional clinical and demographic information. The primary aim of many association studies is to characterize the relationship between the first two of these components, the genotype and a trait. *Pharmacogenomic* investigations aim specifically to analyze how genotypes modify the effects of drug exposure (the third data component) on a trait. That is, these investigations focus on the statistical interaction between treatment and genotype on a disease outcome. While the specific aims of many association studies do not expressly involve the third data component, patient-specific clinical and demographic information, careful consideration of how these factors influence the relationship between the genotype and trait is essential to making valid biological and clinical conclusions. In this chapter, we describe each of these data components, all of which are highly relevant to population-based association studies, and introduce some additional terminology. A discussion of the potential interplay among components of the data and important epidemiological principles, including confounding, effect mediation, effect modification and conditional association, is provided in Section 2.1.2. Further elaboration on the concept of phase ambiguity and appropriate statistical approaches to handling this aspect of the data are given in Chapter 5.

### 1.2.1 Genetic information

Throughout this text, the term *genotype* is defined as the observed genetic sequence information and can be thought of as a categorical variable. The term *observed* is used here to distinguish genotype information from haplotype data, as described below. Humans carry two *homologous chromosomes*, which are defined as segments of deoxyribonucleic acid (DNA), one inherited from each parent, that code for the same trait but may carry different genetic information. Thus, in its rawest form in humans, the genotype is the pair of DNA bases adenine (A), thymine (T), guanine (G) and/or cytosine (C) observed at a location on the organism's genome. This pair includes one base inherited from each of the two parental genomes and should not be confused with the pairing that occurs to form the DNA double helix. These two types of pairing are described further in Section 1.3.1. Genotype data can take different forms across the array of genetic association studies and depend both on the specific organism under investigation and the scientific questions being considered, as we will see throughout this text.

The term *nucleotide* refers to a single DNA base linked with both a sugar molecule and phosphate and is often used interchangeably with the term *DNA base*. *Genes* are defined simply as regions of DNA that are eventually made into proteins or are involved in the regulation of transcription; that is, regions that regulate the production of proteins from other segments of DNA. In *candidate gene* studies, the set of genes under investigation is chosen based on known biological function. These genes may, for example, be involved in the production of proteins that are important components of one or more pathways to disease. In whole and partial *genome-wide association studies* (*GWAS*), segments of DNA across large regions of the genome are considered and may not be accompanied by an *a priori* hypothesis about the specific pathways to disease.

In population-based association studies, the fundamental unit of analysis is the single-nucleotide polymorphism (SNP). A *SNP* simply describes a single base pair change that is variable across the general population at a frequency of at least 1%. The term can also be used more loosely to describe the specific location of this variability. The overriding premise of association studies is that there exists variability in DNA sequences across individuals that captures information on a disease trait. Regions of DNA within and across genes are said to have *genetic variability* if the alleles within the region vary across a population. *Conserved* regions, on the other hand, exhibit no variability in a population. Take the simple example of a single base pair location within a gene. If the genotype at this site is *AA* for all individuals within the population, then this site is referred to as conserved. On the other hand, if *AA*, *Aa* and *aa* are observed, then this site is called variable. Here the letters *A* and *a* are used to represent the observed nucleotides (A, C, T or G). For example, *A* may represent adenine (A) and *a* may represent thymine (T). Further discussion of notation is provided in Section 2.1.1. Highly conserved regions of DNA

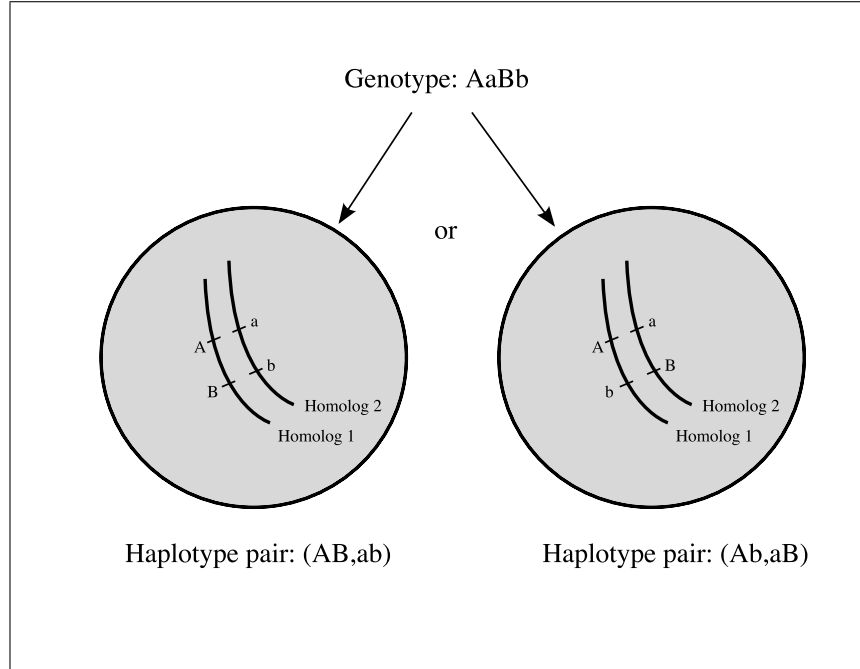


are less relevant in the context of association studies since they will not be able to capture the variability in the disease trait. Studying highly conserved regions would be tantamount in a traditional epidemiological investigation to only recruiting smokers to a study and then trying to assess the impact of smoking on cancer risk. Clearly, multiple levels of the predictor variable, in this case smoking status, are necessary if the goal is to assess the impact of this factor on disease.

*Multilocus genotype* is used to describe the observed genotype across multiple SNPs or genes, though the terms *genotype* and *multilocus genotype* are often used interchangeably. A *locus* or *site* can refer to the portion of the genome that encodes a single gene or the location of a single nucleotide on the genome. Multilocus genotype data consist of a string of categorical variables, with elements corresponding to the genotype at each of multiple sites on the genome. For example, an individual's multilocus genotype may be given by  $(Aa, Bb)$ , where  $Aa$  is the genotype at one site and  $Bb$  is the genotype at a second site. Again the letters  $A$ ,  $a$ ,  $B$  and  $b$  each represent the observed nucleotides (A, C, T or G). Notably, the specific ordering of alleles is non-informative, so, for example, the genotypes  $Aa$  and  $aA$  are equivalent.

The term multilocus genotype should not be confused with the concept of *haplotype*. Haplotype refers to the specific combination of alleles that are in *alignment* on a single *homolog*, defined as one of the two homologous chromosomes in humans. Suppose again that an individual's multilocus genotype is given by  $(Aa, Bb)$ . The corresponding pair of haplotypes, also referred to as this individual's *diplotype*, could be  $(AB, ab)$  or  $(Ab, aB)$ . That is, either the  $A$  and  $B$  alleles are in alignment on the same homolog, in which case  $a$  and  $b$  align, or the  $A$  and  $b$  alleles align, in which case  $a$  and  $B$  are in alignment. These two possibilities are illustrated in Figure 1.2 and described further in Section 2.3.2. This uncertainty is commonly referred to as *ambiguity in allelic phase* or more simply *phase ambiguity*. In general, a multilocus genotype is observable, although missing data can arise from a variety of mechanisms. Haplotype data, on the other hand, are generally *unobservable* in population-based studies of unrelated individuals and require special consideration for analysis, as described in detail in Chapter 5.

This layer of missingness renders population-based association studies unique from family-based investigations. If parental information were available on the individual above, then it might be possible to clarify the uncertainty in allelic phase. For example, if the maternal genotype is  $(AA, BB)$  and the paternal genotype is  $(aa, bb)$ , then it is clear that  $A$  and  $B$  align on the same homolog that was inherited from the maternal side and the  $a$  and  $b$  align on the copy inherited from the paternal side. In population-based studies, family data are generally not available to infer these haplotypes. However, it is possible to draw strength from the population haplotype frequencies to determine the most likely alignment for an individual. This is discussed in greater detail in Chapter 5.



**Fig. 1.2.** Haplotype pairs corresponding to heterozygosity at two SNP loci

The term *zygosity* refers to the comparative genetic makeup of two homologous chromosomes. An individual is said to be *homozygous* at a given SNP locus if the two observed base pairs are the same. *Heterozygosity*, on the other hand, refers to the presence of more than one allele at a given site. For example, someone presenting with the  $AA$  or  $aa$  genotype would be called homozygous, while an individual with the  $Aa$  is said to be heterozygous at the corresponding locus. The term *loss of heterozygosity (LOH)*, commonly used in the context of oncology, refers specifically to the loss of function of an allele, when a second allele is already inactive, through inheritance of the heterozygous genotype.

The *minor allele* frequency, also referred to as the *variant allele* frequency, refers to the frequency of the less common allele at a variable site. Note that here the term *frequency* is used to refer to a population proportion, while statisticians tend to use the term to refer to a count. The terms *homozygous rare* and *homozygous variant* are commonly used to refer to homozygosity with two copies of the minor allele. Consider the simple example of a single-variable site for which  $AA$  is present in 75% of the population,  $Aa$  is present in 20% and  $aa$  is present in 5%. The frequency of the  $A$  allele is then equal to  $(75 + 75 + 20)/2 = 85\%$ , while the frequency of  $a$  is  $(20 + 5 + 5)/2 = 15\%$ . In this case, the minor allele ( $a$ ) frequency is equal to 15%. The *major allele*

is the more common allele and is given by  $A$  in this example. An example of calculating the minor and major allele frequencies in R is provided in Section 1.3.3.

### 1.2.2 Traits

Population-based genetic association studies generally aim to relate genetic information to a *clinical outcome* or *phenotype*, which are both referred to in this text as a *trait*. The terms *quantitative* and *binary* traits refer respectively to continuous and binary variables, where a binary variable is defined as one that can take on two values, such as diseased or not diseased. The term *phenotype* is defined formally as a physical attribute or the manifestation of a trait and in the context of association studies generally refers to a measure of disease progression. In the context of viral genetic investigations, phenotypes typically refer to an *in vitro* measure such as the 50% inhibitory concentration ( $IC_{50}$ ), which is defined as the amount of drug required to reduce the replication rate of the virus by 50%. The term *outcome* tends to mean the presence of disease, though it is often used more generally in a statistical sense to refer to any dependent variable in a modeling framework.

Clinical measures such as total cholesterol and triglyceride levels are examples of quantitative traits, while the indicator for a cardiovascular outcome, such as a heart attack, is an example of a binary trait. In a study of breast cancer, the trait may be defined as an indicator for whether or not a patient has breast cancer. In HIV investigations, traits include viral load (VL), defined as the concentration of virus in plasma, and CD4+ cell count, which is a marker for disease progression. In this text, the terms *trait*, *phenotype* and *outcome* are used broadly to refer to both *in vitro* and *in vivo* clinical measures of disease progression and disease status. Survival outcomes, such as the time to onset of AIDS, time to a cardiovascular event, or time to death, as well as ordinal outcomes, such as severity of disease, are other examples of traits that are also highly relevant to the study of genetic associations with disease. While this text focuses on continuous and binary traits, alternative formulations apply and the general methodology presented is applicable to a wider array of measures.

Traits can be measured cross-sectionally or over multiple time points spanning several weeks to several years. Data measured over time are referred to as *longitudinal* or *multivariate* data and provide several advantages from an analytical perspective, as discussed in detail in several texts, including Fitzmaurice *et al.* (2004). The choice of using cross-sectional or longitudinal data rests primarily on the scientific question at hand. For example, if interest lies in determining whether genotype affects the change in VL after exposure to a specific drug, then a longitudinal design with repeated measures of VL is essential. On the other hand, if the interest is in characterizing VL as a function of genotype at initiation of therapy, then cross-sectional data may be

sufficient. While longitudinal studies can increase the power to detect association, they tend to be more costly than cross-sectional studies and are more susceptible to missing data and the resulting biases. In this text, we focus on the analysis of cross-sectional studies, though the overarching themes and concepts, such as multiple testing adjustments and the need to control type-1 error rates, are equally applicable to alternative modeling frameworks.

### 1.2.3 Covariates

In addition to capturing information on the genotype and trait, population-based studies generally involve the collection of other information on patient-specific characteristics. For example, in relating genetic polymorphisms to total cholesterol level among patients at risk for cardiovascular disease, additional relevant information may include body mass index (BMI), gender, age and smoking status. The additional data collected tend to be on variables that have previously been associated with the trait of interest, in this case cholesterol level, and may include environmental, demographic and clinical factors. Consideration of additional variables in the context of analysis will again depend on the scientific question at hand, the biological pathways to disease and the overarching goal of the analysis. For example, if the aim of a study is to identify the best predictive model (that is, to determine the model that can give the most accurate and precise prediction of cholesterol level for a new individual), then it is generally a good idea to include variables previously associated with the outcome in the model. If the goal is to characterize the association between a given gene and the outcome, then including additional variables, for example self-reported race, may also be warranted if these variables are associated with both the genotype and the outcome. This phenomenon is typically referred to as *confounding* and is discussed in greater detail in Chapter 2. On the other hand, if a variable such as smoking status is in the *causal pathway* to disease (that is, the gene under investigation influences the smoking status of an individual, which in turn tends to increase cholesterol levels), then inclusion of smoking status in the analysis may not be appropriate. In this text, the term *covariate* is used loosely to refer to any explanatory variables that are not of specific independent interest in the present investigation. Covariates are also commonly referred to as *independent* or *predictor* variables.

## 1.3 Data examples

Throughout this textbook, we provide examples using publicly available datasets, including data arising from two human-based investigations and one study involving HIV. Each of these datasets can be downloaded as ascii text files from the textbook website:

<http://people.umass.edu/foulkes/asg.html>

Below we include a summary of each dataset and example code for importing the data into R. Instructions for downloading R, inputting data and basic data manipulation strategies are given in the appendix. Additional elementary R concepts can be found in Gentleman (2008), Spector (2008), Venables and Smith (2008) and Dalgaard (2002). Complete information on all of the variables within each dataset can be found in the associated ReadMe files on the textbook website.

The two settings described in this section, complex disease association studies in humans and HIV genotype–trait association studies, serve as a framework for the methods presented throughout the text. While both the structure of the data and the overarching aims of the two settings are similar, there are a few notable differences worth mentioning. In both settings, belief lies in the idea that genetic polymorphisms (that is, variability in the genetic makeup across a population) will inform us about the variability observed in the occurrence or presentation of disease. Furthermore, this genetic variability in both HIV and humans is introduced through the process of replication. The *rate* at which these two organisms complete one life cycle, however, is dramatically different. While humans tend to replicate over the course of several years, an estimated  $10^9$  to  $10^{10}$  new virions are generated in a single day within an HIV-infected individual. Furthermore, the replication process for HIV, described in more detail below, is highly error-prone, resulting in a mutation rate of approximately  $3 \times 10^{-5}$  per base per replication cycle, see for example Robertson *et al.* (1995).

As a result, there is a tremendous degree of HIV genetic variability within a single human host. That is, each HIV-infected individual carries an entire population of viruses, with each viral particle potentially comprised of different genetic material. In addition, the number of viral particles varies across individuals. Notably, both of these phenomena, genetic variability and the amount of virus in plasma, are influenced by current and past drug exposures. In contrast, humans carry two copies of each chromosome, with the exception of the sex chromosome, one inherited from each parent, and these *tend* to remain constant over an individual's lifetime. While relatively rare, mutations in the human genome do occur within a lifespan as a result of environmental exposure to mutagens. This process is notably slower in humans than in HIV and is not a focal point of this text. Additional details on each of these two settings are provided below.

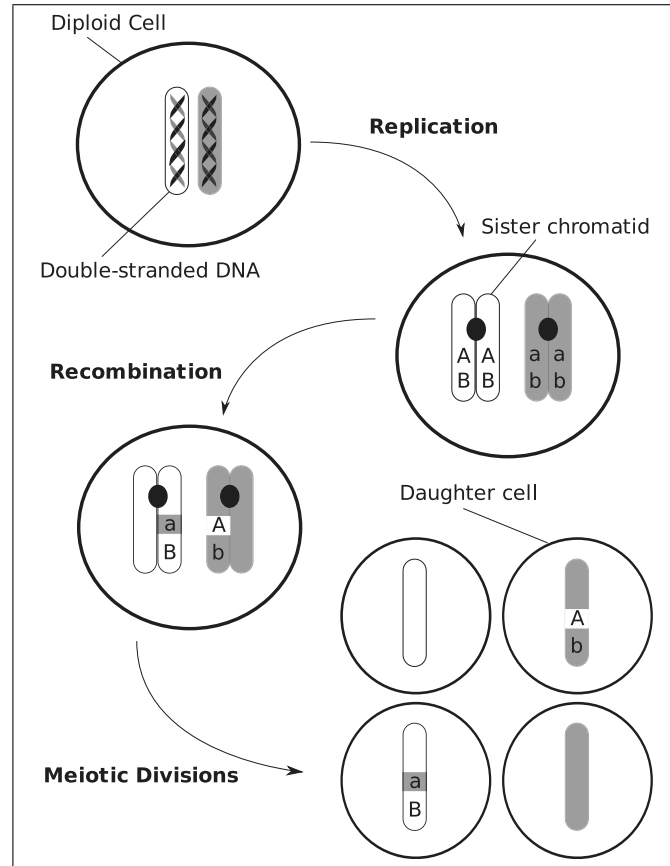
### 1.3.1 Complex disease association studies

Characterizing the underpinnings of complex diseases, such as cardiovascular disease and cancer, is likely to require consideration of multiple genetic and environmental factors. As described in Section 1.1.1, human genetic investigations can involve several stages of processing of human genes, from the

DNA sequence to the protein product, and encompass a wide assortment of study designs. In this text, consideration is given to population-based studies of unrelated individuals, and the primary unit of genetic analysis is the DNA sequence. Humans inherit their genetic information from their two parental genomes through processes termed *mitosis* and *meiosis*. All human cells, with the exception of gametes, contain 46 chromosomes, including 22 homologous pairs, called *autosomes*, and 2 sex chromosomes. Each chromosome is comprised of a DNA double helix with two sugar-phosphate backbones connected by paired bases. In this context, guanine pairs with cytosine (G-C) and adenine pairs with thymine (A-T). This pairing is distinct from the pairing of homologous chromosomes that constitutes an individual's genotype. Notably, the latter pairing is not restricted, so that, for example, genotypes *GT* and *AC* can be observed.

*Mitosis* is a process of cell division that results in the creation of daughter cells that carry identical copies of this complete set of 46 chromosomes. *Meiosis* is the process by which a germ cell that contains 46 chromosomes, consisting of one homolog from each parent cell, undergoes two cell divisions, resulting in daughter cells, called *gametes*, with only 23 chromosomes each. In turn, this new generation of maternal and paternal gametes combines to form a *zygote*. A visual representation of meiosis is provided in Figure 1.3. Notably, prior to the meiotic divisions, each of the two homologous chromosomes are replicated to form *sister chromatid*. Subsequently, in the process of meiosis, cross-over between these maternal and paternal chromatids can occur. This is referred to as a *cross-over* or a *recombination event* and is depicted in the figure, where we see an exchange of segments of the paternal chromatid (shaded) and the maternal chromatid (unshaded). Finally, it is important to note that the 23 chromosomes are combined independently so that there are  $2^{23} = 8,388,608$  possible combinations of chromosomes within a gamete. This phenomenon is commonly referred to as *independent assortment*. The reader is referred to any of a number of excellent textbooks that describe these processes in greater detail. See for example Chapter 19 of Vander *et al.* (1994) and Alberts *et al.* (1994).

Meiosis ensures two things: (1) each offspring carries the same number of chromosome pairs (23) as its parents; and (2) the genetic makeup of offspring is not identical to that of their parents. The latter results from both recombination and independent assortment. An important aspect of meiosis is that whole portions or *segments* of DNA within a chromosome tend to be passed from one generation to another. However, portions of DNA within chromosomes that are far from one another are less likely to be inherited together, as a result of recombination events. In the context of candidate gene studies, the SNPs under investigation can be known *functional* SNPs or what are referred to as *haplotype tagging* SNPs. Functional SNPs affect a trait directly, serving as a component within the causal pathway to disease. Haplotype tagging SNPs, on the other hand, are chosen based on their ability to capture overall variability within the gene under consideration. These SNPs tend to be associated



**Fig. 1.3.** Meiosis and recombination

with functional SNPs but may not be causal themselves. Notably, the length of a gene region can vary as well as the number of measured base pairs within each gene. The latter depends on what are called *linkage disequilibrium* blocks and relate to the probability of recombination within a region. This is described further in Section 3.1.

The structure of human genetics data is similar to that in the HIV setting, with a couple of notable exceptions. First, in human investigations, each individual has exactly two bases present at each location, one from each of the two homologous chromosomes. As described below in Section 1.3.2, in the viral genetics setting, an individual can be infected with multiple strains, resulting in any number of nucleotides at a given site. A second difference is that in many population-based association studies, human genetic sequence data are assumed to remain constant over the study period. One notable exception

is in the context of cancer, in which DNA damage develops, resulting from environmental exposure to mutagens and resulting in uncontrolled cell proliferation. In the complex disease association studies described in this text, the genes under investigation do not vary within the timeframe of study. This is a marked difference from the viral genetic setting, in which multiple genetic polymorphisms can occur within a short period of time, typically in response to treatment pressures. In the following section, we describe the HIV genetic setting in greater detail.

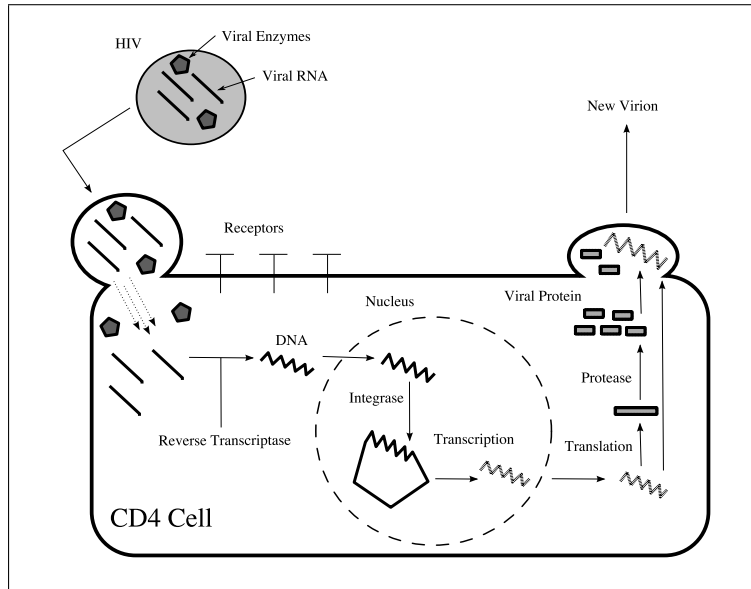
### 1.3.2 HIV genotype association studies

The *Human Immunodeficiency Virus (HIV)* is a retrovirus that causes a weakening of the immune system in its infected host. This condition, commonly referred to as *Acquired Immunodeficiency Syndrome (AIDS)*, leaves infected individuals vulnerable to opportunistic infections and ultimately death. The World Health Organization estimates that there have been more than 25 million AIDS-related deaths in the last 25 years, the majority of which occurred in the developing world. Highly active *anti-retroviral therapies (ARTs)* have demonstrated a powerful ability to delay the onset of clinical disease and death, but unfortunately access to these therapies continues to be severely limited. Furthermore, drug resistance, which can be characterized by mutations in the viral genome, reduces and in some cases eliminates their usefulness. Both vaccine and drug development efforts, as well as treatment allocation strategies in the context of HIV/AIDS, will inevitably require consideration of the genetic contributors to the onset and progression of disease. In this section, the viral life cycle and notable features of the data relevant to these investigations are described.

A visual representation of the HIV life cycle is given in Figure 1.4. As a retrovirus, HIV is comprised of ribonucleic acid (RNA). From the figure, we see that the virus begins by fusing on the membrane of a CD4+ cell in the human host and injecting its core, which includes viral RNA, structural proteins, and enzymes, into the cell. The viral RNA is then reverse transcribed into DNA using one of these enzymes, *reverse transcriptase*. Another enzyme, *integrase*, then splices this viral DNA into the host cell DNA. The normal cell mechanisms for transcription and translation then result in the production of new viral protein. In turn, this protein is cleaved by the *protease (Pr)* enzyme and together with additional viral RNA forms a new virion. As this virion buds from the cell, the infected cell is killed, ultimately leading to the depletion of CD4 cells, which are vital to the human immune system. ARTs, the drugs used to treat HIV-infected individuals, aim to inhibit each of the enzymes involved in this life cycle.

Reverse transcription of RNA into DNA is a highly error-prone process, resulting in a mutation rate of approximately  $3 \times 10^{-5}$  per base per cycle. This, coupled with a very fast replication cycle leading to  $10^9$  to  $10^{10}$  new virions each day, results in a very high level of genetic variability in the viral genome.





**Fig. 1.4.** HIV life cycle

The resulting viral population within a single human host is commonly referred to as a *quasi-species*. While many of these viruses are not viable (that is, they cannot survive with the resulting mutations), many others do remain. Notably, evidence suggests that mutated viruses can be transmitted from one host to another. The composition of a viral quasi-species tends to be highly influenced by current and past treatment exposures. HIV therapies generally consist of a combination of two or three anti-retroviral drugs, commonly referred to as a *drug cocktail*. There are currently four classes of drugs that each target a different aspect of the viral life cycle: fusion inhibitors, nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs). In the presence of these treatment pressures, viruses that are resistant to the drugs tend to emerge as the dominant species within a person. As individuals develop resistance to one therapy, another combination of drugs may be administered and a new dominant species can emerge. Evidence suggests that a blueprint of drug exposure history remains in latent reservoirs in the sense that a resistant species will re-emerge quickly in the presence of a drug to which a patient previously exhibited resistance.

The genetic composition of HIV is a single strand of RNA consisting of the four base pairs adenine (A), cytosine (C), guanine (G) and uracil (U). In general, and for the purpose of this textbook, the *amino acid* (AA) corresponding to three adjacent bases is of interest since AAs serve as the building blocks for proteins. Notably, there is not a one-to-one correspondence between

base triplets and AAs, and thus there are instances in which base information is more relevant, for example in phylogenetic analyses aimed at characterizing viral evolution. There are a total of 20 AAs, though between 1 and 5 are typically observed within a given site on the viral genome across a sample of individuals.

As described above, the viral genome changes over time and in response to treatment exposures. Thus, while viral RNA is single stranded, an individual can carry multiple genotypically distinct viruses, which we refer to as *strains*, resulting from multiple infections or quasi-species that developed over time within the host. Technically, a strain refers to a group of organisms with a common ancestor; however, here we use the term more loosely to refer to genetically distinct viral particles. As a result, multiple AAs can be present at a given site within a single individual. Typically, a frequency of at least 20% within a single host is necessary for standard population sequencing technology to recognize the presence of an allele. Thus, the number of AAs at a given location within an individual tends to range between one and three. In contrast, there are always exactly two alleles present at a given site within an individual for the human genetic setting, one inherited from each of the two parental genomes. Regions of the genome are segments of RNA that generally code for a protein of interest. For example, in the context of studying viral resistance, the *Protease* (Pr) region and *Reverse Transcriptase* (RT) regions are of interest since these code for enzymes that are targeted by ARTs. The Envelope region, on the other hand, may be relevant to studies of vaccine efficacy since it is involved in cell entry. Regions are tantamount to genes in the context of human genetic studies.

### 1.3.3 Publicly available data used throughout the text

#### *The FAMuSS study*

The Functional SNPs Associated with Muscle Size and Strength (FAMuSS) study was conducted to identify the genetic determinants of skeletal muscle size and strength before and after exercise training. A total of  $n = 1397$  college student volunteers participated in the study, and data on 225 SNPs across multiple genes were collected. The exercise training involved students training their non-dominant arms for 12 weeks. The primary aim of the study was to identify genes associated with muscle performance and specifically to understand associations among SNPs and normal variation in volumetric MRI (muscle, bone, subQ fat), muscle strength, response to training and clinical markers of metabolic syndrome. Primary findings are given in Thompson *et al.* (2004). A complete list of associated publications can be found in the ReadMe file on the textbook webpage.

The data are contained in a tab-delimited text file entitled **FMS\_data.txt** and illustrated, in part, in Table 1.1. The file contains information on genotype across all SNPs as well as an extensive list of clinical and demographic factors

Table 1.1. Sample of FAMuSS data

fms.id	actn3_r577x	actn3_rs540874	actn3_rs1815739	actn3_1671064	Term	Gender	Age	Race	NDRM.CH	DRM.CH
1	FA-1801	CC	GG	CC	02-1	Female	27	Caucasian	40.00	40.00
2	FA-1802	CT	GA	TC	02-1	Male	36	Caucasian	25.00	0.00
3	FA-1803	CT	GA	TC	02-1	Female	24	Caucasian	40.00	0.00
4	FA-1804	CT	GA	TC	02-1	Female	40	Caucasian	125.00	0.00
5	FA-1805	CC	GG	CC	02-1	Female	32	Caucasian	40.00	20.00
6	FA-1806	CT	GA	TC	02-1	Female	24	Hispanic	75.00	0.00
7	FA-1807	TT	AA	TT	02-1	Female	30	Caucasian	100.00	0.00
8	FA-1808	CT	GA	TC						
9	FA-1809	CT	GA	TC	02-1	Female	28	Caucasian	57.10	-14.30
10	FA-1810	CC	GG	CC	02-1	Male	27	Hispanic	33.30	0.00
11	FA-1811	CC	GG	CC						
12	FA-1812	CT	GA	TC	02-1	Female	30	Caucasian	20.00	0.00
13	FA-1813	CT	GA	TC	02-1	Female	20	Caucasian	25.00	25.00
14	FA-1814	CT	GA	TC	02-1	Female	23	African Am	100.00	25.00
15	FA-1815									
16	FA-1816	TT	GA	TC	02-1	Female	24	Caucasian	28.60	12.50
17	FA-1817	CT	GA	TC						
18	FA-1818	CT	GA	TC						
19	FA-1819	CT	GG	CC	02-3	Male	34	Caucasian	7.10	-7.10
20	FA-1820	CC	GA	TC	02-3	Female	31	Caucasian	75.00	20.00

for a subset ( $n = 1035$ ) of the study participants. We begin by specifying the web location of the data file as follows:

```
> fmsURL <- "http://people.umass.edu/foulkes/asg/data/FMS_data.txt"
```

We then use the `read.delim()` function to pull the data into R directly from the textbook website:

```
> fms <- read.delim(file=fmsURL, header=T, sep="\t")
```

By specifying `header=T`, we are indicating that the first row of the text file contains the variable names. Alternatively, we could have specified `header=F`, which assumes that the first line of the file is the first record of data. We also indicate with the argument `sep="\t"` that a tab separates each variable within a line of the data. Common alternative specifications are `sep=","` and `sep=" "`, indicating comma and space delimiters, respectively. As described in the appendix, other useful functions for reading data into R include `read.table()` and `read.csv()`. The specifications given above are the default values for `read.delim()` and need not be written out explicitly. We do so for the purpose of illustration.

A portion of the data on the first 20 individuals in this sample are displayed in Table 1.1. Included in this table are the genotypes for four SNPs within the `actn3` gene and a few corresponding clinical and demographic parameters. The variable `Term` indicates the year and term (1—spring, 2—summer, 3—fall) of recruitment into the study, and `Gender`, `Age` and `Race` are all self-declared values of these demographic factors. The percentage changes in muscle strength before and after exercise training are given by `NDRM.CH` for the non-dominant arm and `DRM.CH` for the dominant arm. Generation of the LaTeX code for Table 1.1 is done in R using the `xtable()` function in the `xtable` package. The `print()` function with the `floating.environment` option set equal to `'sidewaystable'` is used to generate a landscape table. Alternatively, we can print the table in R as shown below:

```
> attach(fms)
> data.frame(id, actn3_r577x, actn3_rs540874, actn3_rs1815739,
+           actn3_1671064, Term, Gender, Age, Race, NDRM.CH, DRM.CH)[1:20,]
```

We use the `attach()` function so that we can call each variable by its name without having to indicate the corresponding dataframe. For example, after submitting the command `attach(fms)`, we can call the variable `Gender` without reference to `fms`. Alternatively, we could write `fms$Gender`, which is valid whether or not the `attach()` function was used. A dataframe must be re-attached at the start of a new R session for the corresponding variable names to be recognized. The numbers `1:20` within the square brackets and before the comma are used to indicate that row numbers 1 through 20 are to be printed.

We see from this table that the genotype for `id=FA-1801` at the first recorded SNP (`r577x`) within the gene `actn3` is the pair of bases `CC`. In most

cases, SNPs are biallelic, which means that two bases are observed within a site across individuals. For example, for SNP **r577x** in gene **actn3**, the letters *C* and *T* are observed, while at **rs540874** in gene **actn3**, the two bases *G* and *A* are observed. This pairing is not restricted (that is, *A* can be present with *T*, *C* or *G* within another site), distinguishing this from the pairing of bases that occurs to form the DNA double helix within a single homolog (in which *A* always pairs with *T* and *C* with *G*).

Recall that an individual is said to be *homozygous* if the two observed base pairs are the same at a given site and *heterozygous* if they differ. From Table 1.1, for example, we see that individual FA-1801 from the FAMuSS study is homozygous at **actn3\_rs540874** with the observed genotype equal to *GG*. Likewise, individual FA-1807 is homozygous at this site since the observed genotype is *AA*. Individuals FA-1802, 1803 and 1804, on the other hand, are all heterozygous at **actn3\_rs540874** since their genotypes contain both the *G* and *A* alleles. Determination of a minor allele and its frequency is demonstrated in the following example using data from the FAMuSS study.

*Example 1.1 (Identifying the minor allele and its frequency).* Suppose we are interested in determining the minor allele for the SNP labeled **actn3\_rs540874** in the FAMuSS data. To do this, we need to calculate corresponding allele frequencies. First we determine the number of observations with each genotype for this SNP using the following code:

```
> attach(fms)
> GenoCount <- summary(actn3_rs540874)
> GenoCount
```

AA	GA	GG	NA's
226	595	395	181

The `table()` function in R outputs the counts of each level of the ordinal variable given as its argument. In this case, we see  $n = 226$  individuals have the *AA* genotype,  $n = 595$  individuals have the *GA* genotype and  $n = 395$  individuals have the *GG* genotype. An additional  $n = 181$  individuals are missing this genotype. For simplicity, we assume that this missingness is non-informative. That is, we make the strong assumption that our estimates of the allele frequencies would be the same had we observed the genotypes for these individuals. To calculate the allele frequencies, we begin by determining our reduced sample size (that is, the number of individuals with complete data):

```
> NumbObs <- sum(!is.na(actn3_rs540874))
```

The genotype frequencies for *AA*, *GA* and *GG* are then given respectively by

```
> GenoFreq <- as.vector(GenoCount/NumbObs)
> GenoFreq
```

[1]	0.1858553	0.4893092	0.3248355	0.1488487
-----	-----------	-----------	-----------	-----------

The frequencies of the  $A$  and  $G$  alleles are calculated as follows:

```
> FreqA <- (2*GenoFreq[1] + GenoFreq[2])/2
> FreqA

[1] 0.4305099

> FreqG <- (GenoFreq[2] + 2*GenoFreq[3])/2
> FreqG

[1] 0.5694901
```

Thus, we report  $A$  is the minor allele at this SNP locus, with a frequency of 0.43. In this case, an individual is said to be homozygous rare at SNP rs540874 if the observed genotype is  $AA$ . *Homozygous wildtype*, on the other hand, refers to the state of having two copies of the more common allele, or the genotype  $GG$  in this case.

Alternatively, we can achieve the same result using the `genotype()` and `summary()` functions within the `genetics` package. First we install and upload the R package as follows:

```
> install.packages("genetics")
> library(genetics)
```

We then create a `genotype` object and summarize the corresponding genotype and allele frequencies:

```
> Geno <- genotype(actn3_rs540874, sep="")
> summary(Geno)
```

Number of samples typed: 1216 (87%)

Allele Frequency: (2 alleles)

	Count	Proportion
G	1385	0.57
A	1047	0.43
NA	362	NA

Genotype Frequency:

	Count	Proportion
G/G	395	0.32
G/A	595	0.49
A/A	226	0.19
NA	181	NA

Heterozygosity (Hu) = 0.4905439

Poly. Inf. Content = 0.3701245

Here we again see that  $A$  corresponds to the minor allele at this SNP locus, with a frequency of 0.43, while  $G$  is the major allele, with a greater frequency of 0.57.  $\square$

*The Human Genome Diversity Project (HGDP)*

The Human Genome Diversity Project (HGDP) began in 1991 with the aim of documenting and characterizing the genetic variation in humans worldwide (Cann *et al.*, 2002). Genetic and demographic data are recorded on  $n = 1064$  individuals across 27 countries. In this text, we consider genotype information across four SNPs from the v-akt murine thymoma viral oncogene homolog 1 (AKT1) gene. In addition to genotype information, each individual's country of origin, gender and ethnicity are recorded. For complete information on this study, readers are referred to <http://www.stanford.edu/group/morrinst/hgdp.html>. Data are contained in the tab-delimited text file `HGDP_AKT1.txt` on the textbook website. Again we begin by specifying the location of the data:

```
> hgdpURL <- "http://people.umass.edu/foulkes/asg/data/HGDP_AKT1.txt"
```

Then we apply the `read.delim()` function to read the data into R:

```
> hgdp <- read.delim(file=hgdpURL, header=T, sep="\t")
```

Data on the first 20 observations in this dataset are provided in Table 1.2. Here the variable `Population` refers to ethnicity, `Geographic.origin` is the country of origin and `Geographic.area` is a more general description of location for the individuals in this cohort.

*The Virco data*

Several publicly available datasets that include viral sequence information, treatment histories and clinical measures of disease progression for HIV-infected individuals are downloadable at the Stanford Resistance Database: <http://hivdb.stanford.edu/>. In this text we consider a data set generated by Virco<sup>TM</sup>, which includes protease (Pr) sequence information on 1066 viral isolates and corresponding fold-resistance measures for each of eight Pr inhibitors. Fold resistance is a comparative measure of responsiveness to a drug, where the referent value is for a *wildtype* or *consensus* virus. The consensus AA at a site on the viral genome is defined as the AA that is most common at this site in the general population. The data are comma delimited and contained in the file `Virco_data.csv` on the textbook website. We use the `read.csv()` function in R to read in the data:

```
> vircoURL <- "http://people.umass.edu/foulkes/asg/data/Virco_data.csv"
> virco <- read.csv(file=vircoURL, header=T, sep=",")
```

Note that we now indicate `sep=","` since the data are comma delimited. This is the default for the `read.csv()` function. Complete information on the variables in the database and associated publications can be found on the Stanford Resistance Database website. A sample of the data on a select set

Table 1.2. Sample of HGDP data

Well ID	Gender	Population	Geographic.origin			Geographic.area			AKT1		
			Central	African	Republic	Central	Africa	CA	CT	TT	AA
1 B12	HGDP00980	F	Biaka Pygmies								
2 A12	HGDP01406	M	Bantu	Kenya			Central Africa	CA	CT	TT	AA
3 E5	HGDP01266	M	Mozabite	Algeria (Mzab)			Northern Africa	AA	TT	TT	AA
4 B9	HGDP01006	F	Karitiana	Brazil			South America	AA	TT	TT	AA
5 E1	HGDP01220	M	Daur	China			China	AA	TT	TT	AA
6 H2	HGDP01288	M	Han	China			China	AA	TT	TT	AA
7 G3	HGDP01246	M	Xibo	China			China	AA	TT	TT	AA
8 H10	HGDP00705	M	Colombian	Colombia			South America	AA	TT	TT	AA
9 H11	HGDP00706	F	Colombian	Colombia			South America	AA	TT	TT	AA
10 H12	HGDP00707	F	Colombian	Colombia			South America	AA	TT	TT	AA
11 A2	HGDP00708	F	Colombian	Colombia			South America	AA	TT	TT	AA
12 A3	HGDP00709	M	Colombian	Colombia			South America	AA	TT	TT	AA
13 A4	HGDP00710	M	Colombian	Colombia			South America	AA	TT	TT	AA
14 F5	HGDP00598	M	Druze	Israel (Carmel)			Israel	AA	TT	TT	AA
15 G11	HGDP00684	F	Palestinian	Israel (Central)			Israel	AA	TT	TT	AA
16 C2	HGDP00667	F	Sardinian	Italy			Southern Europe	AA	TT	TT	AA
17 E10	HGDP01155	M	North Italian	Italy (Bergamo)			Southern Europe	AA	TT	TT	AA
18 B7	HGDP01415	M	Bantu	Kenya			Central Africa	AA	TT	TT	AA
19 B8	HGDP01416	M	Bantu	Kenya			Central Africa	AA	TT	TT	AA
20 G4	HGDP00865	F	Maya	Mexico			Central America	AA	TT	TT	AA



of variables is given in Table 1.3. The variable **SeqID** is the sequence identifier, and **IsolateName** is the name given to the corresponding isolate. The drug-specific fold-resistance variables are labeled **Drug.Fold**, so, for example, Indinavir (IDV) fold resistance is given by the variable **IDV.Fold**. A higher fold-resistance value indicates that the corresponding isolate is more resistant (less sensitive) to the indicated drug than a wildtype sequence based on an *in vitro* assay.

The genotype information is available in two formats. The first representation is given by the variables with names that begin with the letter **P** and followed by a number. This number refers to the amino acid position within the Pr region of the viral sequence. For example, the variable **P10** represents the tenth AA position within the Pr region of the viral genome. A “—” in the data table indicates the presence of the population consensus AA, while a letter indicates a mutation in the form of the AA corresponding to this letter. For example, for **SeqID==3852**, a variant AA is observed at site 10 in the form of Isoleucine (*I*). A total of 99 **P** variables are included in this dataset, corresponding to the 99 AA sites in the protease region of the viral genome. An alternative formulation of the data is given by the variable **MutList**, which is a list of all the observed mutations. These data are coded by a letter, followed by a number, followed by another letter. The number is again the AA location, the first letter is the consensus AA at this site and the letter following the number is the AA(s) that are observed at the corresponding location. For example, *L10I* indicates that AA *I* is present in place of leucine (*L*) at site 10.

Table 1.3. Sample Virco data

SeqID	IsolateName	IDV	Fold	P10	P63	P71	P82	P90	CompMutList	
1	3852 CA3176	14.20	I	P	-	-	M	L10I, M46I, L63P, G73CS, V77I, L90M, I93L		
2	3865 CA3191	13.50	I	P	V	T	M	L10I, R41K, K45R, M46I, L63P, A71V, G73S, V77I, V82T, I85V, L90M, I93L		
3	7430 CA9998	16.70	I	P	V	V	A	M	L10I, I15V, K20M, E35D, M36I, I54V, R57K, I62V, L63P, A71V, G73S, V82A, L90M	
4	7459 Hertogs-Pt1	3.00	I	P	T	-	M	L10I, L19Q, E35D, G48V, L63P, H69Y, A71T, L90M, I93L		
5	7460 Hertogs-Pt2	7.00	-	-	-	A	-	K14R, I15V, V32I, M36I, M46I, V82A		
6	7461 Hertogs-Pt3	21.00	I	P	V	V	A	M	L10I, K20R, M36I, N37D, I54V, R57K, D60E, L63P, A71V, I72V, V82A, L90M, I93L	
7	7462 Hertogs-Pt4	8.00	-	P	-	A	-	M36I, G48V, I54V, D60E, I62V, L63P, V82A		
8	7463 Hertogs-Pt5	100.00	I	-	V	A	M	L10I, I13V, M36I, N37D, G48V, I54V, D60E, Q61E, I62V, I64V, A71V, V82A, L90M, I93L		
9	7464 Hertogs-Pt6	18.00	-	P	-	A	-	V32I, M46I, L63P, V82A, I93L		
10	7465 Hertogs-Pt7	15.00	-	I	V	A	M	E34K, R41K, K43R, I54V, I62V, L63I, A71V, T74S, V82A, L90M		
11	7466 Hertogs-Pt8	4.00	I	P	-	-	-	L10I, E35D, M36I, G48V, D60E, L63P, H69Y		
12	7467 Hertogs-Pt9	45.00	-	P	V	-	-	I13V, K14R, K20M, E35D, M36I, N37D, K45R, L63P, H69X, A71V, I84V, L89X		
13	15492 RC-V33778	1.00	X	-	V	-	-	L10X, I15V, I50V, I62V, A71V, I72V, N83Z		
14	15493 RC-V213888	1.00	F	A	-	-	-	L10F, I13V, L33F, M46X, I50V, L63A, T74S, V77I, L89M		
15	15494 RC-V207648	2.00	F	-	-	-	-	L10F, V32I, M46I, I47V, I62V		
16	15495 RC-V022292	3.00	-	P	V	A	M	E34Z, R41K, K43R, I54V, I62V, L63P, A71V, V82A, L90M, I93L		
17	15498 RC-V020855	1.00	I	X	X	-	-	L10I, G48V, I54X, L63X, I64V, A71X, I93L		
18	15499 RC-V216965	1.00	-	T	V	M	X	L33X, K43Z, M46V, I50V, Q58E, D60E, L63T, I64V, A71V, I72Z, V77I, V82M, L90X		
19	15500 RC-V020829	0.50	I	P	-	-	-	L10I, D30N, E35D, M36V, P39Z, L63P, N88D, I93L		
20	15501 RC-V020834	1.00	-	P	-	-	M	E35D, M36I, G48V, L63P, H69Z, L90M		

## Problems

- 1.1.** State the primary analytic considerations that distinguish population-based and family-based investigations.
- 1.2.** Define and contrast the following terms: (a) genotype, (b) haplotype, (c) phase, (d) homologous, (e) allele, and (f) zygosity.
- 1.3.** Based on the FAMuSS data, determine the minor allele and its frequency for the `actn3_1671064` SNP. Report these frequencies overall and stratified by the variable labeled `Race`. Interpret your findings.
- 1.4.** Using the HGDP data, summarize the genotype frequencies for the SNP labeled `AKT1.C6024T`, overall and by geographic area, using the variable named `geographic.area`. Interpret the results.
- 1.5.** Report the observed proportion of mutations at sites 1, 10, 30, 71, 82 and 90 in the Protease region of the HIV genome for the Virco data using the variables labeled `P1`, `P10`, `P30`, `P71`, `P82` and `P90`. Explain your findings.

Applied Statistical Genetics with R  
For Population-based Association Studies

Foulkes, A.S.

2009, XXIII, 252 p., Softcover

ISBN: 978-0-387-89553-6