

# Sociological Methods & Research

<http://smr.sagepub.com>

---

## How Is a Statistical Link Established Between a Human Outcome and a Genetic Variant?

Guang Guo and Daniel E. Adkins  
*Sociological Methods Research* 2008; 37; 201  
DOI: 10.1177/0049124108324526

The online version of this article can be found at:  
<http://smr.sagepub.com/cgi/content/abstract/37/2/201>

---

Published by:



<http://www.sagepublications.com>

Additional services and information for *Sociological Methods & Research* can be found at:

**Email Alerts:** <http://smr.sagepub.com/cgi/alerts>

**Subscriptions:** <http://smr.sagepub.com/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>

**Permissions:** <http://www.sagepub.com/journalsPermissions.nav>

**Citations** <http://smr.sagepub.com/cgi/content/refs/37/2/201>

# How Is a Statistical Link Established Between a Human Outcome and a Genetic Variant?

Guang Guo

*University of North Carolina at Chapel Hill*

Daniel E. Adkins

*University of North Carolina at Chapel Hill*

The objective of this article is to provide a nontechnical and intuitive introduction to the basic concepts and techniques that are used to establish statistical connections between genetic variants and human phenotypes. Depressive symptoms and delinquent behaviors that are of interest to sociologists are a subset of such human phenotypes. This article focuses on basic linkage analysis and association studies, the essential ideas behind the methods, and a very limited amount of molecular genetics needed for understanding the ideas. The article is written with those social scientists in mind who are interested in the topic but not yet ready to engage the vast and rapidly developing primary literature (journal articles).

**Keywords:** *society and genetics; behavior genetics; association studies; linkage analysis*

The objective of this article is to provide a nontechnical and intuitive introduction to the basic concepts and techniques that are used to establish statistical connections between genetic variants and human outcomes. Depressive symptoms and delinquent behaviors that are of interest to sociologists are examples of such human outcomes. This article does not address the enormously difficult and complicated issues of the biochemical processes through which variations in the deoxyribonucleic acid (DNA) sequence are translated into phenotypic variations. The article

---

**Author's Note:** Please address correspondence to Guang Guo, University of North Carolina, Department of Sociology, 155 Hamilton Hall CB #3210, Chapel Hill, NC 27599-3210; e-mail: [guang\\_guo@unc.edu](mailto:guang_guo@unc.edu).

is written with those social scientists in mind who are interested in the topic but not yet ready to engage the vast and rapidly developing primary literature.

The extant secondary literature consists of a large number of review articles, Web pages, and textbooks on molecular genetics, human genetics, linkage analysis, and association studies—the topics our article is devoted to. The wealth of information, however, may sometimes appear overwhelming to those who merely want a limited exposure in an efficient way. Our article focuses on the basic linkage analysis and association studies, the essential ideas behind the methods, and a very limited amount of molecular genetics needed for understanding the ideas. The introduction is inevitably incomplete and at times imprecise, but it is hoped that we will be able to provide a short piece of reading that maximizes the gain for the readers we have in mind.

Detecting the relationships between alleles and phenotypes has a great deal to do with the complexity of the phenotypes. Human phenotypes can be classified into one of two broad categories—Mendelian (or monogenic) and complex (or polygenic). In the Mendelian case, a single gene or allele may unequivocally determine the expression of a disease or phenotype. These phenotypes are likely to show clear patterns of inheritance in families and thus make gene-mapping efforts relatively straightforward; for example, see mapping studies on cystic fibrosis (Collins and Morton 1998) and neurofibromatosis (Barker et al. 1987). The spectacular advances in molecular genetics during the past two or three decades are largely confined to Mendelian cases.

Mendelian phenotypes, however, are rare (Lifton and Jeunemaitre 1993). The vast majority of phenotypes, such as cancer, blood pressure level, height, depression, and delinquency are subject to the influences of multiple genes, multiple environmental factors, and interactions among the two (Lander and Schork 1994). Genetic effects on complex traits are much weaker than those on Mendelian traits. Establishing a statistical link to a complex trait is much more difficult than establishing such a link to a Mendelian trait. Unraveling the genetic and environmental origins of these complex traits is now the focus of much molecular genetics research.

Genetic mapping of complex human outcomes is thorny. Even with experimental organisms (e.g., *drosophila*), it has been difficult to establish the links with the genes that are responsible for simple and highly heritable traits (Mackay 1996; Weiss and Terwilliger 2000) in spite of the many advantages that animal experiments have over human observational

studies: Inbred experimental organisms allow for much better control of both genetic and environmental factors.

Interestingly, given adequate training, quantitative social scientists may find a gene–environment analysis of complex human phenotypes rather similar, at one level, to what they spend their lifetime doing: deciphering how a large number of factors jointly shape a human outcome.

## Basic Genetics

### Human Genome

All organisms, including humans, have a genome containing all of the blueprints for the development and growth of an organism. The blueprints contained in a genome are coded in its DNA and are divided into discrete units called genes. Genes specify life primarily through directing the production of proteins. One of the most amazing things about life is that DNA from all organisms (e.g., humans, mice, worms, or trees) is made up of the same chemical and physical components. The total number of genes in the human genome is estimated at some 30,000, which is much lower than previous estimates of 80,000 to 140,000 and only one third greater than that of the simple roundworm *C. elegans*, which has about 20,000 genes (Claverie 2001).

Within virtually every human cell, there is a membrane-bound center or nucleus that contains the vast majority of our genetic information encoded on long molecules of DNA. DNA is composed of two linked strands. Each strand consists of two basic parts—(1) the four chemical bases that compose the genetic code: adenine and guanine, which are known as *purines*, and cytosine and thymine, which are known as *pyrimidines*, and (2) the sugar-phosphate “backbone” that structures the DNA molecule and attaches successive bases. The two strands are connected to each other by chemical pairing of each base on one strand to a specific partner on the other strand. Adenine (A) pairs with thymine (T), and guanine (G) pairs with cytosine (C). Thus, A · T and G · C base pairs are said to be complementary.

The structure of DNA (known as a double helix) resembles a slightly twisted ladder in which the two long sidepieces are sugar-phosphate backbones and the rungs are complementary base pairs (Figure 1). The complementary base pairing makes DNA an appropriate molecule for carrying genetic information: One strand of DNA can act as a template for the synthesis of a complementary strand. Thus, the DNA sequence is readily duplicated and perpetuated across generations of cells.

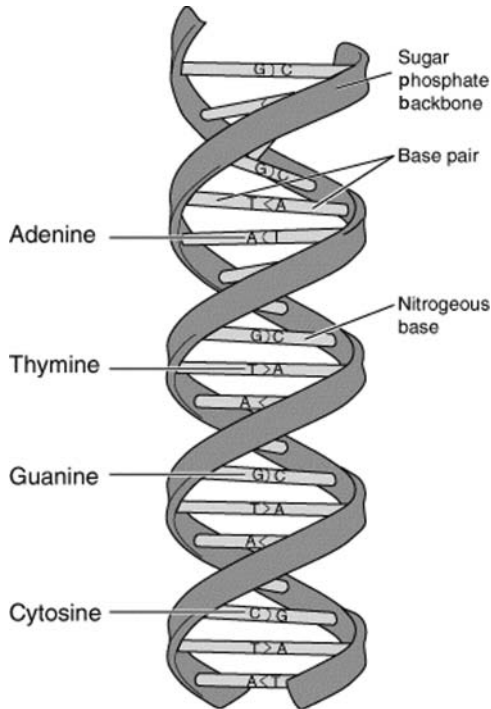
In most cells, DNA is packaged into structures known as chromosomes, located within the cell nucleus. In the form of chromosomes, DNA is elaborately coiled around proteins, such that the DNA molecule may be compacted to about 1/10,000 of its stretched-out length (Figure 2). The human genome consists of 23 chromosomes. Most human cells are *diploid*, meaning that they possess two copies of each of their 23 chromosomes, one set from each parent. Notable exceptions to this include gametes (sperm and egg), which have only a single set of chromosomes and are thus known as *haploid*. Among these chromosomes, 22 pairs are non-sex chromosomes, known as *autosomes*, with the remaining pair (i.e., sex chromosomes) specifying the sex of the individual.

The human genome consists of some 3 billion base pairs such as A, C, T, and G. These bases form DNA sequence that is arranged in a particular order along the DNA strand (e.g., ATTCGGCA). The order contains the exact codes for the blueprints mentioned earlier. Equating one DNA base to one English letter, the 3 billion bases would fill in 42 copies of *Webster's Third New International Dictionary*—a total of 116,000 pages (Mange and Mange 1999). Genes are segments of DNA in which a discrete, consecutive sequence of these bases encodes the directions for the synthesis of proteins. Gene sizes vary greatly, with an average of some 3,000 bases ranging from about 100 bases for tRNA<sup>lyr</sup> to more than 2 million bases for the dystrophin gene (Tennyson, Klamut, and Worton 1995). Genes occupy only about 2 percent of the human genome, with vast non-coding regions in between. The DNA in the noncoding regions provide chromosomal structural integrity and regulate protein production.

The coding sections of most vertebrate genes are split into exons, or expressed DNA sequences, and introns, or intervening sequences. DNA “instructions” produce proteins through two primary steps, transcription and translation. At Step 1, the instructions encoded by a gene are transcribed onto a smaller piece of genetic material known as messenger ribonucleic acid (mRNA) within the cell nucleus. The mRNA transcript is the exact complement of the entire gene, including both exon and intron sequences. After detaching from the DNA, the mRNA is often *spliced*. Part of the splicing process involves the removal of introns and the joining together of adjacent exons (Figure 3). Geneticists are now beginning to find that introns, once thought to be “junk” DNA, actually have important functions in regulating DNA activation (Zuckerkindl 2002).

The edited mRNA transcript then migrates out of the nucleus to the location of translation or protein production, the ribosome. Here, the mRNA is translated into protein in units of three consecutive RNA bases, called *codons*,

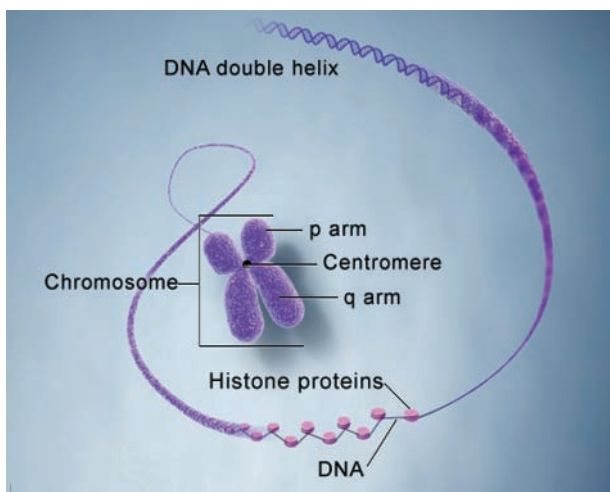
**Figure 1**  
**Double Helix Structure of DNA**



where each codon specifies a particular amino acid (Figure 4). In this way, ribosomes synthesize proteins one amino acid at a time, with the sequence of amino acids exactly determined by the sequence of the codons in the mRNA. Protein products are frequently subject to further modification after translation, before going on to serve their myriad of functions in maintaining life. It was formerly thought that each gene is coded for a single protein; however, it is known now that in some cases a single transcript can be edited in several ways, each specifying a different protein (Strachan and Read 1999).

About 99.9 percent of nucleotide bases are exactly the same in all individuals. However, 0.1 percent of 3 billion bases is still a very large number and leaves a lot of room for genetic differences across individuals. DNA variations occur at about 3 million bases in a human genome, and

**Figure 2**  
**DNA Packaged Into Chromosomes**



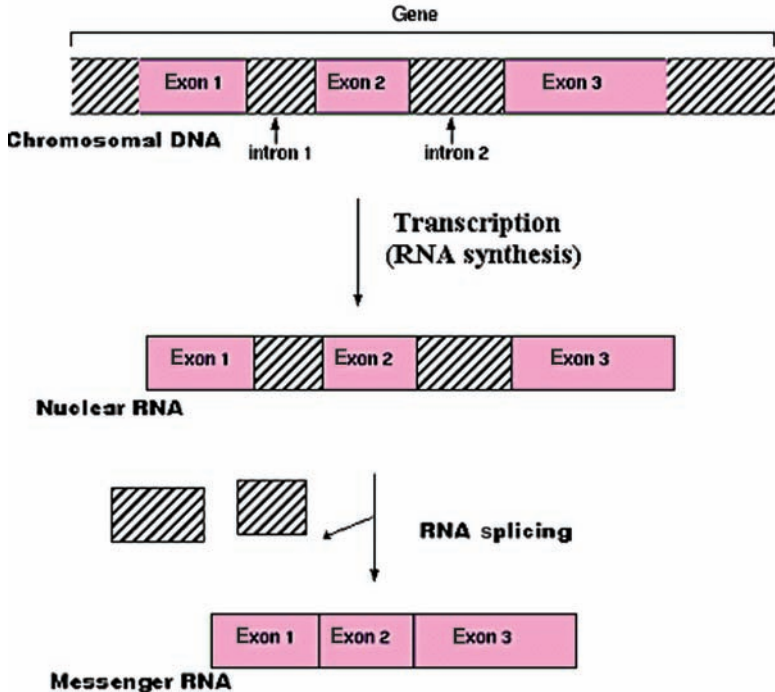
Source: U.S. National Library of Medicine.

these bases are known as single-nucleotide polymorphisms, or SNPs. These polymorphic sites influence gene expression and serve as markers for genetic research (Risch and Merikangas 1996).

### Sources of Human Genetic Variations

Human genetic diversity is driven by two processes: human reproduction and mutation. Through human reproduction we mix and recombine DNA in each generation, thus diversifying the genomes of our progeny. The genetic process driving this diversification is known as *meiosis*. Meiosis is the type of cell division that gives rise to sperm and egg cells (i.e., gametes), whereas mitosis is the normal type of cell division. In meiosis, DNA is diversified by two mechanisms: independent assortment and recombination. Independent assortment refers to the process through which the 23 chromosome pairs from the mother or the father are split up—with 23 chromosomes, 1 of each pair, going on to form a gamete. When the pairs separate during meiosis, for each of the 23 homologous pairs, the choice of which chromosome enters the gamete is independent.

**Figure 3**  
**Transcription of RNA From DNA and RNA Splicing**

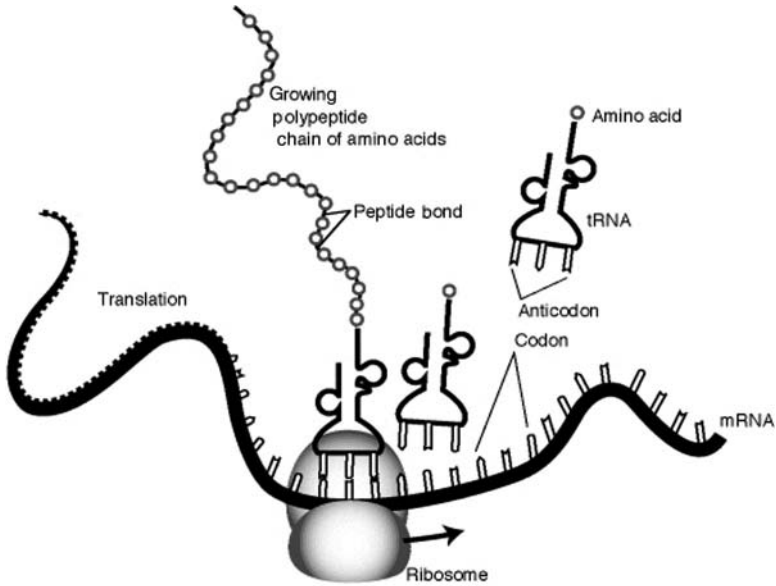


Thus, due to this independent assortment, each individual (mother or father) produces 1 of  $2^{23}$ , or about 8.4 million, possible different parental chromosomes. Therefore, no two eggs or sperms produced by one individual are likely to be identical.

Besides this independent assortment of chromosome pairs, recombination or crossovers result in a further round of genetic diversification. Recombination involves exchanges of chromosome parts between maternal and paternal homologous chromosomes. These homologues are two copies of a chromosome in a diploid cell; one was inherited from the father and the other from the mother (Figure 5).

When each pair of chromosomes exchanges DNA during recombination, the process takes place in segments that are often several thousands of base pairs long. The particular DNA sequences at two or more

**Figure 4**  
**Translation of mRNA Into Proteins**

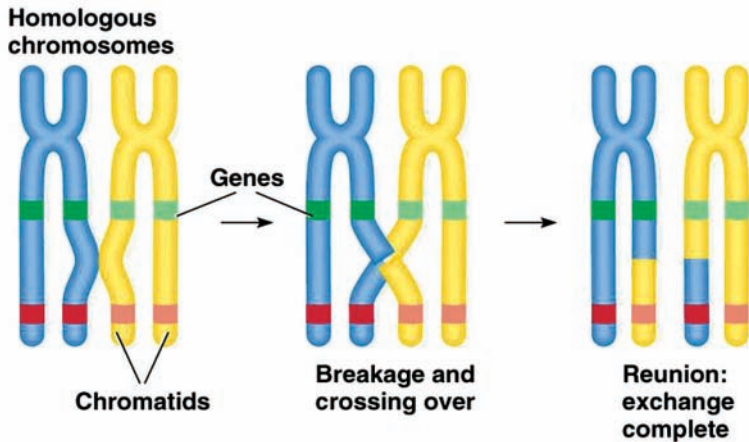


Note: mRNA = messenger RNA; tRNA = transfer RNA.

neighboring loci within the segment are thus spatially linked. The phenomenon of DNA being spatially dependent due to recombination is known as *linkage disequilibrium* and is the basis for contemporary molecular genetic analysis, which attempts to link molecular genetic variants to a human outcome. DNA markers or genetic polymorphisms may be used to detect the approximate location of a putative gene via the association between such markers and the traits of interest. Even when the marker is not within the gene of interest, the approximate location is suggested because the marker is spatially linked to the putative gene.

In addition to meiosis, which maintains genetic diversity by reshuffling existing genomes, mutation creates truly new genetic sequences. A mutation is a permanent change, a structural alteration, in the DNA. The causes of mutations are referred to as *mutagens* and include many factors such as radiation and mutagenic chemicals. Additionally, mutations are sometimes attributed to random chance events.

**Figure 5**  
**Chromosomal Crossovers Result in Genetic Recombination**



There are two places where mutations can be introduced and carried into the next generation: in the gametes and the germ line cells that produce them. Although the vast majority of mutations are deleterious, they are nonetheless one of the principle vehicles of evolution. The rare occurrence of beneficial mutations provides a species with the opportunity to adapt to new environments in many ways, such as protection from new pathogens. Thus, as new environments arise, individuals carrying certain mutations that enable an evolutionary advantage will survive to pass this mutation on to their offspring.

### SNPs and Other Polymorphisms

Most of our DNA sequence is identical to the DNA sequence of others. However, there are inherited regions of DNA that can vary from one individual to another. Variations in DNA sequence among individuals are termed *polymorphisms*. Sequences with a high degree of polymorphism are very useful for DNA analysis that attempts to link human outcomes to variations in DNA sequence.

For social scientists, a genetic polymorphism may be understood as a genetic variable, that is, a DNA measure that varies across individuals. More

technically, a genetic polymorphism refers to a locus of DNA sequence in the genome at which two or more variants exist in natural populations. Several classes of genetic polymorphisms have been observed in the human genome. VNTR, STR, and SNPs are classes of DNA polymorphisms. VNTR, an acronym for variable number of tandem repeats and also called minisatellite DNA, is a chromosomal locus at which a particular repetitive sequence is present in different numbers in different individuals of a population. STR, an acronym for short tandem repeats and also called microsatellite DNA, is similar to VNTR. A VNTR is characterized by repeating units between 10 and 200 bases, whereas STR has repeating units of 2 to 5 bases.

The most important class of polymorphisms is SNPs, which are naturally occurring variants affecting a single nucleotide. For example, two sequenced DNA sections from two individuals, CCAATTA and CCAACTA, contain a difference in a single nucleotide. In this case, the SNP has two alleles: C and T.

SNPs have a number of important advantages over other types of genetic polymorphisms in genetic studies of complex human outcomes. SNPs are the most numerous type of polymorphism in the human genome, occurring once every several hundred base pairs throughout the genome (Cargill et al. 1999). SNPs are also found throughout the human genome. They are found in exons, introns, promoters, and enhancers; they are also found in segments between genes. Of particular importance is the impact of a single base pair substitution on a human outcome. An SNP in a coding region could potentially affect a corresponding protein, an SNP in an intron could affect splicing, and an SNP in a promoter region may influence related gene expression (Schork, Fallin, and Lanchbury 2000).

All SNPs originally arise from mutational events. However, by the time an SNP is measured and analyzed, the mutational event is generally many generations past. The vast majority of SNPs are rare (International HapMap Consortium 2005). About 40 to 50 percent of SNPs are expected to have a minor allele frequency of less than 0.05. Currently, major efforts are being made in the use of SNPs to identify loci influencing human traits and diseases.

## Linkage Analysis

### Two-Point Linkage Analysis

Until fairly recently, linkage analysis was the most commonly used approach to identify genes responsible for observed traits, particularly

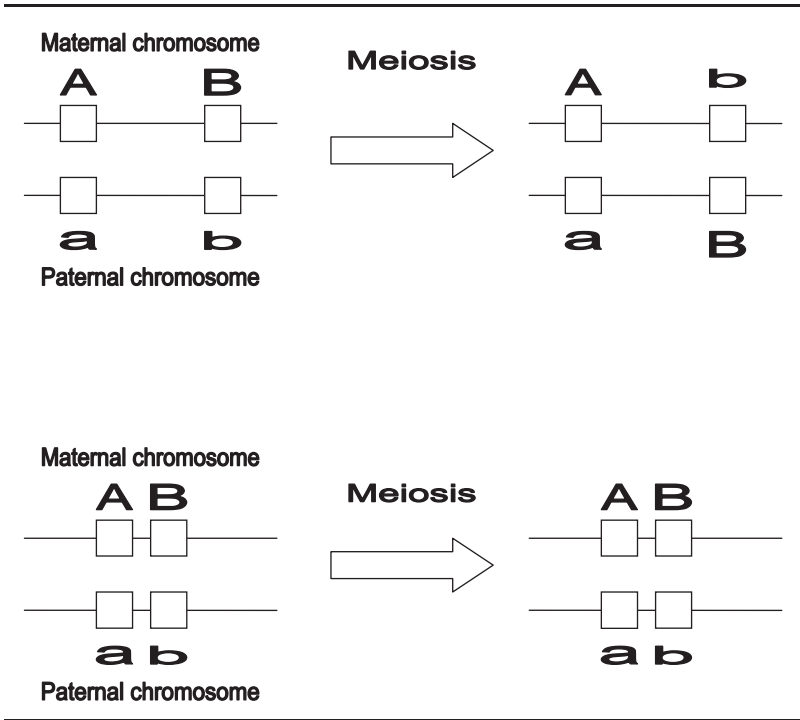
diseases. Although each particular linkage analysis method consists of a somewhat different set of technical details, the basic principle is similar. Genes that are located close to one another on the same chromosome are more likely to segregate, or to hang together, than genes far away on the same chromosome or genes on different chromosomes during meiosis or gamete formation. Thus, if a trait under consideration is found to segregate with, or to be passed on to the next generation together with, a genetic marker whose position is already known, then it can be inferred that the putative gene for the trait is located near the genetic marker.

Figures 6 and 7 illustrate the basic idea of linkage analysis intuitively. Figure 6 describes the observation that alleles that are close together on the same chromosome are more likely to cosegregate, or remain together, during meiosis than alleles that are further away on the same chromosome. Thus, when a marker allele (A) shows a strong tendency to be coupled with a disease, we would infer that the locus of a putative allele responsible for the dominant disease is in relatively close proximity to marker Allele A (Figure 7).

In Figure 7, Allele A is always coupled with the disease. This would happen if the locus of A is the very locus of the disease-causing gene (without considering other complicating factors such as penetrance and genetic heterogeneity). But this is extremely unlikely. We almost never know how far our genetic marker is from the putative gene. It would be lucky enough if the marker is close but at some distance from the locus of the putative gene for the disease on the same chromosome. Then recombination would be possible and Allele A would not always be observed in combination with the disease.

The objective of a linkage analysis is to find out how close the marker is to the putative gene. This can be accomplished by estimating the recombination fraction or the proportion of children in which there is a recombinant event between loci of interest, such as between a marker allele and a disease allele. This involves classifying the children into recombinants and nonrecombinants. In Figure 7, all the children in Generation III are nonrecombinants or the product of nonrecombinant gametes from their father. Nonrecombinants are individuals who receive from a parent a combination of alleles that has not been affected by one or more crossovers. Figure 8 is the same as Figure 7 except for the additional two recombinants: the fifth (BD) and sixth children (CD) in Generation III. These two recombinants are the result of the formation of new alleles in offspring not present in parents. Both of these children (BD and CD) are affected; however, in these two recombinant children, Allele A is not coupled with the disease, presumably

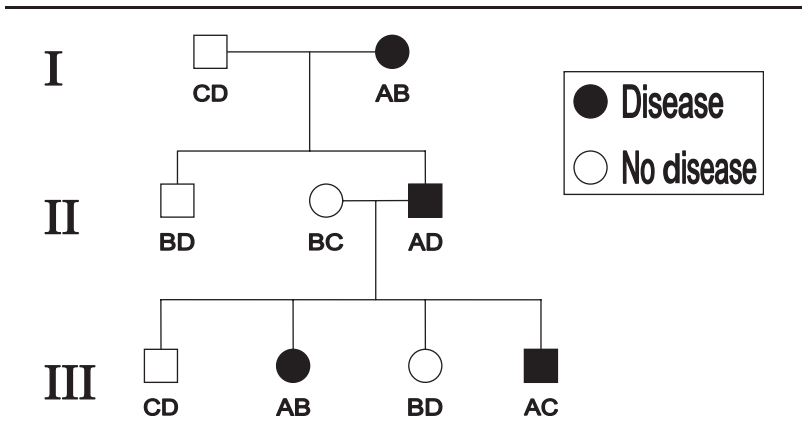
**Figure 6**  
**Proximate Alleles Tend to Cosegregate During Meiosis**



resulting from the processes of either crossing over or independent assortment. The presence of recombinants indicates that the disease gene is likely to be far away from the marker allele.

The recombination fraction  $\theta$  is .5 on average when two loci are on different chromosomes. Independent segregation ensures that the marker loci and the disease loci are coinherited 50 percent of the time on average. Alleles of two loci on opposite ends of the same chromosome behave in a similar manner as alleles on different chromosomes. Recombination fractions for two unlinked loci are .5, and they are never more than .5 (Strachan and Read 1999:270). The recombination fraction  $\theta$  is less than .5 when two loci are close together on the same chromosome. The closer two loci are on a chromosome, the less likely a crossover will separate them and the smaller the  $\theta$ .

**Figure 7**  
**Allele A Cosegregates With a Dominant Disease**



Genetic distance and physical distance between two loci are distinct concepts although the two are correlated. Genetic distance is defined by the expected number of crossovers per meiosis. One morgan (M) refers to the length of DNA on which one crossover per meiosis is expected on average. Physical distance is measured in terms of DNA base pairs.

The standard two-point linkage analysis uses the classic logarithm of the odds (LOD) score method developed by Morton (1955) about 50 years ago. Let  $\theta$  represent the probability of being a recombinant for a gamete and  $1 - \theta$  represent the probability of being a nonrecombinant. Then the likelihood function for  $\theta$  is

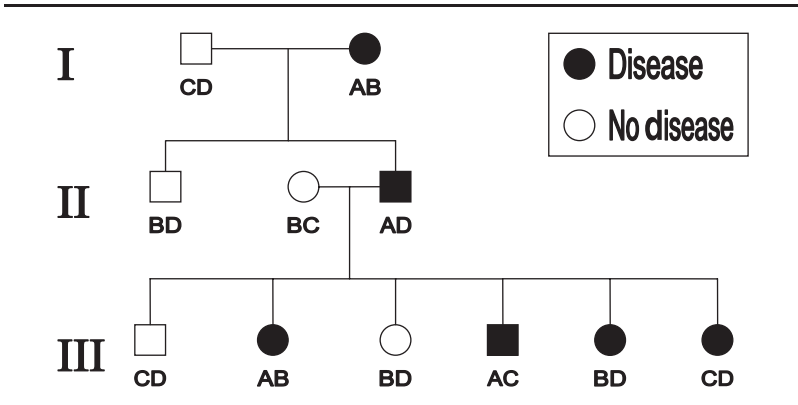
$$L(\theta) = (\theta)^R (1 - \theta)^{N - R},$$

where  $R$  is the number of recombinants and  $N$  is the total number of gametes including both recombinants and nonrecombinants. In linkage analysis, the LOD score is defined as the common (base 10) logarithm of the likelihood ratio of  $L(\theta)$  to  $L(\theta = .5)$ :

$$Z(\theta) = \log_{10}(L(\theta)/L(\theta = .5)).$$

The LOD score is the difference between  $\log_{10}L(\theta)$  and  $\log_{10}L(\theta = .5)$ , where  $\theta$  in  $\log_{10}L(\theta)$  is set at a value between 0 and .5 such as .1, .2, .3, .4, and .5 and  $\theta$  in  $\log_{10}L(\theta = .5)$  is fixed at .5. The LOD score is thus a function of  $\theta$  arbitrarily set between 0 and .5. The estimated  $\hat{\theta}$  is the value

**Figure 8**  
**Allele A Cosegregates in Some, but Not All, Cases With a Disease**



of  $\theta$  that maximizes the LOD score  $Z(\theta)$ . Positive LOD scores are considered evidence for linkage, and negative LOD scores are considered evidence for the absence of linkage.  $Z(\theta) = 3$  is considered strong evidence for linkage, indicating that the linkage hypothesis is 1,000 times as likely as the hypothesis that the two loci are not linked.

Table 1 shows the calculation of the LOD scores based on the hypothetical pedigree in Figure 8. The table includes the LOD scores for eight families. We assume that the first six of the eight families are exactly the same and the last two of the eight families are exactly the same. All meioses are phase known in the first six families: The evidence is available that the father in Generation II must have inherited A with the disease. Four of the six meioses in Generation III (Figure 8) in each of the first six families are nonrecombinants (CD, AB, BD, and AC), and two are recombinants (BD and CD). The LOD score for each of the six families is calculated accordingly.

The last two families in Table 1 are also based on the pedigree in Figure 8 except that we assume that the genotypes for Generation I are unknown. This would be the case when the two individuals in Generation I in Figure 8 were deceased and their genotype was unavailable at the time of the study. Then, the phase for Families 7 and 8 becomes unknown. The father in Generation II could have inherited either Allele A or D together with the disease. Thus, either CD, AB, BD, and AC are nonrecombinants when A is inherited with the disease, or CD, AB, BD, and AC are

**Table 1**  
**Logarithm of the Odds (LOD) Score Calculation**  
**for Eight Families Based on Figure 8**

Family	Recombination Fraction ( $\theta$ )				
	0.01	0.1	0.2	0.3	0.4
1	-2.211	-0.377	0.021	0.141	0.123
2	-2.211	-0.377	0.021	0.141	0.123
3	-2.211	-0.377	0.021	0.141	0.123
4	-2.211	-0.377	0.021	0.141	0.123
5	-2.211	-0.377	0.021	0.141	0.123
6	-2.211	-0.377	0.021	0.141	0.123
7	-2.512	-0.673	-0.254	-0.087	-0.018
8	-2.512	-0.673	-0.254	-0.087	-0.018
LOD score	-18.292	-3.606	-0.385	0.671	0.701

recombinants when D is inherited with the disease. The two cases are each weighted by one half in the calculation for the last two families. The total LOD scores are plotted against the recombination fraction in Figure 9, which does not indicate any evidence for linkage. An additional two dozen families similar to the first six would have produced such evidence.

So far, our discussion has dealt with the classical type of linkage analysis in which markers, one at a time, are tested for linkage to a trait caused by a single chromosomal location, known as *two-point mapping* of Mendelian traits. Currently, most linkage studies of Mendelian traits analyze the linkage between several markers and the trait of interest simultaneously. This approach, known as *multipoint mapping* of Mendelian traits, has several advantages over two-point mapping including allowing (1) a more definitive chromosomal ordering of the linked loci and (2) the extraction of increased information when markers are not fully informative.

Once linkage has been confirmed, the search for the susceptibility gene within the linked region can begin. In common disorders, the region of linkage is generally large, which means that the search could involve looking for one gene in a region of 20 to 30 million base pairs, which is likely to contain 500 to 1,000 genes.

### Complications of Classic Linkage Analysis

Linkage analysis has a number of limitations. First, incomplete penetrance is a major problem for linkage analysis. The penetrance of a trait for

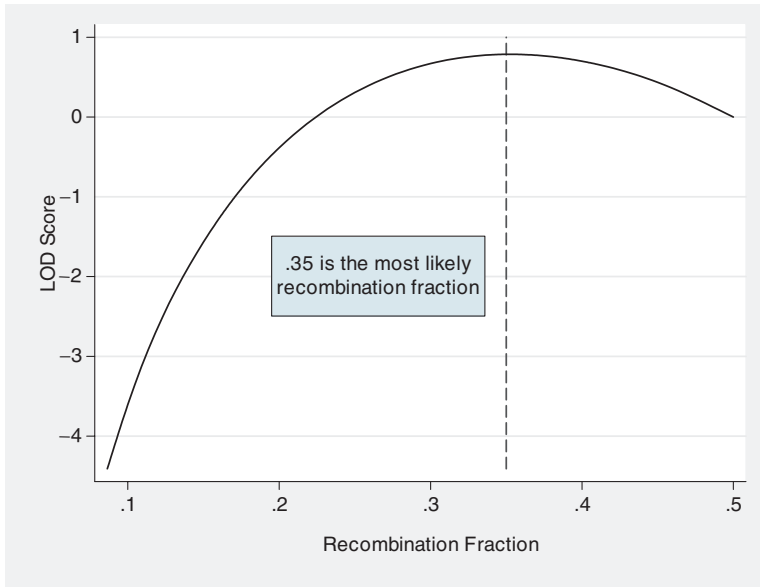
a given genotype is defined as the probability that a person with the genotype manifests the trait. In our numerical example (Figure 8 and Table 1), the penetrance is assumed to be 1 or 100 percent. Incomplete penetrance can be viewed as a form of variable expressivity. It may occur for many reasons, both genetic and environmental. Very often the presence and absence of a trait depends also on an unusual genetic background, a particular lifestyle, or just chance. The LOD-score linkage analysis requires specifying precisely the penetrance of each genotype. If no allowance is made for unaffected people who are nonpenetrant gene carriers, then these individuals may be incorrectly scored as recombinants. For complex diseases such as obesity and schizophrenia, the problems are much more intractable. For complex human outcomes, the penetrance is generally unknown.

*Phenocopies* are individuals who do not have a high-risk genotype, but who develop a disorder phenotypically indistinguishable from those phenotypes that are caused by the genotype. Breast cancer is an example. More than 90 percent of breast cancers are sporadic, that is, nonhereditary. These patients are phenocopies. If these breast cancer patients are included in the study for identifying *BRCA1* or *BRCA2*, then these individuals are likely to be wrongly scored as recombinants. The particular solution to phenocopies in the study on *BRCA1* was to calculate the LOD score only using early-onset breast cancer patients (Hall et al. 1990). In practice, the difficulties created by phenocopies are similar to those created by incomplete penetrance.

Given a sufficient number of meioses, the main obstacle in linkage analysis of Mendelian traits is *nonallelic heterogeneity*. Distinct DNA sequences at the same locus (allelic heterogeneity) or different loci (nonallelic heterogeneity) can be responsible for indistinguishable phenotypes. In linkage analysis, a marker close to a particular disease locus will demonstrate linkage in families where the disease is caused by alleles at that locus. In other families, where the disease is caused by alleles at other loci, the marker will show no linkage with the disease. Tuberous sclerosis is a case in point. Because mutations at two loci (*TSC1* at 9q34 and *TSC2* at 16p13) on Chromosomes 9 and 16 are responsible for the condition, the mapping took years of coordinated work in spite of the fact that the condition is clearly dominant and that large families are available for the studies (O'Callaghan 1999).

Breast cancer is also a relevant example. The two identified genes for breast cancer, *BRCA1* and *BRCA2*, are located on Chromosomes 13 and 17, respectively. A linkage analysis using breast cancer patients with *BRCA1* and patients with *BRCA2* may get significant signals from two loci

**Figure 9**  
**Plotting the Logarithm of the Odds (LOD) Scores Against**  
**the Recombination Fractions Based on the Calculation in Table 1**



in the genome if the sample is large and if phenocopies are not a problem. In most cases, signals from neither locus may be large enough and null results may be reported.

### Affected Sibling-Pair Linkage Analysis

In cases where the familial data do not satisfy the assumptions of classical, parametric linkage analysis, one long-standing nonparametric approach has been affected sibling (sib)-pair analysis. Because affected sib-pair analysis is model free, it can be performed without making any assumptions about the genetics of the disease. Thus, it has been used as one of the main tools for seeking genes conferring susceptibility to common non-Mendelian diseases such as diabetes or schizophrenia.

By definition, complex phenotypes are characterized by low penetrance, genetic heterogeneity, and lack of simple mode of inheritance (dominant or

recessive). The sib-pair method first developed in Penrose (1935) has been used to address these problems. The method relies on the concept that certain DNA segments between a pair of siblings are identical by descent (IBD) because the segments are inherited from the same parent. Figure 10 illustrates the distinction between being IBD and being identical by state. The A in II2 and the A in II3 are IBD since both As are unmistakably derived from their Father I1. However, III1 and III2 may not share one allele IBD even though the two siblings apparently share an A allele. It is unclear whether the A allele in III2 is from I1 or I4.

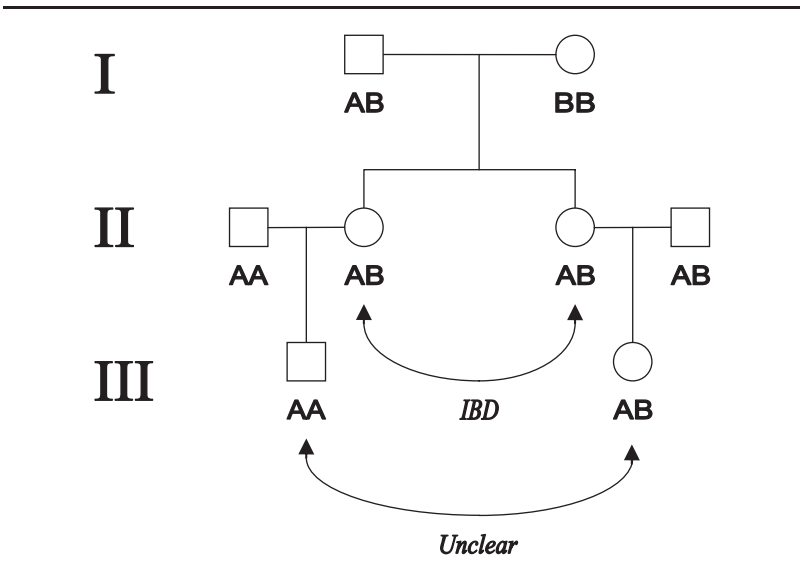
At a random DNA segment, pairs of full biological siblings are expected to share zero, one, and two parental alleles or marker alleles IBD with the probabilities of .25, .50, and .25, respectively. The sib-pair method uses full sibling pairs where both siblings are affected by a disease. The basic idea is that these affected siblings should share a higher proportion of marker allele IBD than by chance if the marker is linked to the disease locus. This should be true regardless of whether the disease is dominant or recessive. The sib-pair method has not proved particularly effective in mapping complex phenotypes. One drawback of the method is its weak power for complex phenotypes. In an influential article, Risch and Merikangas (1996) showed that the procedure would require an impractically large number of affected sibling pairs to locate a disease locus associated with a relative risk of less than 3.

## Association Studies

The majority of contemporary genetic studies that attempt to establish a statistical link between genetic variants and complex human outcomes are association studies. Social scientists routinely investigate whether Factor A is statistically associated with Factor B. The “association” in association studies literally means a statistical association as social scientists know it. In the simplest case where only one SNP is involved, the question is whether the categorical SNP variable is statistically correlated with a continuously or discretely measured human outcome.

Association studies are also referred to as linkage disequilibrium analysis. Linkage disequilibrium is simply association among tightly linked SNPs. Association studies and linkage studies share the same goal: to find the loci of the putative genes responsible for a human phenotype via one or more genetic markers whose loci are known. Both linkage and association

**Figure 10**  
**Distinction Between Identical by State and Identical by Descent (IBD)**



studies rely on similar principals: DNA in adjacent segments tends to be coinheritred by the next generation.

As we showed earlier, linkage analysis searches for DNA segments that are inherited together over the past few generations in family pedigrees of known ancestry. Naturally, the number of recombination events available for analysis in the past few generations is limited. A linkage analysis-identified genomic region is usually quite wide, containing several hundreds to several thousands.

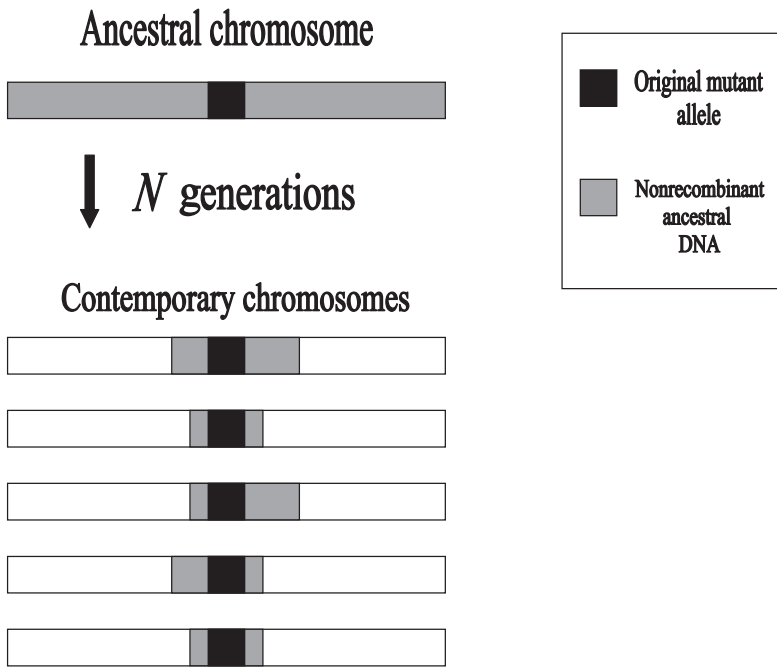
The standard analysis data for association studies consist of a sample of independent and unrelated individuals from a human population. Ostensibly, the data contrast from the pedigrees of linkage analysis. In fact, there is a great deal of similarity between the two types of data. The unrelated individuals in an association study may be viewed as coming from a gigantic pedigree that has accumulated over a large number of generations. Thus, at a fundamental level, associations studies are a sort of linkage analysis performed on one or more huge pedigrees.

Figure 11 illustrates the origin of linkage disequilibrium. In the figure, the original mutation is indicated in black and the rest of the ancestral chromosome is indicated in grey. The putative allele first appeared in a population through mutation or migration. Over generations, the linkage disequilibrium in the ancestral chromosome tends to decay; that is, the flanking stretches of ancestral haplotypes tend to become shorter and shorter. The rate of the decay in linkage disequilibrium is a function of generation  $t$  and recombination rate  $r$  between markers as described in the equation  $D = (1 - r)^t D_0$ . The contemporary individuals who carry the mutant alleles (Figure 11) will also carry segments of the ancestral chromosome (grey) that is adjacent to the mutant alleles. However, the DNA segments tend to be small. It is generally accepted that the segments do not extend beyond tens to hundreds of kilobases (kb; thousands of base pairs) as compared with the much wider segments of DNA (thousands of kb) considered in linkage analysis.

Association studies are often considered more attractive than linkage analysis for mapping complex phenotypes in recent years because of superior statistical power. Association studies are thought particularly promising for detecting associations between common genetic variants and common diseases (Cardon and Abecasis 2003; Risch and Merikangas 1996). More recently, whole genomewide association studies have increasingly become the focus of attention. Genomewide association studies aim at a near “thorough” survey of all known human genetic variants. Typically, these studies would survey 100,000 to 500,000 SNPs, as compared with traditional association studies that usually investigate a few to several dozens of SNPs in a single study. Naturally, given the recent development of genomewide studies, many issues are still unresolved concerning optimal research design and analysis strategies.

Genomewide studies are becoming feasible because of recent advances in high-throughput genotyping technology and the work by the International HapMap Project (International HapMap Consortium 2003, 2005). The HapMap Project has cataloged millions of SNPs and haplotypes across diverse populations (Africans from Nigeria, Chinese and Japanese from Asia, and European Americans from Utah). The high-density SNP genotyping across the genome provides information on SNP validation, frequency, and the correlation structure of alleles in the genome. The information can be used to identify subsets of highly informative “tag” SNPs. All genotyping data from the HapMap project are freely available on the Web, and it is hoped that this public resource will increase the power and efficiency of genetic association studies.

**Figure 11**  
**Origin of Linkage Disequilibrium**



### Sources of False-Positive (Negative) Findings in Association Studies

Association studies are based on the premise that a large sample of present-day unrelated individuals are in fact related a number of generations ago, thus providing sufficient recombination events (hence, more statistical power than linkage analysis) for analysis. However, many other events could have happened between the ancestral mutation and the present-day linkage disequilibrium. The level of linkage disequilibrium at the present day could be a function of genetic drift, population growth, population structure, migration, selection, mutation, and gene conversion. There is an enormous concern about false-positive or -negative findings from association studies in the genetics community. In this section, we describe two major sources of false-positive findings.

## Population Stratification

A statistical association between genetic polymorphisms and a human outcome can be produced by confounding effects, just as a statistical association in a social science study can be produced by omitted confounding factors. One such confounding factor is population stratification, which is present when individuals with the outcome and individuals without the outcome have different genetic backgrounds or have different ancestral population origins (Allison and Neale 2001; Cardon and Palmer 2003; Ewens and Spielman 1995). More specifically, population stratification exists when a population consists of multiple subgroups, each with a different allele frequency. Within each subgroup, the allele is independent of the outcome, but when the sample that consists of the multiple subgroups is analyzed without the knowledge of the subgroups, erroneous associations between the allele and the outcome may be concluded. The problem becomes more serious for large-scale association studies (Reich and Goldstein 2001).

A classic hypothetical example involves a false-positive result between a genetic variant and chopstick usage in San Francisco. Suppose that only European and Chinese Americans live in San Francisco, and Allele A in a particular gene among Chinese Americans has a much higher frequency than among European Americans. Suppose that all the Chinese use chopsticks and all Europeans do not. Then an association analysis that is unaware of the ethnic grouping in San Francisco would find a false relationship between Allele A in the gene and chopstick usage (Hamer and Sirota 2000). When analysis is done within the Chinese or European Americans, the relationship between the genetic variant and chopstick usage is nonexistent.

The simplest solution is to stratify an analysis by self-reported ethnicity. Tang et al. (2005) showed a near-perfect correspondence between the four self-reported ethnic categories (European Americans, African Americans, East Asians, and Hispanics) and the categories determined by 326 microsatellite markers. However, there may exist “cryptic” or unobserved population structures within each self-reported ethnicity (Pritchard, Stephens, Rosenberg, et al. 2000). In such a case, neither the number of groups nor the group membership of each individual is observed. When genotyping data for mother–father–child trios are available, unobserved population structures can be addressed by the transmission-disequilibrium test (TDT; Spielman, McGinnis, and Ewens 1993) for binary outcomes or QTDT (Allison 1997) for quantitative outcomes. A number of techniques have been developed as a general solution to unobserved population structures,

including genomic control (Devlin, Bacanu, and Roeder 2004; Devlin and Roeder 1999), structured association (Pritchard, Stephens, and Donnelly 2000; Pritchard, Stephens, Rosenberg, et al. 2000), and principal components analysis (Price et al. 2006). A large number of unlinked genetic markers in the range of 100 to 300 are required for any of the three general approaches.

## Multiple Testing

Multiple testing is another serious concern in genetic association studies. An association study investigates whether 1 or more genetic variants are associated with a human trait or disease. Such a study often involves tests of a large number of genetic variants, reaching 100,000 to 500,000 in a genomewide association study and creating a grave problem of multiple comparisons. As the number of statistical tests carried out increases, the number of false positives would increase proportionally. If 100 statistical tests are performed on the data set and the tests are independent, on average, about five false positives would appear even if the genetic variants and the trait are completely unrelated.

In the genetic literature, an adjustment is normally required to control for the increased error rate when multiple tests are conducted. The classic approach of the Bonferroni procedure overcorrects the problem under the assumption of independent tests because many of the genetic variants may not be independent. Several advances have been made in recent years. Some of the recent focuses (Benjamini et al. 2001; Hochberg 1988; van den Oord and Sullivan 2003) are on controlling the false discovery rate (FDR, defined as the percentage of statistical tests deemed significant that are false positives) rather than the traditional Type I error rate (the probability of making at least one false-positive inference). By controlling the FDR, we assure ourselves that on average, only about 5 percent of the total positive discoveries are false. This preserves greater power to detect true positives than the more traditional Bonferroni-type procedures. Another approach uses permutation testing (Edgington 1980; Nichols and Holmes 2002). This approach takes advantage of the correlation structure between the tests (testing SNPs in linkage disequilibrium will produce correlated tests of significance) in the multiple adjustment procedure. The permutation test computes significance by counting the number of ways the data can be permuted that produce results more extreme than observed (as in the Fisher exact test). Much less power is lost in this way of correcting for multiple testing.

## Conclusion

In this article, we aimed to provide a nontechnical and intuitive introduction to a number of statistical strategies that are used to link genetic variants and human outcomes. Interested readers should consult the primary literature (journal articles) for a more rigorous, detailed, and up-to-date treatment of the subjects, especially those who seriously contemplate performing genetic studies themselves.

It should be pointed out that significant statistical findings alone are rarely, if ever, considered proof of a link between a genetic variant and a human complex phenotype. This contrasts with the usual practice in social sciences. Repeated significant results in social sciences showing a connection between, for example, parental education and children's education attainment are often considered sufficient evidence for such a connection. The credibility of the evidence is not only from the replicated statistical results but also from real-life observation. Drawing from personal experiences, most people would probably agree that a higher level of parental education would lead to a higher level of education in children on average. Such confirmation from life experiences is not available for interpreting genetic findings. Genotypes are not visible in everyday life. To develop a credible story that supports statistical findings, other evidence is needed, such as those from animal studies and biochemical studies.

Although statistical evidence is rarely sufficient, it is usually necessary. The human phenotypes that interest social scientists are almost always complex. These phenotypes are subject to the influences of multiple genes, multiple environmental factors including social and cultural contexts, and the interaction between the two classes of influence. Social scientists have expertise in social and cultural contexts as well as in analyzing complicated human outcomes. This expertise will enable them to make important contributions required to understand how genetic variants are linked to some of the complex human outcomes.

## References

- Allison, D. B. 1997. "Transmission-Disequilibrium Tests for Quantitative Traits." *American Journal of Human Genetics* 60:676-90.
- Allison, D. B. and M. C. Neale. 2001. "Joint Tests of Linkage and Association for Quantitative Traits." *Theoretical Population Biology* 60:239-51.

- Barker, D., E. Wright, K. Nguyen, L. Cannon, P. Fain, D. Goldgar, et al. 1987. "Gene for von Recklinghausen Neurofibromatosis Is in the Pericentromeric Region of Chromosome 17." *Science* 236:1100-1102.
- Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi, and I. Golani. 2001. "Controlling the False Discovery Rate in Behavior Genetics Research." *Behavioural Brain Research* 125:279-84.
- Cardon, L. R. and G. R. Abecasis. 2003. "Using Haplotype Blocks to Map Human Complex Trait Loci." *Trends in Genetics* 19 (3): 135-40.
- Cardon, L. R. and L. J. Palmer. 2003. "Population Stratification and Spurious Allelic Association." *Lancet* 361:598-604.
- Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, et al. 1999. "Characterization of Single-Nucleotide Polymorphisms in Coding Regions of Human Genes." *Nature Genetics* 22:231-38.
- Claverie, J. M. 2001. "Gene Number: What If There Are Only 30,000 Human Genes?" *Science* 291:1255-57.
- Collins, A. and N. E. Morton. 1998. "Mapping a Disease Locus by Allelic Association." *Proceedings of the National Academy of Sciences of the United States of America* 95:1741-45.
- Devlin, B., S. A. Bacanu, and K. Roeder. 2004. "Genomic Control to the Extreme." *Nature Genetics* 36:1129-30.
- Devlin, B. and K. Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55:997-1004.
- Edgington, E. 1980. *Randomization Tests*. New York: Marcel Dekker.
- Ewens, W. J. and R. S. Spielman. 1995. "The Transmission Disequilibrium Test: History, Subdivision, and Admixture." *American Journal of Human Genetics* 57:455-64.
- Hall, J. M., M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey, et al. 1990. "Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21." *Science* 250:1684-89.
- Hamer, D. and L. Sirota. 2000. "Beware the Chopsticks Gene." *Molecular Psychiatry* 5:11-13.
- Hochberg, Y. 1988. "A Sharper Bonferroni Procedure for Multiple Tests of Significance." *Biometrika* 75:800-802.
- International HapMap Consortium. 2003. "The International HapMap Project." *Nature* 426:789-96.
- . 2005. "A Haplotype Map of the Human Genome." *Nature* 437:1299-320.
- Lander, E. S. and N. J. Schork. 1994. "Genetic Dissection of Complex Traits." *Science* 265:2037-48.
- Lifton, R. P. and X. Jeunemaitre. 1993. "Finding Genes That Cause Human Hypertension." *Journal of Hypertension* 11:231-36.
- Mackay, T. F. C. 1996. "The Nature of Quantitative Genetic Variation Revisited: Lessons from *Drosophila* bristles." *Bioessays* 18:113-21.
- Mange, E. J. and A. P. Mange. 1999. *Basic Human Genetics*. 2d ed. Sunderland, MA: Sinauer Associates.
- Morton, N. E. 1955. "Sequential Tests for the Detection of Linkage." *American Journal of Human Genetics* 7:277-318.
- Nichols, T. E. and A. P. Holmes. 2002. "Nonparametric Permutation Tests for Functional Neuroimaging: A Primer With Examples." *Human Brain Mapping* 15:1-25.
- O'Callaghan, F. J. K. 1999. "Tuberous Sclerosis: Epidemiological Research Is Needed to Complement New Findings in Genetics." *British Medical Journal* 318:1019-20.
- Penrose, L. S. 1935. "The Detection of Autosomal Linkage in Data Which Consists of Pairs of Brothers and Sisters of Unspecified Parentage." *Annals of Eugenics* 6:133-38.

- Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies." *Nature Genetics* 38:904-9.
- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155:945-59.
- Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly. 2000. "Association Mapping in Structured Populations." *American Journal of Human Genetics* 67:170-81.
- Reich, D. E. and D. B. Goldstein. 2001. "Detecting Association in a Case-Control Study While Correcting for Population Stratification." *Genetic Epidemiology* 20:4-16.
- Risch, N. and K. Merikangas. 1996. "The Future of Genetic Studies of Complex Human Diseases." *Science* 273:1516-17.
- Schork, N. J., D. Fallin, and J. S. Lanchbury. 2000. "Single Nucleotide Polymorphisms and the Future of Genetic Epidemiology." *Clinical Genetics* 58:250-64.
- Spielman, R. S., R. E. McGinnis, and W. J. Ewens. 1993. "Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM)." *American Journal of Human Genetics* 52:506-16.
- Strachan, T. and A. P. Read. 1999. *Human Molecular Genetics*. 2d ed. New York: John Wiley.
- Tang, H., T. Quertermous, B. Rodriguez, S. L. R. Kardia, X. Zhu, A. Brown, et al. 2005. "Genetic Structure, Self-Identified Race/Ethnicity, and Confounding in Case-Control Association Studies." *American Journal of Human Genetics* 76:268-75.
- Tennyson, C. N., H. J. Klamut, and R. G. Worton. 1995. "The Human Dystrophin Gene Requires 16 Hours to Be Transcribed and Is Cotranscriptionally Spliced." *Nature Genetics* 9:184-90.
- van den Oord, E. and P. F. Sullivan. 2003. "False Discoveries and Models for Gene Discovery." *Trends in Genetics* 19:537-42.
- Weiss, K. M. and J. D. Terwilliger. 2000. "How Many Diseases Does It Take to Map a Gene With SNPs?" *Nature Genetics* 26:151-57.
- Zuckerandl, E. 2002. "Why So Many Noncoding Nucleotides? The Eukaryote Genome as an Epigenetic Machine." *Genetica* 115:105-29.

**Guang Guo** is a professor of sociology and a faculty fellow at the Carolina Center for Genome Sciences at the University of North Carolina at Chapel Hill. He has made sustained and focused efforts to integrate social sciences and genetics in his career. He is a guest editor of a special issue for two major sociology journals (*Social Forces* and *Sociological Methods & Research*) on sociology and biology/genetics, and he served recently on the National Committee on Gene-Environment Interaction for Health Outcomes at the Institute of Medicine, National Academies 2005-2006. In addition to publications in sociological literature, he has published numerous articles in genetics journals.

**Daniel E. Adkins** is a postdoctoral researcher at the Virginia Commonwealth University Center for Biomarker Research and Personalized Medicine, while finalizing his PhD in sociology at the University of North Carolina at Chapel Hill. His research focuses on quantitative method development and application in biomarker/genetic research and the social sciences. He has forthcoming articles in *Social Forces*, *Sociological Theory*, and *Research in Social Stratification and Mobility*.