Formal Genetics of Humans: Multifactorial Inheritance and Common Diseases

8

Andrew G. Clark

Abstract The study of the genetics of complex traits is made complicated by the fact that the traits themselves are influenced by an interplay of many genes with many environmental factors. In this chapter the historical concepts of quantitative genetics, including additive variance and heritability, will be developed to underscore how important it is to understand that the root of the problem is to explain how genes contribute to the variance in a trait. With molecular genetic markers, such as SNPs, it is possible to test whether there are differences in the measured phenotype among the genotypes at the genetic marker, and this serves as a crude test of association. Many interesting challenges arise when such a test is expanded to 1 million markers spanning the entire chromosome, a design known as a genome-wide association study (GWAS). Complications due to population stratification, admixture, genotype x environment interaction, epistasis, and rare alleles are all considered. Methods that test association by use of excess of allele sharing in siblings (affected sib methods) or other relatives, or by excess cotransmission of alleles and a disease state (transmission disequilibrium test) have their own set of advantages and disadvantages. The chapter closes with some considerations of why the powerful methods presented here nevertheless leave much of the genetic variance in complex traits unexplained.

Content

A.G. Clark (✉)
Department of Molecular Biology and Genetics,
Cornell University, Ithaca, NY 14853, USA
e-mail: ac347@cornell.edu

M.R. Speicher et al. (eds.), *Vogel and Motulsky's Human Genetics: Problems and Approaches,*
DOI 10.1007/978-3-540-37654-5_8, © Springer-Verlag Berlin Heidelberg 2010

8

8.1 Genetic Analysis of Complex Traits

A primary goal of genetic analysis is to understand the causal relationships that connect the observed variation in phenotypes to the underlying genetic variation in the population. The simplest case was that observed by Gregor Mendel, where a single gene with two different co-dominant alleles presents a one-to-one correspondence between genotype and phenotype. In this situation, the ability to predict offspring ratios from any given parental phenotypes is very good. Most human traits do not follow these simple rules of transmission, but instead have a more complex association between genotype and phenotype. We become convinced that there is at least some genetic aspect to the transmission, because there is *familial resemblance*. These traits aggregate in families, but do not segregate like a single Mendelian gene. Such traits include stature and body proportions, facial features, skin color, and blood pressure. Many diseases may have a complex nexus of causes, but often the liability may differ between individuals and may be genetic in origin. In earlier years, a Mendelian framework was often superimposed naively on such data, with no testing of the formal requirements for simple modes of inheritance. We will show in this chapter that a fruitful way to approach the genetics of complex traits is to fit the data on individual genotypes and phenotypes to specific models that consider different ways in which the genetic variation may be causing the phenotypic variation. One outcome of this kind of model fitting is to map the genes responsible for the variation. But by the very nature of complex traits, there is also a role of environmental effects on the traits, and the observation that different genotypes respond differentially to environmental pressures means that the inferences about the genotype-phenotype association depends on the environmental context. Let us first consider some basic principles about variation at the phenotypic and genetic levels.

8.1.1 Variation in Phenotypic Traits

A fundamental idea to focus on in considering complex traits is that the primary feature that is of importance is among-individual *variation*. Nearly every trait shows some level of variation among individuals, from overall body size measurements, to the most minute features, such as fingerprints. Most biochemical traits also display variation, including the levels of many components of the blood (cholesterol, hemoglobin) and ranging up to the activities of metabolic enzymes in the liver. It is only because there is variation among individuals that there is an opportunity to identify underlying genes that themselves harbor genetic variation in the form of differences in DNA sequences. These gene variations in turn may mediate the phenotypic variation. Just because we can identify mouse mutants in orthologous genes that have profound effects on a particular phenotype, this does not guarantee that the population will harbor natural polymorphisms in that gene, which in turn will influence trait variability. Similarly, the genes that appear to be most responsible for variation in a trait may play a part in the mechanism for that aspect of biology that seems totally peripheral, or in many cases one has no clue why gene X influences trait Y. This level of decoupling of genetic and phenotypic variation may seem unnerving at first, but for many attributes of profound medical importance there is important phenotypic variability (such as susceptibility to atherosclerosis) and excellent understanding of relevant pathways, but relatively great uncertainty about the causes of variation.

8.1.2 Familial Resemblance and Heritability

Before we make the leap from phenotypic variation to seeking to find the genes responsible for that variation, there is one other attribute of the trait that is of vital

importance. The trait might actually not have any genetic variation responsible for the phenotypic variation, but may instead be driven entirely by environmental influences, such as diet or exercise levels. Fortunately there is a rich history of study of the problem of detecting a role of genes in complex traits simply by asking whether the degree of resemblance among relatives is elevated above what one would see by chance.

A fundamental idea in quantitative genetics is that variability in a trait can be partitioned into components that contribute to that variability. We seek to explain the variability in the phenotypic measures in terms of both genetic and environmental factors. Environment is considered as a sort of trash-bin term to encompass all nongenetic factors that influence the phenotypic value. The simplest statement of a model is that the phenotypic value of an individual is composed of the sum of the genotypic value plus the environmental value:

$$P = G + E$$

where P = phenotypic value, G = genotypic value, and E = environmental value.

The phenotypic values of all individuals in a population have a mean and a variance around this mean. The variance is distinguished from other measures of variability by one mathematical property: different variances can be added to give a total variance and, conversely, a total phenotypic variance V_P can be broken down into its components, such as the genotypic variance V_G and the environmental variance V_E:

$$V_P = V_G + V_E$$

The idea that the sum of normally distributed factors yields a normal distribution whose variance is the sum of the variances of the components is true in the limit with many factors, and is a central idea in statistics (indeed, it is called the Central Limit Theorem) (Fig. 8.1).

However, the addition rule for variances applies only if genotypic and environmental values are independent of each other, i.e., when they are not correlated. If there is a correlation between the two, the covariance of G and E must be added:

$$V_P = V_G + V_E + 2\,\mathrm{Cov}_{GE}$$

Let us take an example from the area of genetics that first introduced these concepts – agricultural studies. It is normal practice in dairy husbandry to feed cows according to their milk yield. Cows that produce more milk are given more food. Such correlations of genetic and environmental factors tend to inflate the variance. Whether human societies present environmental perks to individuals in a way that is correlated with genetic proclivity is open to discussion. In any event, any such correlation between genetic and environmental variation should be identified, as it can cause serious problems in the modeling if it is ignored.

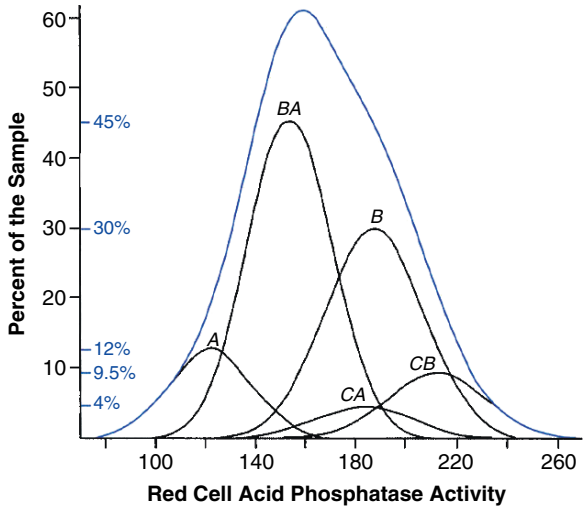

Fig. 8.1 The population distribution of acid phosphatase activity. The bell-shaped curve of total enzyme level (acid phosphatase) may be the sum of enzyme activities for acid phosphatase for genotypes of several polymorphic alleles, each having different overlapping acid phosphatase activity. Most phenotypes have a distribution in a population that results from summing over heterogeneous collections of genotypes (from Harris et al. 1968 Ann N Y Acad Sci. 151:232–242)

Another assumption is that specific differences in environments have the same effect on the various genotypes. When this is not so, there is an interaction between genotype and environment, giving an additional component to the variance V_{GE}. A prime example of genotype by environment interaction occurs with adverse reactions to drugs by a subset of individuals with a susceptible genotype. In the laboratory, where multiple replicate experiments may be run with identical genotypes of plants or animals, genotype × environment interaction is measured by testing the same genotypes across a range of environments.

The genotypic value V_G can be subdivided into several components: an additive component (V_A) and a component (V_D) measuring the deviation attributable to dominance and epistasis (V_I) from the expectation derived from the additive model. The dominance variance is contributed by heterozygotes (*Aa*) that are not exactly

intermediate in value between the corresponding homozygotes (*aa* and *AA*). The variance contributed by epistasis refers to the action of genes that affect the expression of other genes. Hence, the concept of additive variance does not imply the assumption of purely additive action of the genes involved. Even the action of genes showing dominance or epistasis tends to have an additive component. The whole genotypic variance can be written:

Phenotypic variance	Genetic variance	Environmental variance	Genetic × environmental covariance
$V_P =$	$V_A + V_D + V_I$	$+ V_E$	$+ \mathrm{Cov}_{GE}$

To estimate these various components of variance, one measures the phenotypes of individuals that have different known relationships to one another. There are simple algebraic relationships between the correlations of phenotypic measures among relatives and these components of variance. The one that we will focus on is the relationship between parents and offspring. Sir Francis Galton observed a nearly linear relationship between points that represent family groups plotted as follows. Define the *x*-axis as the average of the two parents' phenotypes (also called the midparent), and the *y*-axis as the average of the offspring phenotype. If each point represents a nuclear family, then in a population, such points will fall along a line whose slope is called the "heritability." If the slope is 1.0, this would mean that the average of the offspring is always equal to the average of the parents, and so the resemblance is perfect. More typically, the slope might be about 0.5, meaning that for every increase by a factor of 2 in the phenotype of the midparent, the offspring mean would increase by 1. Galton called this line through the scatter of points (Fig. 8.2) a "regression" line, because the offspring tend to be less deviant from the population mean than do the parents. As Galton put it, the offspring *regress* toward the population mean.

The heritability, as calculated by the midparent-offspring regression can also be written as:

$$h^2 = \frac{V_A}{V_P}$$

When it is written in this way, it is clear that heritability can be considered as the proportion of the total phenotypic variance that is explained by additive genetic effects. This expression varies between 0 and 1, and many morphological traits, such as height, have a heritability in the range of 0.7–0.8, implying that the bulk

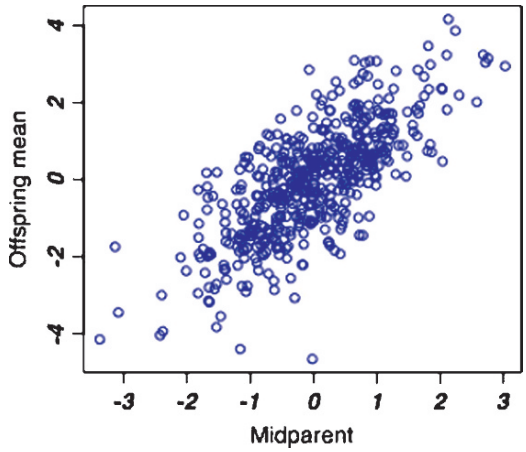

Fig. 8.2 The parent-offspring regression. The *x*-axis plots the average phenotypic measure of the two parents (the midparent) and the *y*-axis is the average phenotype of the offspring from each couple. Thus, each *point* on the plot represents a single nuclear family. The slope of the regression line through these points is the narrow-sense heritability. Francis Galton constructed many such scatter-plots, and inferred the degree of familial resemblance in this way

of the variance is genetic in origin. This very same term, often called the narrow sense heritability, has also been used by plant and animal breeders to predict the outcome of artificial selection for such economically useful traits as milk production in cows and egg laying in chickens. A high proportion of additive genetic variance implies that the trait will respond rapidly to selection. Heritability of many complex disorders, such as diabetes, is more in the range of 30–50%, implying that there clearly is a genetic component, but that this only explains part of the variation in disease risk.

It is worth stating carefully some of the properties of heritability:

(a) Heritability is a ratio. A ratio changes when either the numerator or the denominator changes. There is an increase in h^2 when the numerator (V_G, genotypic, or V_A, additive, variance) increases, or when the denominator (V_P, phenotypic variance) decreases. We could also say that a reduction in environmental variability will actually increase the heritability!

(b) The estimation of heritability is based on theoretical correlations between relatives. These correlations are valid only for random mating. Assortative mating leads to other correlations and, unless taken into consideration, produces systematic errors in the estimation of h^2. The correlations resulting from assortative mating were first calculated by Fisher [7]. These correlations can be used for adjustment of h^2.

(c) An estimation of h^2 is strictly valid only when the assumption is made that covariance and interaction between genotypic and environmental values are 0.

Correlations between relatives do not prove that there is genetic variability; they may also be caused by common environmental influences within families. In animal breeding, where the environment can be controlled, this factor might either be neglected or quantified. In humans, this is almost impossible. One of the major areas of research in human complex trait genetics today is to develop better automated methods for measuring differences in the environments that individuals have experienced. For example, many chemical exposures can be assessed by directly measuring residues in the bloodstream.

8.1.3 The Special Case of Twins

The use of twins has been much more popular in the past as a means for understanding genetic transmission of traits and diseases, but twin studies remain an excellent tool for developing concepts of genetic transmission. Identical or monozygotic (MZ) twins represent a wonderful experiment in nature (Fig. 8.3), since their genetic identity implies that differences between MZ

Fig. 8.3 Monozygotic twins have long fascinated human geneticists, and the questions about their biology change with advancing technologies. Initially interest focused on gross morphological similarities that were easy to measure. Now these questions center on issues of differential epigenetic modifications, differences in somatic mutations, altered patterns of X-inactivation, and similarities in brain activity as measured by functional MRI

twins must be due to accumulated perturbations of the environment [15]. In discussions on methods of quantitative genetics the use of twin data for quantitative assessment of the degree of genetic determination has been mentioned repeatedly. Indeed, twin investigations have played a major role in the history of human genetics. Especially in the field of behavior genetics, much of our current understanding is based on twin data. Therefore critical assessment of the twin method, its advantages, and limitations, is well motivated.

The twin method for assessing heritability is based on the biological observation that MZ twins originate from splitting of one zygote into two identical clones. It follows that any phenotypic differences between MZ twins must be largely caused by environmental influences. Somatic mutations may arise that generate differences between MZ twins, and efforts to quantify differences in somatic mutations between twin pairs using modern genomics technologies are under way in several laboratories. Environmental differences may manifest themselves by altering the epigenetic states of chromosomal regions, and an active area of research is to quantify the magnitude of differences in DNA methylation and histone acetylation between MZ twin pairs [2].

The degree of phenotypic similarity between MZ twins can be contrasted to the similarity between dizygotic (DZ) twins. Assuming that DZ twins are influenced by the same environmental differences but have only one-half of their genes in common by descent, the greater degree of resemblance of MZ twins provides a kind of measure of heritability. This heritability, however, is not the same as the parent-offspring regression approach mentioned above. Instead, the heritability one gets from twins is *broad-sense heritability*:

$$h_B^2 = \frac{V_G}{V_P}$$

where V_G and V_P refer to the total genotypic and phenotypic variance, respectively. This broad-sense heritability can be estimated from MZ and DZ twin pairs by calculating the average correlation between pairs of MZ twins (r_{MZ}) and the average correlation between pairs of DZ twins (r_{DZ}). The broad sense heritability is then $h_B^2 = 2(r_{MZ} - r_{DZ})$. It takes some algebra to show exactly why this is so, and it is of course true only when there is no shared environment effect. If there is a shared environment effect and it is measurable, one can adjust the heritability downward using another formula.

The above model for estimating heritability from twins makes some key assumptions about the biology that deserve to be considered carefully. In particular, twins have a unique shared environment that nontwins do not, and one has to worry whether that shared time in utero may influence their degree of resemblance. Because they have shared nutrition and environmental stresses, this shared environment might be expected to inflate the resemblance of twins. Whether the resemblance is augmented more in MZ twins than DZ twin depends on the details of how the environment is experienced (e.g., in one chorion or in two chorions).

One appreciates the effect of the uterine environment simply by examining medical attributes of twins and nontwins. Twins suffer from a higher frequency of abnormalities during pregnancy and at birth. Their lower birthweight can be attributed only partly to the shorter duration of gestation. The stillbirth rate and infant mortality in early life are considerably higher in multiple births than in single ones; in later years, twins run a higher risk than nontwins of becoming mentally retarded, which is presumably at least partly due to complications during pregnancy and at birth. Even the mean IQ of both MZ and DZ twins is slightly lower than that of control populations.

Some features of twins result in a higher chance that they *differ* in traits. X-inactivation in females occurs at the division of the zygote after X-inactivation (and is a fairly disruptive process). It therefore may happen that all cells in which a certain X-linked gene has been inactivated end up in one twin, while all the cells with active X chromosomes are found in the co-twin. This phenomenon leads to clinical expression of X-linked traits (such as Duchenne muscular dystrophy or color blindness) in only one member of a female twin pair that is heterozygous for the X-linked trait. A striking example is that two of the MZ Dionne quintuplets were color blind! The roles of X-inactivation and of intrauterine effects on epigenetic modifications are two of the many processes that occur during development and result in altered resemblance between twins. Thus, twins remain a fascination for geneticists, although simple calculation of heritability based on twin resemblance is clearly fraught with problems.

8.1.4 Embedding a Single Measured Gene Influencing a Continuous Trait

Consider a trait that has important medical consequences, where the trait has continuous phenotypic variation but we also know about an underlying mechanism for the trait, and we have managed to identify a gene whose variation influences the trait. As an illustration, consider the example of warfarin dose and the *VKORC1* polymorphism. Warfarin is an important anticoagulant drug that is used for heart disease patients and other circumstances where it is important to "thin" the blood to prevent thrombosis (clotting). The problem with warfarin has been that it has a narrow range of dose within which it is effective – too low a dose and it fails to delay clotting time, but at too high a dose it leads to hemorrhagic complications. For each patient there is a period where the optimal dose for that patient must be determined by approximate testing of coagulation status. To make matters worse, there is wide variability among individuals in the best therapeutic dose.

Rieder et al. [17] did a retrospective study on a large cohort of individuals who had been on warfarin therapy. These individuals had been through the battery of tests to determine their correct warfarin dose, and this was the phenotype being considered. The target of warfarin is the vitamin K epoxide reductase complex 1 (*VKORC1*), and a first guess might be that there could be polymorphism in this gene that requires different doses of warfarin for effectiveness. The study was a stunning success, finding mean differences in optimal warfarin dose across genotypes.

Other studies had associated warfarin dose with the cytochrome P450 2C9 (*CYP2C9*) gene, and this raises the question of whether there might be other genes elsewhere in the genome that also contribute to variation in optimal warfarin dose. Cooper et al. [3] did a scan of 181 European warfarin users and a replication sample of 374 individuals. They tested 550,000 SNPs and found that *VKORC1* had by far the strongest association ($P=6.2\times10^{-13}$) and that a SNP in *Cyp2C9* has moderate significance ($P<10^{-4}$). Because none of the other SNPs attained significance in this study, the conclusion was that common SNPs with large effects on optimal warfarin dose are unlikely to be discovered outside of *VKORC1* and *CYP2C9*. In Sect. 8.1.5 we will see how to partition the variance in a trait, and to determine what fraction of the total variance in a phenotype is attributable to one or two major genes.

8.1.5 A Model for Variance Partitioning

The preceding section showed how the continuously varying phenotype can be thought of as having multiple causal factors that determine the phenotype. If one is lucky and has a handle on one of those factors, it is possible to determine what fraction of the total variation is explained by that one factor. Let us consider a model that seeks to explain variation in continuous traits as the sum of the effects over many loci. We can further assume that, as in the above section, we have a handle on one of the loci. Among a collection of individuals whose genotype is *aa*, we can define the mean phenotype as $-a$. For the *Aa* heterozygotes, let the mean phenotype be d, and for the *AA* homozygote, let the mean be $+a$. If the frequencies of the *A* and *a* alleles are p and q, an the population is in Hardy–Weinberg equilibrium, then the mean phenotype for the whole population is:

$$p^2a + 2pqd + q^2(-a) = a(p-q) + 2pqd$$

If we were to plot the phenotypes for these three genotypic classes on the y-axis, and label the x-axis with the genotypes *aa*, *Aa*, and *AA* at coordinates 0, 1, and 2 (think of this as measuring 0, 1, and 2 copies of the *A* allele), then a regression through these points has many useful attributes. The increase in phenotype for each addition of an *A* allele is the "average effect of an allelic substitution" and has the value $\alpha = a + d(q\text{-}p)$. The y-axis values for the points on the regression line are $-2p\alpha$, $2pqd$, and $2q\alpha$. These are the "breeding values," a term from classic animal breeding analysis. They give the value of each genotype if the allelic substitutions were purely additive. But because there is dominance, we can calculate the deviation of each observed phenotype from this regression line fit (like a residual in a regression). These are the dominance deviations, and they are $-2p^2d$, $2pdq$ and $-2q^2d$, respectively.

From the breeding values and the dominance deviations we can calculate two important attributes of this trait. The additive genetic variance is the variance in breeding values. This is the sum of the squared deviations from the mean, weighted by the population frequencies or:

$$V_A = p^2(2q\alpha)^2 + 2pq[(q-p)\alpha]^2 + q^2(-2p\alpha)^2 = 2pq[a + d(q-p)]^2$$

8

Recall that one of the definitions of narrow-sense heritability is the additive genetic variance divided by the phenotypic variance. This formula for the additive variance makes it clear that the additive genetic variance depends on allele frequencies, and it drops to zero with either $p=0$ or $p=1$. The variance in dominance deviations is the dominance variance. This is:

$$V_D = p^2(2q^2d)^2 + 2pq(2pdq)^2 + q^2(2p^2d)^2 = (2pqd)^2$$

The above formulae show clearly how the variance components are impacted by allele frequencies, and how heritability itself also varies with allele frequencies. The important point to remember about these measures of quantitative genetics is that these are parameters of a model, and the numbers have meaning only so far as the model explains the data. In many circumstances in plant and animal breeding for agricultural purposes, we have excellent data demonstrating the utility of the models. Human quantitative genetics cannot assess whether the model fits nearly as thoroughly, both because the environment is less well controlled and because the only crosses that can be observed are those drawn from a large, essentially randomly mating population.

In the absence of epistasis or genotype×environment interaction, the total genetic variance is the simple sum of the additive variance and dominance variance: $V_G = V_A + V_D$. This splitting of a variance into two parts is called variance partitioning, and a key part of modern quantitative genetics is to partition variance into components that have biological meaning. For example, in the *VKORC1* example, the total genetic variance in optimal warfarin dose can be partitioned into a component of variance attributable to the *VKORC1* gene, and another component that accounts for the rest of the genome. For further development of the models for partitioning quantitative genetic variation, see [6].

8.1.6 Relating the Model to Data

When there is a measured genotype that it is suspected is involved in a trait, the above model suggests a straightforward way to test what the effects of that gene on the phenotype are. We have to emphasize that this test is valid under the assumption that the effects across genes are additive. If this assumption is not right, then the inferred effects of the measured gene will not be valid – and the estimates can either spuriously overestimate the effects or underestimate them. Thus, a lot hinges on the validity of the assumption of additive effects.

First we bin the individuals in the population into the three bins based on their genotypes at the measured locus. If we plot them as in Fig. 8.4, it can be seen that one way to test the null hypothesis of no effect of this gene would be to perform a linear regression and test whether the regression coefficient (the slope) differs from zero. A nonzero slope indicates that the gene has an additive effect on this trait. A really wonderful aspect of this approach to the problem is that the slope is proportional to the additive effect contributed by this locus. Similarly, if the data are plotted with two *x*-values, where genotypes *AA* and *aa* are on the left, and *Aa* on the right, then the regression through these points will have zero slope if the heterozygotes are intermediate between the two homozygous classes. This would be true if there were zero dominance. So the estimator for the dominance effect is simply the regression coefficient obtained from the data when arranged in this way.

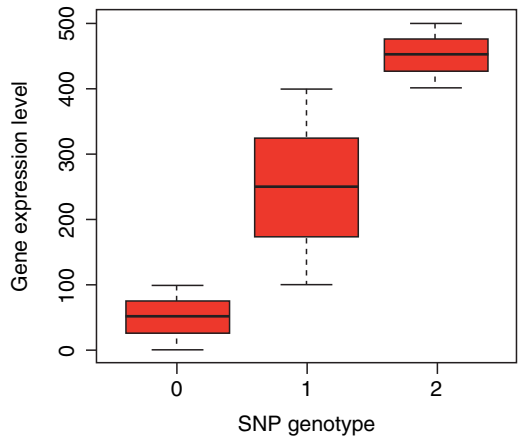

Fig. 8.4 Association of a single SNP with a continuous trait can be assessed by the regression plot depicted here: the *x*-axis has discrete elements for each genotype at one locus, and the *y*-axis is the continuously variable phenotypic measure. The slope of the regression line through these points yields the additive component of variance. If the genotype at this one SNP has no effect on the phenotype, this slope will be zero. The test of significance for whether the regression coefficient is greater than zero is the formal statistic test for whether this SNP shows an association with the phenotype

8.1.7 Mendelian Diseases Are Not Simple

While it is possible to trace the transmission of simple Mendelian traits and show that the trait is consistent with a major gene that is transmitted in the same way as smooth vs wrinkled peas, humans are so acutely aware of subtle phenotypic differences that the full spectrum of phenotypes associated with a major gene is almost never simple. And the departures from the pure Mendelian pattern are not always subtle. There are cases of individuals who are homozygous for a disease-causing allele, but are nevertheless perfectly healthy. Is this an example of a variant allele with reduced penetrance? Or, as is more often the case, are there other genes in the genome conferring a modifying influence, virtually suppressing the disease phenotype in these individuals? Mendelian disorders are not simply composed of two alleles, one healthy and one diseased; rather, a multitude of mutations that knock out function can result in disease, and there is also typically a series of alleles in the healthy group. In this case we have a mutation-selection balance between a fully functional gene and a rainbow series of alleles of reduced function. Heterozygotes may have even more intermediate phenotypes. It should be clear how this presents a situation where, despite the primary role of one major gene, there is nevertheless a continuous spectrum of disease severity in the population.

8.2 Genetic Polymorphism and Disease

Much of our understanding about the genetic basis of complex chronic diseases is based on our knowledge of Mendelian disorders, coupled with experiences in quantitative genetics of agricultural and laboratory organisms. We see that the complex disorders aggregate in families but do not segregate as Mendelian genes do, and so the inevitable conclusion is that the genetic basis involves many genes. In order to find those genes and to better understand the transmission of the disorder, we must construct a model for the genetic architecture. There may, for example, be a single major gene that accounts for most of the disease risk, but a series of modifier genes may temper the expression of this major gene. Or there may be ten genes, each of which is equally important in determining the trait. The frequency of the high-risk alleles may be very low, which may happen if there is natural selection driving them to low frequency in a mutation-selection balance, or they may have more intermediate frequency if they have little influence on reproductive fitness. In the next sections we will examine properties of polymorphisms in human genes and their impact on complex diseases.

8.2.1 Finding Genes Underlying a Complex Trait

In the preceding sections of this chapter we saw the consequences of a gene that has an effect on a trait and how its effect is added to the mix of effects of other genes and environment to add to the among-individual variability in the trait. Now imagine the situation in which you have no information about any of the underlying genes. The only data you have are the measurements of phenotypes of many individuals. You can determine that the trait has a heritable component because of the fact that relatives have correlated phenotypes. It is also clear that, if you did have measurements of the genotypes of a gene that happened to influence the trait (let the genotypes be *AA*, *Aa*, and *aa*), you might be able to see this from the fact that the phenotypes of these three genotypic groups might be different. The challenge is to identify an efficient way to find such genes.

8.2.2 Limitations of Pedigree Analysis

Probably the first approach one would consider for mapping the genes that underlie variation in a trait would be linkage analysis using pedigrees. This is a fundamental approach in human genetics, and it has a long history of success. As soon as one suspects a genetic basis for a syndrome, one has a collection of cases, and so by acquiring DNA samples from relatives, it becomes possible to test the linkage of the syndrome to anonymous marker loci throughout the genome. Methods for performing linkage analysis are provided elsewhere in this text, but there are several attributes of

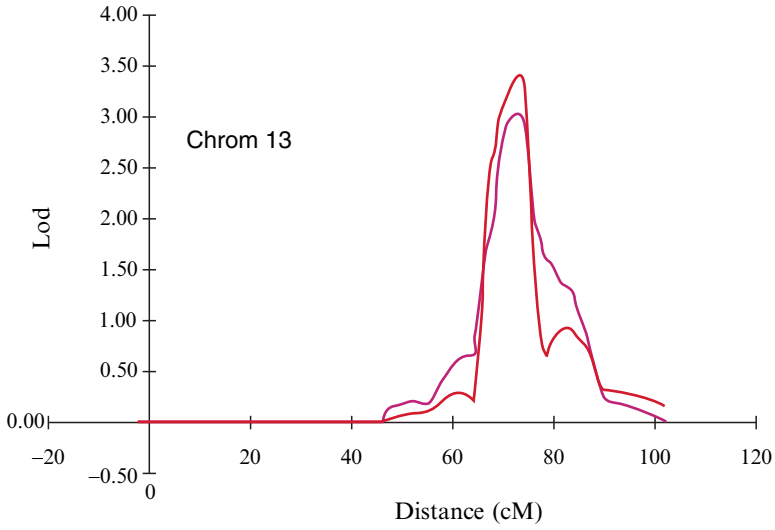

Fig. 8.5 Linkage analysis in pedigrees produces log-odds (LOD) plots like this one. In this example, there were 396 people in 22 families that were identified as having bipolar disorder in at least two members per family (these are called multiplex families). As described in the text, a model is fitted that provides the LOD score, representing the likelihood of obtaining such data given linkage at each position sliding along the chromosome. The salient feature to note is that the width of the LOD peak is nearly 20 cM or 20 Mbp across. This implies that there is relatively little confidence in the location, apart from there being a gene somewhere within that 20-Mbp region. (After [4])

linkage analysis that especially pertain to finding genes for complex traits that render linkage analysis somewhat less than ideal. First, if multiple genes are involved in a trait, the transmission pattern in a pedigree may be highly complex, and we may fail to detect the impact of any single marker through its marginal effect on risk. Even more serious is the fact that the resolution, in terms of accuracy of pinpointing the location of a gene on the genome, is limited by a combination of the sample size, the number of markers, and the number of meiotic exchanges represented by the pedigree. Typical pedigree studies have a mapping resolution of no better than 10 or 20 cM (centiMorgans), which is equivalent to approximately 10–20 Mbp of DNA sequence. This span typically encompasses dozens of human genes, and so one is left with a particularly challenging fine-mapping problem (Fig. 8.5).

8.2.3 A Prevailing Model: Common Disease Common Variants

For genetic association to be found by linkage disequilibrium, a fundamental constraint is that the rare allele must be relatively common (greater than about 10%) or the power to detect the association will be very low. Given that this approach can only find relatively common alleles, one can ask just how badly association mapping works for rare alleles. After all, common alleles, all else being equal, will contribute more to the total population variance in the trait, and will hence have a greater population attributable risk (defined elsewhere).What then are the prospects that common diseases will be caused by these relatively common alleles? Some Mendelian disorders can provide useful insight. If the disease is associated with a change in the environment, such as presence of malarial parasites, then alleles that may cause a disease (sickle cell anemia) may be driven to high frequency by the presence of a worse disease (malaria) against which they confer resistance. This kind of counterselection results in a heterozygote advantage, and any disease associated with alleles showing heterozygote advantage, either now or in the recent past, would be expected to have common alleles. The rapid expansion of the human population and the fact that many human populations have gone through population bottlenecks can also drive deleterious alleles to relatively high frequencies by drift and founder effects. In short, it was plausible that many diseases might have relatively common alleles as an underlying genetic cause. But these arguments do not make it particularly convincing that most complex diseases would be driven mostly by common alleles.

In the end it seems clear that successful identification of the common alleles causing disease would be the most desirable place to begin, since they likely harbor more of the population risk, and diagnostic tests that identify these tests are likely to identify more at-risk individuals than would tests for very rare alleles [1]. Now that more than 300 genome-wide association studies have been completed (http://www.genome.gov/gwastudies/), we can see that in no case was a very

large portion of the total variance explained by the associated SNPs. While there are many success stories of finding well-replicated associations between disorders and common SNPs, the effect sizes of those SNPs are all very small. That the common SNPs do not explain much of the variation in risk does not imply that the Common Disease Common Variant Hypothesis is totally in error, however, because it is possible that the variance explained is eroded by the fact that we are looking at effects of marker SNPs, and perhaps not the actual SNPs causing the variation in risk. But the fact that so little of the variance in risk is explained is unfortunate, and it suggests that myriad rare alleles of larger effect might contribute a substantial portion of disease risk in humans.

8.2.4 Affected Sib-pairs

For a brief period in the 1990s, the affected sib-pairs method was very popular, and it met with some success in mapping genes for some traits (more than 600 papers applying affected sib methods appear in PubMed; see [20] for a review of methods). The basic idea is that because full pedigrees are time consuming, expensive, and difficult to collect, one could collect the single kind of relative best matched for age and environment, namely siblings. The principle behind mapping with affected sib-pairs is to score genetic markers throughout the genome in a collection of sibs, and then to scan the genotype data to identify regions of the genome that show an excess of genetic identity between the sibling pairs.

To make sense of affected sib-pair methods we need the concept of *Identity By Descent*. Two alleles sampled from either two individuals or the same individual are said to be identical by descent if they can be traced back to a single ancestor. If two parents have genotypes A_1A_2 and A_3A_4, then a pair of siblings may both be A_1A_3, in which case they share two alleles that are IBD, or they may be A_1A_3 and A_1A_4, in which case they share one allele IBD. Finally, the two siblings may be A_1A_3 and A_2A_4, in which case they share zero alleles IBD. If you consider all possibilities, you find that ¼ of the time they share two alleles IBD, ½ the time they are expected to share one allele IBD, and ¼ of the time they are expected to share no alleles IBD. In table form it looks like this:

	Count of alleles IBD		
	0	1	2
Observed	n_0	n_1	n_2
Expected	$n/4$	$n/2$	$n/4$

where $n=n_0+n_1+n_2$ is the total count of sib-pairs in the study. The test of association is to perform a simple Chi-square test. If the null hypothesis is rejected, and if there is an excess count of those sharing one and two alleles, then this SNP shows a positive association with the disorder. It is not so easy to explain the case when the null hypothesis is rejected with an excess of cases sharing zero alleles. It does not imply that the SNP has protective effects. There are many extensions of this simple affected sib-pair test, including use of LOD scoring, application to continuously varying traits, and application to cases where other circumstances result in an empirical deviation from ¼ : ½ : ¼ for the expected allele sharing.

The basic idea of affected sib-pair mapping is to find regions of the genome where affected sibs have an elevated chance of sharing more alleles than this null model. The LOD score equivalent to the Chi-square can be plotted for each SNP as one scans along the chromosome, resulting in plots remarkably like the LOD score plots from full pedigree mapping efforts. Affected sib-pair methods retain the advantage in being much faster and easier to collect than full pedigrees.

8.2.5 Transmission Disequilibrium Test

The problem of hidden population stratification was seen as a serious limitation of direct association testing, because any such stratification could result in false-positive test results that would be difficult to identify without a full independent replication study. The Transmission Disequilibrium Test (TDT) is one of the simplest designs that is immune to the problem of population stratification. Since it was first introduced by Spielman et al. [21], there have been dozens of extensions to allow a similar test approach to apply to other scenarios. We will focus on just the simplest application, since it shows why the test works so well.

Suppose our sample consists of trios, each of parents and an affected offspring. The essence of the TDT is to ask whether the two alleles at a heterozygous SNP are transmitted at a 50:50 ratio to the affected offspring. If the SNP is linked to a mutant allele at a disease-

causing gene, then the transmission will be distorted. The test is essentially a Chi-square test for the co-transmission of the SNP and the disease state. If the count of trios where the *A* allele is transmitted is n_A, and the count of trios where the *a* allele is transmitted is n_a, then the Mendelian expectation is that each count would be $(n_A+n_a)/2$, so that the Chi-square is

$$X^2 = \frac{\left[n_A - \left(\frac{n_A+n_a}{2}\right)\right]^2 + \left[n_a - \left(\frac{n_A+n_a}{2}\right)\right]^2}{\left(\frac{n_A+n_a}{2}\right)} = \frac{(n_A-n_a)^2}{(n_A+n_a)}$$

This remarkably simple test has many positive attributes, not the least of which is the virtual immunity to distortions caused by population stratification. Its simplicity and robustness explain in part why it has been applied in nearly 1,200 published studies in human genetics.

8.2.6 Full-Genome Association Testing

In a major paradigm-shifting paper, Risch and Merikangas [18] pointed out the statistical limitations for mapping by determining linkage in pedigrees and carefully showed how we might be able to map in humans purely by association testing. This approach would work if there was relatively little linkage disequilibrium (LD) between SNPs or other genetic variants that are far apart along the chromosome. The hope was that a signature of high LD between a marker and a disease would indicate that the disease had to have risk factors mapping close to the SNP. This strongly motivated the quest for better understanding of LD across the human genome, and eventually led to completion of the human HapMap project [19]. The HapMap project provided us with a map of some 8 million markers and information on the pattern of LD across them in three human population samples. It also stimulated commercial entities to develop methods for genotyping those SNPs with high accuracy and low cost (see Sect. 8.3.2).

Risch and Merikangas [18] made the case for genome-wide association testing by showing that for a given sample size, one could have a greater probability of detecting association (higher power) by doing an association study than by doing a pedigree study. They considered a range of allele frequencies and genotypic relative risks for the disease-causing alleles, and several scenarios for the markers to be scored. It is impressive to see how accurately they foresaw the problems of testing 1,000,000 markers, estimating that a significance level of $a = 5 \times 10^{-8}$ would be needed to have a low probability of false positives. In Table 8.1,

Table 8.1 Sample sizes needed to detect a gene that elevates the risk of a complex disease under different assumptions of frequencies, genotypic relative risks, and testing approaches. (from Risch and Merikangas [18])

					Association			
Linkage					Singletons		SibPairs	
Genotypic risk ratio (γ)	Frequency of disease allele A (*p*)	Probability of allele sharing (*Y*)	No. of families required (*N*)	Probability of transmitting diseases allele A (*P*(tr-A)	Proportion of heterozgous parents (Het)	(*N*)	(Het)	(*N*)
4.0	0.01	0.52	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.5	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296.71	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0,667	0.5	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620.807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2216	0.253	941
	0.50	0.51	17,997	0.600	0.5	949	0.49	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

From [18], the paper that convinced the human genetics community that by scoring genotypes and phenotypes in direct association tests we ought to be able to identify genetic variants responsible for disease. The genotypic risk ratio (γ) is the ratio of risk of genotypes AA:aa

reproduced from their paper, you can see the massive reduction in sample size needed in an association study relative to a pedigree study for the same chance of finding a disease gene.

Note that association testing works by demonstrating a statistical correlation between allelic states of an anonymous marker and a putative risk-elevating locus. This approach is quite distinct from linkage-based mapping methods. The latter rely on identification of recombination events within the sample, and noting that two genes are closely linked if there are relatively few such recombination events. Because linkage methods rely on counting recombination events, the resolution comes from having a large number of such events. Even the largest pedigrees might have only a few thousand recombination events, and this limits the resolution and the statistical confidence in map distances obtained in linkage studies. Association studies seem to depend solely on the statistical correlation of allelic states, but behind this test is the idea that the correlations arise from a combination of low rates of recombination in the ancestral history of the variation and from random genetic drift. Genes that are far apart will have allelic states randomized relative to one another by recombination over a few generations. If the genes are close together, drift can generate LD, and recombination will be very slow to erode it, so at equilibrium there is a tendency for tightly linked genes to display LD.

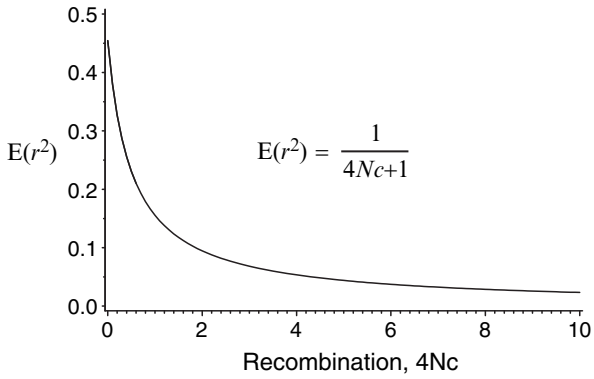

Fig. 8.6 Under the population genetic model, in which there is a balance between mutation, neutral drift, and recombination, there arises an equilibrium level of linkage disequilibrium (LD) as is plotted here. LD is here measured as the correlation coefficient r^2, as described in the text. The theory says that the expected value of r^2, or $E(r^2) = 1/(4N_e c + 1)$, where N_e is the effective population size, and c is the recombination rate. Note that the terms appear as the product $N_e c$, so that one expects the same LD if one doubles the population size and halves the recombination rate. The theory shows that there is a strong inverse relation between $4N_e c$ and LD

8.3 LD Mapping and Genome-Wide Association Studies

8.3.1 *Theory and How It Works: HapMap and Genome-Wide LD*

The basic principle behind LD mapping, also called association mapping, rests on a few key assumptions. Suppose a population is in a state of near equilibrium, with relatively little mixing through migration, so that the resulting genetic variation in the population is in Hardy–Weinberg proportions. In a population that has a steady rain of mutations, there will be a balance between the input of variation by mutation and its loss by random genetic drift. Some of the mutations have a deleterious effect; other mutations have no measurable effect; and very rarely some will be advantageous. Because there is recombination occurring in each generation, the statistical association between mutant alleles will tend to erode over time; however, the effect of random drift is to keep the LD from completely decaying to zero. Instead, there is a balance between mutation, drift, and recombination that produces a steady state level of LD. An approximate relation at steady state is $E(r^2) = 1/(1 + 4N_e c)$, indicating that the expected linkage disequilibrium as measured by r^2 is a simple function related inversely to a term with $4N_e c$, where N_e is the effective population size and c is the recombination rate [14, 22]. According to this theory, one would get the same LD if one halved the recombination rate and doubled the population size, so long as $4N_e c$ is kept the same (Fig. 8.6).

Empirically, the data on human LD support the idea of association mapping very well. In particular, one does find SNPs that are in strong pairwise LD, but basically this only happens if the SNPs are in close physical proximity along the genome (Fig. 8.7). When a pair of SNPs is farther apart than 100 kb or so, they only very rarely have strong LD. This means that a strong association between a disease and an SNP provides fairly convincing evidence that a gene associated with elevating disease risk must reside near the marker SNP.

8

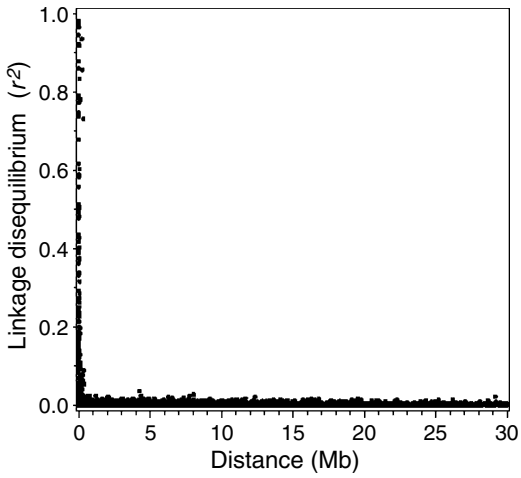

Fig. 8.7 A plot of the pairwise LD for a collection of SNPs from an early SNP study. In this study, the SNP genotypes were determined at several thousand SNPs in a few hundred people, and for each pair of SNPs it was possible to calculate the distance between them (for pairs on the same chromosome) and the level of LD between that SNP pair, showing clearly that SNPs that are far apart almost never have appreciable LD

According to this theory, if one imagines that there are SNPs responsible for disease, then there ought to be a statistical association between case/control status and the genotypes at nearby SNPs. Table 8.2 shows the steps for a genome-wide association test. The quality checking step is particularly vital, because despite the impressive gains in genotyping technologies, artifacts always creep into these studies and any slight perturbation from perfect genotyping calls can and usually does result in false-positive calls. Nearly every GWAS study had a moment of amazement when so many positive signals of association were seen, only for the number to dwindle as quality testing revealed more and more to be artifactual.

8.3.2 Technology: The Fantastic Drop in Genotyping Costs

One cannot overstate the importance of developments in the technology for large-scale molecular biology in accelerating the rate of discovery in human genetics. This is nowhere more true than in the area of genome-wide association testing. As recently as 2002, it cost about 1 U.S. dollar to score the genotype of an individual at one targeted nucleotide in the genome. Just 5 years later, one could score 1 million SNPs for $ 400, a 2,500-fold reduction in cost. This came about through development of mass manufacture of high-quality microarrays and methods to label and hybridize DNA to these arrays that gave highly accurate genotype calls. Competition among multiple manufacturers for competing technologies probably helped to drive the costs down as they drove speed and accuracy up. The next frontier is whole-genome sequencing at costs comparable to those of a CAT scan, and the human genetics community seems to have a consensus that this will happen within the next few years. Returning to the problem of mapping genetic variants that are associated with risk of complex diseases, even if we had complete DNA sequences of all the individuals in the case-control GWAS studies, many of the barriers to identification of genes responsible for inflated risk would still be there.

Table 8.2 Steps for a genome-wide association study

1. Identify the sample. Should be from a homogeneous population. Clearly defined cases and controls matched for gender and age.
2. Score the genotypes. Today this is almost universally done by applying standard commercial SNP genotyping chips from Affymetrix or Illumina.
3. Quality checking. It is necessary to take the genotype calls through rigorous testing for Hardy–Weinberg departures, spurious heterogeneity across runs, clustering of artifacts with cases, etc. Generally poor-quality DNA means removing some individuals, and some SNPs need to be removed.
4. Perform first-pass statistical inference**.** Nearly everyone starts with single-SNP tests, such as the Cochran–Armitage trends test.
5. Double-check all positives. The vast and overwhelming majority of positive hits seen at the first pass are errors of some sort. Disbelieve them until you fail to prove that they are errors.
6. Perform validation study. Standard practice is to repeat the study in another population to see that the same result is repeated.
7. Perform additional statistical inference. One can check for genotype x environment and epistatic effects, although the power will be low.

8.3.3 Case-Control Studies

Despite the fact that complex disorders are intrinsically embedded in likely interactions with environmental factors, the easiest design to begin genome-wide studies that identify genes associated with the disease is the case-control design. Because these tests entail examination of so many SNPs (typically 500,000 or 1 million SNPs), it is necessary to have large sample sizes so that the *P*-values of tests are sufficiently small, even when effect sizes are moderate, for the statistical tests to retain significance in the face of so many simultaneous tests. For example, the Wellcome Trust Case Control Consortium examined 2,000 cases for each of seven different disorders, and these were each contrasted against 3,000 controls [25]. With a complex disorder it becomes necessary to dichotomize individuals into these two bins, and it is crucial that this be done rigorously and homogeneously across the study., Other variables, such as sex, age, diet, etc. must either be randomized, controlled (e.g., by examining one sex only), or done as matched cases and controls, where the matching is for as many of these ancillary variables as possible. But case-control studies have a solid place in the history of medical research, and the simplicity of their design and ready access to samples stratified in this simple way means they are likely to continue to be useful. In addition, the first-pass statistical tests are very simple indeed.

8.3.4 Statistical Inference with Genome-Wide Studies

If the individuals in the study are placed into discrete bins of ;cases' and "controls," then the simplest way to consider the data is as a 3×2 table:

	AA	*Aa*	*aa*
Cases	n_{11}	n_{12}	n_{13}
Controls	n_{21}	n_{21}	n_{22}

It is legitimate to perform a 3×2 contingency Chi-square test on these data, provided the cell counts are sufficiently large (above 5 or so). For many SNPs one finds that the rare homozygous class has only a few observations, and in these cases one has to be careful about the aberrant behavior of the test statistic with small cell counts. One common way to solve the problem of small cell counts is to perform a permutation test to estimate the probability of a more extreme table. Another approach is to pool cells (e.g., the rarest genotype class, or column, could be pooled with the heterozygotes, yielding a 2×2 table).

Because the three genotypes are not totally independent categories, but rather there is an underlying order to them, a test more appropriate than the 3×2 contingency Chi-square is the Cochran–Armitage trend test. This test assumes that there is a linear trend in the phenotypes as one progresses from *AA* to *Aa* and *aa*, and obtains an asymptotically Chi-square test statistic under this model. Its primary advantage is in statistical power, because it effectively saves a degree of freedom. Just as for the contingency Chi-square, the significance test for the Cochran–Armitage trend test can be based on a permutation, and this allows it to be used even when cell counts are small. One needs to have *P*-values below 10^{-6} to attain significance across the whole study, and the Wellcome Trust Case Control Consortium was successful in achieving this for more than 80 SNPs across the seven disorders they mapped by GWAS (Fig. 8.8).

The genome-wide SNP chips are not successful at producing a reliable genotype call for every SNP in every individual, and the resulting missing data can be a challenge for analysis. One of the interesting features of dense SNP data is that because nearby SNPs are in LD, when one SNP call is missing, there is often some ability to predict the value of the missing genotype by use of the flanking SNPs. This "guessing the missing data" is known in statistics as *imputation* [8, 11]. While it sounds suspicious to fill in the missing data in this way, it is easy enough to test how well it works – simply take a large data set, blind yourself to some of the known genotype calls, and determine whether the imputation procedure gets the correct genotype call. When this is done, the misclassification error rate can be as low as 1%. With genome-wide SNP chips, whose density is one SNP every 3 kb on average, the imputation error rate varies with population but is typically less than 3%. Depending on the analysis, this can make a big difference. For the Wellcome Trust case-control study, use of imputed genotype calls often produced SNPs whose association *P*-values were more significant than the nonimputed SNPs.

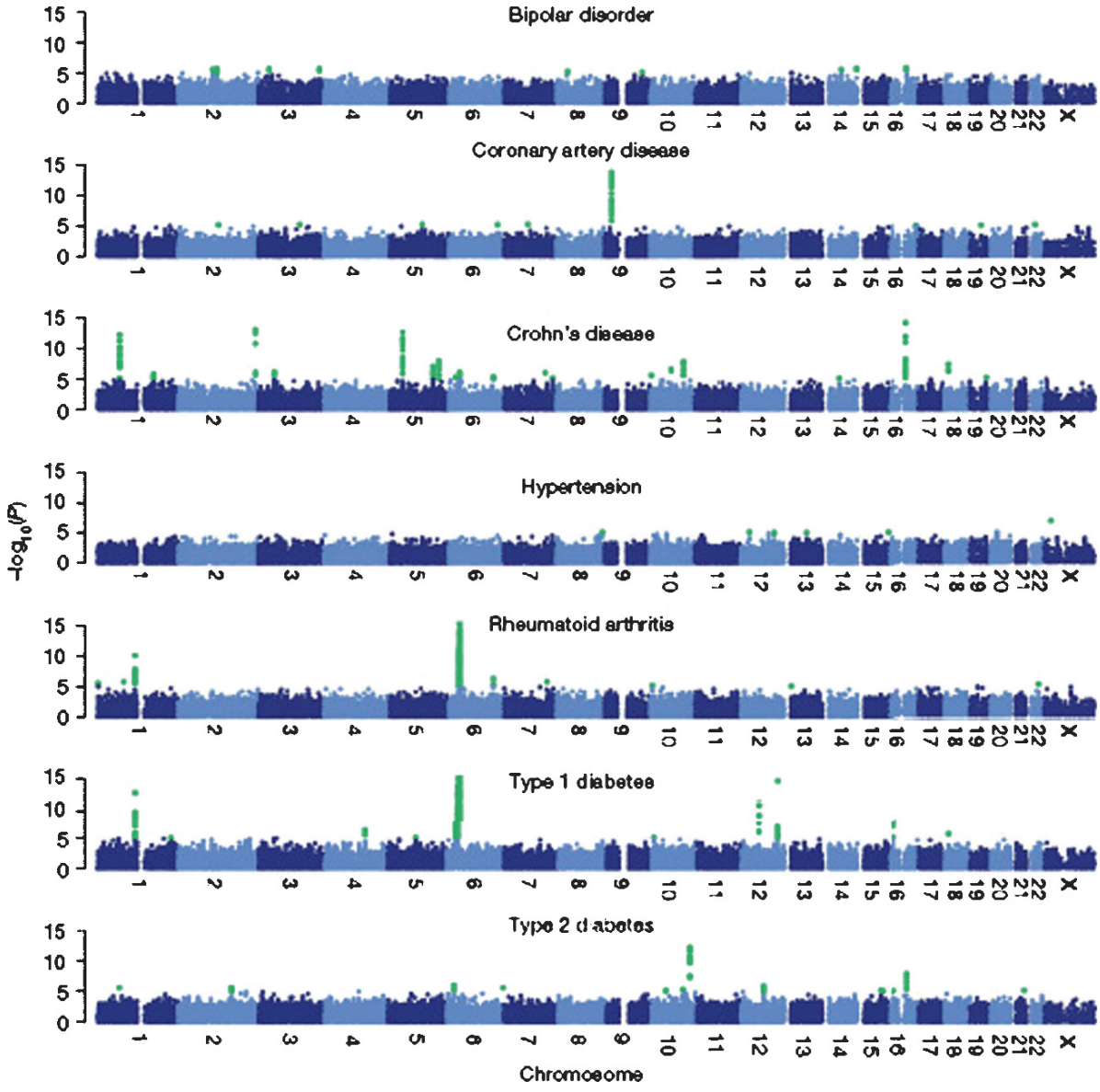

Fig. 8.8 Results observed by the Wellcome Trust Case Control Consortium in a large multi-disorder genome-wide association study. This study examined seven different complex disorders and performed genome-wide association tests for all traits using a common panel of healthy control individuals. Each of these plots (such plots are sometimes called Manhattan plots) shows the results of all 500,000 significance tests for association between each of the 500,000 SNPs and the specified disease. The *y*-axis of each plot is $-\log_{10}$ (*P*-value), so that a value of 6 implies a *P*-value of 10^{-6} (such an event would be seen by chance alone once out of every 1 million trials)

8.3.5 Replication and Validation

A problem with performing 500,000 tests at once is that one expects that 25,000 will be "significant" at $P<0.05$ by pure chance. Even when stringent criteria are applied to control for the false-positive rate, such as Bonferroni correction or use of False Discovery Rate, it is inevitable that if one places all the tests in rank order from the lowest to the highest *P*-value, that in amongst the significant tests at the lowest *P*-value range, there will be many tests that are spuriously considered positive. It is felt that the only way around this problem, to distinguish false positives from true posi-

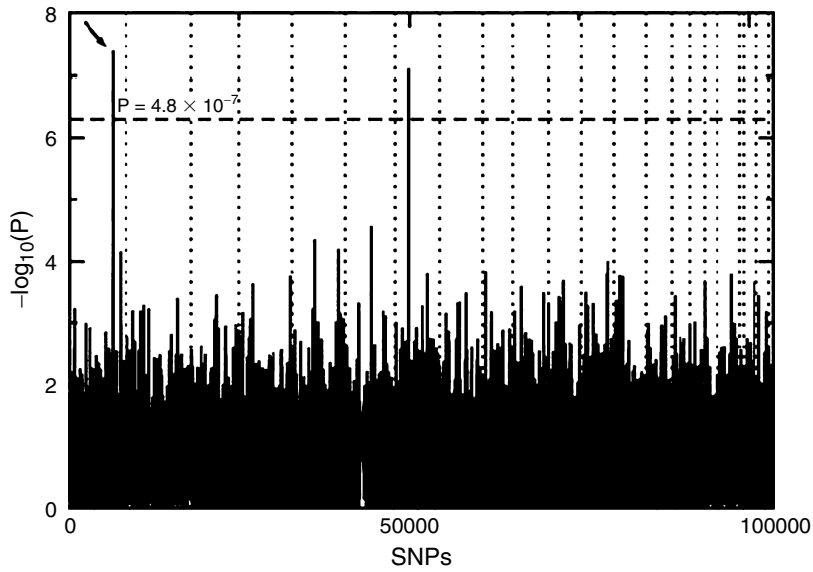

Fig. 8.9 A plot similar to that in Fig. 8.8, showing the outstandingly strong signal from the association of macular degeneration with complement factor H. The *dashed line* is the Bonferroni critical value for $P<0.05$, implying that any point above this line would be expected to occur by chance only 1 out of 20 times even after doing the 100,000 tests. (After [9])

tives, is to "replicate" the study. The word "replicate" is placed in quotes because of course there is no way to truly replicate a human study. Each individual is unique and each set of environmental circumstances is unique. At best, a second study on a similar but independent second population sample might identify overlapping sets of genomic regions harboring variation associated with disease risk. If so, this does indeed lend support to the initial positive result. The rub is that the second population is not identical, and the differences in genotypic and environmental composition between the two studies may in fact account for the difference between the results. That is, it may truly be a positive in the first study and not in the second. For now we hope that this is relatively rare, and are forced to rely on replication as a signature of real and repeatable effects.

8.3.6 Age-Related Macular Degeneration and Complement Factor H

In the early days, when the human genetics community was coming to grips with the idea that genome-wide association studies might actually work, Klein et al. [9] published a paper that showed that it could work far better than anyone could have hoped. The disorder was age-dependent macular degeneration, and they applied a simple case-control design. What was remarkable about the study was that they genotyped only 116,204 SNPs (using one of the early commercial chips) in a ridiculously small sample of 96 cases and 50 controls. To have a test that remains significant in the face of 116,204 tests would require an odds ratio of something like 6.0, and in fact, this is just what they found (Fig. 8.9). The positive hit was in the gene for complement factor H, and the result immediately sent the AMD community scrambling to understand the role of this immunity factor in macular degeneration risk.

8.4 Admixture Mapping and Population Stratification

8.4.1 How to Quantify Admixture

Before considering how to use admixture for mapping purposes, first consider how one might try to determine the degree of admixture of an individual's genome, and whether it is possible to infer which alleles came from which population. If one could identify the "parental" populations from which the admixed population derives, then the first thing to do is to estimate allele frequencies in the parental and admixed populations. In the extreme example where the allele frequencies are 0 and 1 in the parentals, it is easy to see that the allele frequency in the admixed population directly gives an estimate of the proportion of the alleles derived from the second population. If instead the allele frequencies in the two parental populations are p_1 and p_2,

and the frequency in the admixed population is p_a, then the admixture proportion, a, giving the proportion of the alleles derived from the second population, is:

$$\alpha = \frac{|\, p_a - p_1 \,|}{|\, p_2 - p_2 \,|}$$

It turns out this is a maximum-likelihood estimator for this simple single gene case. The situation gets more interesting when we have genome-wide data. For each region of the genome it is possible to estimate the proportion derived from each parental population, but what we really want is to identify for each individual the population of origin of that individual's two alleles. This is much easier with runs of SNP alleles along the chromosome, or haplotype segments. Based on the frequencies in the two parental populations, there are methods that produce reasonably accurate calls of the stretches of the genotype derived from each parental population. One effective approach applies a Markov hidden Markov model to the genotype data [22].

8.4.2 Using Admixture for Mapping

If two different populations have differing risk of a complex disorder, and there is an admixed population that also manifests the disorder, if one could identify regions of the genome derived from each population for each admixed individual, then a means of mapping might be to look for an association between disease status and population-of-origin of genomic segments. These methods are still being refined, but they appear to be very promising, especially in populations with variation in the degree of mixing of the two genomes [23]. It is good to have large blocks of unrecombined chromosomal segments to attain power, but more finely diced genomic regions are needed in order to map with fine resolution. Also, the method works best when the parental populations are well defined, and when there are only two parental populations that are widely separated from each other historically (to maximize allele frequency differences).

A reasonable target for admixture mapping methods are diseases that differ in incidence between the two parental populations. End-stage kidney disease has a lifetime incidence of about 1.5% in Europeans and about 7.5% in African Americans. At the outset we do not know whether there is a genetic basis for this, but admixture mapping could in principle identify genetic factors if they exist. One particular form of end-stage kidney disease that shows strong familial clustering is focal segmental glomerulosclerosis (FSGS). Relative to Europeans, African Americans have a fourfold increased risk for FSGS and an 18- to 50-fold increased risk for HIV-1-associated FSGS. For this reason, Kopp et al. [10] identified 190 African-American cases and 222 controls for FSGS, obtained genome-wide SNP data and applied admixture mapping. On chromosome 22 they found a region with a LOD score of 9.1, implying that African ancestry for this chromosomal region inflated the risk of FSGS by more than ninefold. Subsequent genotyping of additional SNPs in additional samples narrowed the mapping to the gene *MYH9*. The precise mutation(s) responsible for the elevated risk of African alleles are still not known, but this success and the relative ease of application of admixture mapping in studies of African American population samples, make it likely that we will see many future successes in its application.

8.4.3 The Perils of Population Stratification

Many complex disorders display a wide range of incidences across different human populations. At the outset we cannot say whether the difference in incidence is due to a difference in gene frequencies or whether differences in environmental exposures account for the variation in disease risk. Sometimes a population will face a change in an environmental factor, and then the role of environment can become starkly clear. For example, the increase in saturated fat consumption in the diet of Chinese, especially in large cities, is being accompanied by a sharp increase in cardiovascular disease [24]. The increase in protein content of the diet in post-World War II Japan was accompanied by an astonishing increase in the average stature of that population. But in addition to such clear environmental effects, many genes have allele frequencies that differ among populations, and whenever we try to do association tests when there are differences in disease incidence and allele frequencies, we must be wary of a serious artifact.

Suppose two populations have disease incidences of 4% and 20%. These two populations have been isolated geographically for thousands of years, and many alleles differ in frequency. Suppose one particular gene has allele frequencies of 0.10 and 0.30 in the two populations. Now imagine that there was a large influx of individuals from the second population into the first population, and the population sample consists of a 50/50 mix of individuals from the two populations, but investigators were unable to keep track of the ancestral origin of each individual. The population sample contains hidden stratification of these two population groups. The allele frequency in the sample would be $(0.10+0.20)/2=0.15$, and the disease incidence would likewise be the average of the two populations or 12%. But, assuming that there is zero association between this gene and the disease, the table of genotype and phenotype frequencies would be:

	AA	*Aa*	*aa*
Diseased	16	144	320
Healthy	76	856	2,580

This table was constructed by calculating the Hardy–Weinberg proportions in each population (frequencies of 0.01, 0.18, and 0.81 in one population and 0.04, 0.32, and 0.64 in the other), taking the average frequencies across the two populations for each genotypic class, and then calculating the disease incidence for each genotype. The Chi-square test of heterogeneity is $\chi^2=10.53$, for which $P<0.005$. We have generated an association that appears significant purely due to the fact that the population with the higher disease incidence happened by chance to have a higher allele frequency for this SNP. In fact, for any SNP having an allele frequency difference of sufficient magnitude between the two populations, there will be this same kind of spurious association. This is why it is so crucial to avoid hidden population stratification in association testing.

8.4.4 How to Correct for Hidden Population Stratification

Fortunately, there are ways to identify the problem of hidden population stratification that allow some degree of correction of the false positives it causes. First note that a mixture of two populations having different allele frequencies results in genotype frequencies that depart from Hardy–Weinberg proportions. The easiest way to see this is to imagine populations with allele frequencies 0 and 1. A mixture of the two would give 50% *AA*, 0% *Aa*, and 50% *aa* individuals. The allele frequency is 50%, but there is a massive deficit of heterozygotes (or excess of homozygotes). One way to tease apart the sample into its original populations is to try to find clusters of individuals each of which form a Hardy–Weinberg population. This is the basic idea behind the program STRUCTURE, which is widely used in heterogeneous population samples to try to understand its partitioning into units [16].

Another approach, first used in 1978 by L.L. Cavalli-Sforza's group [12], is to apply a principal components analysis to the genotype data. This is a multivariate statistical procedure that identifies linear combinations of the SNPs that explain the most among-individual variability (arbitrarily number coded as, for example 0, 1, and 2 for the three genotypes). Generally there are multiple orthogonal sets of "axes" or vectors of SNPs that are needed to describe the variation. What PCA does is provide the weightings for each SNP and each such principal component. In the end, one can simply plot these principal components for each individual, and to the extent that individuals are more genetically similar to each other, they will fall closer together in these plots. If there are separate clusters of individuals, as there might be if there were discrete populations, these would appear as clusters in the PCA plot. Recently this method was applied to a sample of some 7,000 individuals from Europe genotyped at 500,000 SNPs [13], and the plot of the first two principal components produces an astonishingly good reproduction of the geographic map of Europe (Fig. 8.10). What does this imply? Just that there is a measurable isolation by distance among Europeans, and that historically people have tended to marry and settle down not far from their birth place.

To use PCA for association testing, one could identify the discrete clusters and use this as a covariate in the analysis, trying to explain as much of the variance in disease risk by population of origin first, and then explaining the remainder with the allele frequencies. Alternatively, one could directly use the principal components loadings as cofactors in the association analysis. This is an area of active research, and some of the newer approaches for dealing with genetic ancestry

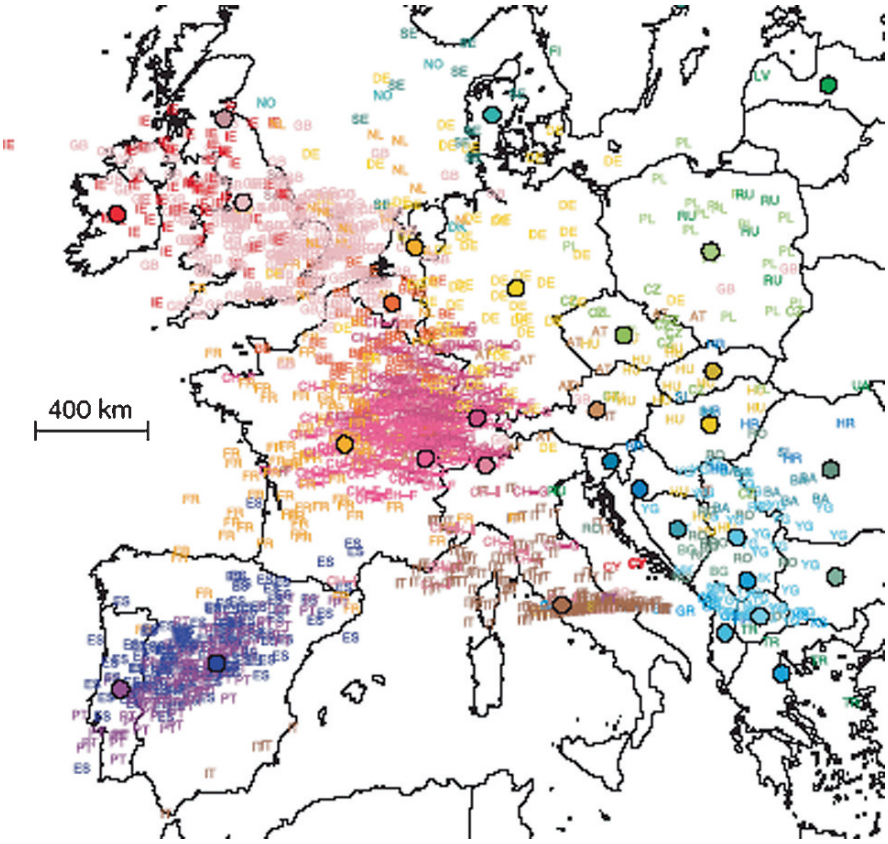

Fig. 8.10 The principal components plot from a study of 500,000 SNPs across a European sample of nearly 7,000 individuals. (From [13]). The raw genotype data were analyzed by Principal Components Analysis to try to find collections of SNPs that explain the most variance. A Principal Component is a combination of weightings of a subset of SNPs, and so after the PCA is run, each individual has a value for each principal component (PC1, PC2, PC3, etc.). If one plots a point (*x, y*) for the values (PC1, PC2) for each individual, one gets a plot like that shown. Note the impressive correspondence to the map of Europe, indicating that simple geographic distance is well correlated with the degree of genetic difference between individuals living that distance apart

and population structure in association studies are presented in Chap. 20 in this volume.

8.5 Complications

The models that we have presented up to now were purposely simplified so that the principle concepts would be clear. We assumed that the effects of many genes were additive, and proceeded to fit real data to this model without particularly questioning whether the model was correct. In fact, several factors can contribute to departures from this simple additive model, and many people think that these departures are virtually ubiquitous. Departures from additivity do not bring to a halt hopes of finding genes that act on complex traits, but they do make the problem more challenging.

8.5.1 Genotype by Environment Interaction

One of the challenges of studying the genetics of complex traits in humans is that we can never measure the same genotype in more than one controlled environment. Monozygotic twins at least give us some idea of the impact that different environments may have as a zygote undergoes development and eventually manifests mature phenotypes. With model organisms, where it is possible to produce many individuals with the same genotype, a very simple experiment produces a profoundly important result. The experiment is to simply rear the set of genotypes in two or more environments. Figure 8.11 shows an example of one such experiment, where a set of *Drosophila* lines were reared at two different temperatures, and body mass was measured in the resulting adult flies. As you can see, some lines gain weight

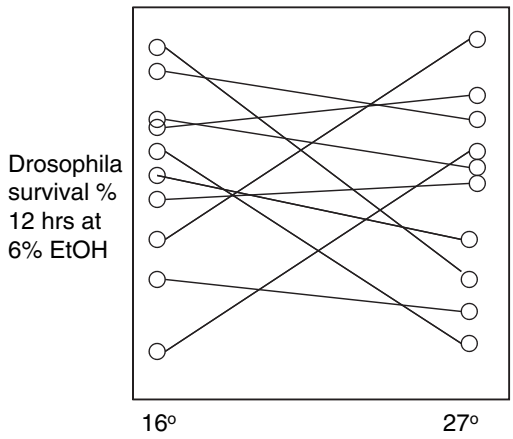

Fig. 8.11 A typical study of genotype × environment interaction obtained from model organism studies where the same genotypes can be reared in two or more environments. This kind of design nearly ubiquitously shows crossing of the mean phenotype lines, indicating a nonlinear effect of the environment attributable to a genotype × environment background. (Data from Kristi Montooth)

when moving from the low to the higher temperature, and other lines do the opposite.

Whenever the lines connecting the mean phenotypes across a range of environments cross, as they do in Fig. 8.11, this is a form of genotype × environment (G×E) interaction. More formally, we could set up an analysis of variance of these data, where the two factors are genotype and environment, and the interaction term in the analysis of variance would quantify the degree of G×E interaction. The impressive feature of this simple experiment is that whenever an experiment of this sort is done having any power at all, the observation of significant G×E is nearly universal.

Human examples of G×E interaction are a bit harder to find, but this is only because one has to define genotypes by particular targeted subsets of SNPs. A clear example of human G×E comes from drug responses. The particular example of *VKORC1* and warfarin response is a prime example. The phenotype of clotting time shows a strong interaction between genotypes at *VKORC1* and the environment of having taken warfarin. Given the ubiquity of G×E in animal and plant studies, one might expect that a differential response to drugs, varying with the genotype of the patient, is probably also nearly universal.

A particularly good example of a genotype by environment is seen in α_1-antitrypsin. The antiproteolytic activity of human serum was detected in 1897, and in 1900 Landsteiner showed this activity to be located in the albumin fraction. Antiproteolytic activity is measured by hydrolysis of artificial substrates by trypsin in the presence of the serum to be tested. The concentration of antiproteolytic activity increases quickly, for example with bacterial infection, after injection of typhoid vaccine, and during pregnancy. Interindividual differences in levels of antiproteolytic activity in the blood were first observed in 1963. A simple recessive mode of inheritance was proposed for low α_1-antitrypsin levels. Many different alleles have been discovered that vary widely in their activity levels. The gene is located on 14q31-32; it spans 10.2 kb, and has five exons. Two variants, Z and S, are especially important because the α_1-antitrypsin level is appreciably reduced relative to the common M type.

Subcutaneous injection of typhoid vaccine and diethylstilbestrol leads to a 100 % increase in activity of subjects with the MM type. Heterozygotes of the MZ type show a moderate increase, whereas in homozygotes of the ZZ type hardly any increase is seen. Many studies have shown that the rate of obstructive pulmonary disease in these ZZ individuals is at least 15 times the rate in the general population. Among ZZ homozygotes only 70–80% develop obstructive emphysema, and in heterozygotes the frequency is much lower. When a patient is exposed to recurrent bronchial irritation, such as that caused by smoking or frequent infections, these enzymes cause digestive damage to the lungs. Tobacco smoking enhances the danger of bronchial infections and hastens the progress of the disease. Once we are in the era of widespread genotyping for medical diagnostics, individuals who are found to be ZZ homozygotes and possibly ZM heterozygotes ought to get extra guidance regarding their exceptional risk of COPD, especially if they smoke.

The a_1-antitrypsin polymorphism is an example in which there is a subset of genotypes with heightened environmental sensitivity. The associated diseased condition can be thought of as one of reduced penetrance, and that penetrance is increased by an environmental trigger. The genetics of COPD appears to be complex, but for individuals with the ZZ genotype of a_1-antitrypsin, the disease is practically Mendelian. This is one of the more hopeful situations motivating the study of genotype × environment interactions – many diseases that we think of as complex and unpredictable may prove to have a simple gene of large effect whose otherwise low penetrance is triggered by

an identified environmental factor. Such situations are also highly sought after because they provide a means whereby early genotypic analysis may result in an ability to give advice about environmental hazards that could greatly impact disease prevalence.

8.5.2 Epistasis

In the context of complex traits, epistasis is the situation when the risk of the disorder departs from an additive effect across two or more risk-elevating SNPs. Table 8.3 makes the situation clear. If one locus has marginal phenotypes (means across all other factors) of a_1, a_2, and a_3, and the other locus has marginal phenotypes of b_1, b_2, and b_3, then the two-locus genotypes have phenotypes that might fit the additive pattern as depicted in Table 8.3. Any departure from this additivity is an example of epistasis. One extreme example is where all the genotypes in the table have one phenotype, but the *aabb* genotype in the lower right corner has a radically different genotype. Consider two parallel pathways, where the organism requires the product of one or the other pathway, and the *aa* genotype knocks out one pathway, and the *bb* genotype knocks out the other. In this case, all the genotypes except *aabb* would get the required product, but the *aabb* doubly homozygous mutant would fail in both pathways and would produce the extreme phenotype. This kind of epistasis is rampant in model organisms, but when we try to test for it in human complex traits, it is not so easy to find. The reason is primarily due to the greatly reduce statistical power to detect such interaction effects. Given this low statistical power, it is premature to conclude that epistasis is not very prevalent in humans.

Table 8.3 Two-locus genotypes and additive genotypic effects[a]

	BB	*Bb*	*bb*
AA	$a_1 + b_1$	$a_1 + b_2$	$a_1 + b_3$
Aa	$a_2 + b_1$	$a_2 + b_2$	$a_2 + b_3$
aa	$a_3 + b_1$	$a_3 + b_2$	$a_3 + b_3$

[a]Define (a_1, a_2, a_3) as the effect of genotypes *AA*, *Aa*, and *aa* on the phenotype, and (b_1, b_2, b_3) as the effects of genotype *BB*, *Bb*, and *bb*, then the matrix below gives the expected genotypic effects for the nine pairwise genotype combinations assuming that the two loci have additive effects. These genotypic effects would be equivalent to the measured phenotypes in the environmental effect is zero

It has been argued that epistasis is especially likely to be found for phenotypes that are closely related to molecular function. The argument is that molecular biology is loaded with intermolecular interactions, and so if there is polymorphism in pairs of molecules that interact in some key pathway, then it is all the more likely that those variants may display an interaction in disease risk. Following this reasoning, Dimas et al. [5] examined pairs of SNPs for possible interactions in driving transcript abundance. They used the genome-wide expression data generated by the Sanger Centre in the 210 cell lines from the unrelated individuals whose genotypes were scored in the HapMap study. Reasoning that coding SNPs might be compensated for by flanking SNPs, they specifically looked for coding-flanking SNP pairs that influenced transcript abundance in nonadditive ways. After identifying non-synonymous SNPs that affect expression and flanking SNPs that also affect expression, they performed an ANOVA test for each SNP pair to detect main effects and pairwise interactions. At a significance level of $P<0.001$ they expected 331 such interactions by chance, but observed 412. In this set were several cases of strong and highly significant interactions. Although the final conclusion does not overwhelmingly suggest that pairwise interactions are rampant in the human genome, the test had relatively low power given the small sample sizes. As our ability to apply tests of epistasis to larger samples targeted at specific pathways improves, it does seem likely that epistatic interaction among human genetic variants will be seen to play as important a role as has been found in genetic model organisms.

8.6 Missing Heritability: Why is so Little Variance Explained by GWAS Results?

One of the more surprising results from the genome-wide association studies has been that they uniformly find only SNPs of very small effect, and that even the sum of the effects of all the SNP associations that are found only explains a small proportion of the total genetic variance. This implies that if one has the SNP genotype for all the SNPs that impact a trait, one still has rather poor ability to predict the phenotype. This is surprising in light of the density of SNP genotypes obtained (one every 3 kb on average) and the large

sample sizes (in some studies in excess of 30,000). The most dramatic example of this poor prediction ability is the case of body height (stature). The heritability of stature in humans is approximately 80%, making it one of the more strongly heritable complex phenotypes that we know. Despite this, even the top 20 SNPs found to be associated with stature explain less than 5% of the variance. Because we know from the heritability studies that there are genetic factors explaining the familial resemblance, this problem is sometimes called "missing heritability"; or, by analogy with dark matter in astrophysics, it is also called "dark heritability."

There are several reasons why a GWAS study may fail to explain more of the genetic variance in a complex trait. First, the SNPs that are used as markers are not expected to be the causal factors that drive the phenotype, but instead are correlated with the trait-affecting SNPs. This indirect association would erode the prediction power. Second, the SNPs that are used as markers are only quite common, because they were chosen from the HapMap studies, which specifically sought to catalog common SNPs. If much of the variance in traits is driven by rare SNPs, the correlation between the SNP markers that were used and these rare SNPs could be quite low. Third, it is clear that the complex traits that are studied include an environmental component, and if there are genotype × environment interactions (G×E), each SNP genotype will be averaged across all the environments, so that its effect would appear to be eroded compared with an SNP that had no such G×E interaction. Soon we hope to have the means to directly test for G×E interactions, but the primary challenge that must be tackled is to have accurate and meaningful measurements of the environment. Fourth, the statistical models have only made use of single SNPs at a time, and the trait may instead be driven by interactions among SNPs, or epistasis. It is also possible that there are other sources of heterogeneity, including epigenetic differences among individuals.

8.7 Concluding Remarks

The human genetics community is striving to improve methods for identification of genes that underlie complex genetic disorders and to understand how the effects of genes combine to produce inflated risk of disease. As part of the effort to better understand the role of rare alleles, the 1000 Genomes Project (www.1000genomes.org) was launched to provide the stimulus to accelerate the development of sequencing technologies that reduce the cost while increasing the speed and accuracy of whole-genome resequencing methods. Statistical methods need to be developed that accommodate the known complexities that may connect variation at the genotypic and phenotypic levels. While we can have confidence that methods of genome-wide association testing based on full genome sequence will be developed and improved in the near future, the prediction of an individual's disease risk given only his or her genome sequence may never attain useful accuracy (apart from extreme alleles that are nearly deterministic for some disorders, such as Mendelian disorders), especially if the disorder is heavily impacted by stochastic environmental factors, or by complex interactions between genotype and environment. But prediction of individual risk could make an enormous difference to public health, especially if environmental amelioration of that risk were possible, and so the drive to maximize prediction accuracy will motivate work in this area for years to come.

References

1. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888
2. Cheung VG, Bruzel A, Burdick JT, Morley M, Devlin JL, Spielman RS (2008) Monozygotic twins reveal germline contribution to allelic expression differences. Am J Hum Genet 82:1357–1360
3. Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, Schwarz UI, Ritchie MD, Stein CM, Roden DM, Smith JD, Veenstra DL, Rettie AE, Rieder MJ (2008) A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 112:1022–1027
4. Detera-Wadleigh SD, Badner JA, Berrettini WH et al (1999) A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2. Proc Natl Acad Sci USA 96:5604–5609
5. Dimas AS, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P, Tavaré S, Dermitzakis ET (2008) Modifier effects between regulatory and protein-coding variation. PLoS Genet 4:e1000244
6. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman Group Ltd, London
7. Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52:399–433

8. Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. PLoS Genet 4(12):e1000279
9. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
10. Kopp JB, Smith MW, Nelson GW, Johnson RC, Freedman BI, Bowden DW, Oleksyk T, McKenzie LM, Kajiyama H, Ahuja TS, Berns JS, Briggs W, Cho ME, Dart RA, Kimmel PL, Korbet SM, Michel DM, Mokrzycki MH, Schelling JR, Simon E, Trachtman H, Vlahov D, Winkler CA (2008) MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. Nat Genet 40:1175–1184
11. Marchini J, Howie B (2008) Comparing algorithms for genotype imputation. Am J Hum Genet 83:535–539
12. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201:786–792
13. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456:98–101
14. Ohta T, Kimura M (1969) Linkage disequilibrium due to random genetic drift. Genet Res 13:47–55
15. Poulsen P, Vaag A (2003) The impact of genes and pre- and postnatal environment on the metabolic syndrome. Evidence from twin studies. Panminerva Med 45:109–115
16. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
17. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, McLeod HL, Blough DK, Thummel KE, Veenstra DL, Rettie AE (2005) Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. N Engl J Med 352:2285–2293
18. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
19. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, The International HapMap Consortium (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918 [Abstract] [PDF]
20. Schaid DJ, Olson JM, Gauderman WJ, Elston RC (2003) Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. Hum Hered 55:86–96
21. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516
22. Sved JA (1968) The stability of linked systems of loci with a small population size. Genetics 59:543–563
23. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. Am J Hum Genet 79:1–12
24. Wang Y, Mi J, Shan XY, Wang QJ, Ge KY (2007) Is China facing an obesity epidemic and the consequences? The trends in obesity and chronic disease in China. Int J Obes 31:177–188
25. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature 447:661–678