Chapter 15

Mapping Genes Conferring Susceptibility to Complex Diseases

15

KEY CONCEPTS

- Many common diseases are multifactorial, with a variety of environmental and genetic factors each reducing or increasing an individual's susceptibility to that disease. They are also usually complex, having many different possible causes.
- The major genetic effects on human health and disease come from genetic factors that affect susceptibility to common complex diseases, rather than those that cause Mendelian diseases. A main aim of much current human genetic research is to identify such factors.
- Evidence for the role of genetic factors in many common complex diseases comes from studies of families, twins, and adopted people. Such studies need careful interpretation to disentangle genetic effects from the effects of a shared family environment.
- Linkage analysis has shown only very limited success in mapping susceptibility factors for common complex diseases. Standard lod score analysis cannot be used because it is not possible to specify certain parameters for the susceptibility factors, such as the mode of inheritance, gene frequencies, or penetrances. Instead, non-parametric methods are used, which do not require a detailed genetic model to be provided. Non-parametric linkage methods take affected relatives and look for chromosomal segments that they share more often than expected by chance. Affected sib pairs are the most frequently used sets of relatives.
- An alternative approach to finding susceptibility factors is to seek populationwide statistical associations between a certain genotype and a disease. Such associations may arise because many supposedly unrelated people share a chromosome segment inherited from a distant common ancestor who carried a susceptibility factor. However, not all populationwide associations have a genetic cause.
- The International HapMap Project has defined the ancestral chromosome segments in four human populations, and cataloged markers (tagging SNPs) that can be used to identify them.
- Shared ancestral chromosome segments are extremely small (typically a few kilobases). Thus, seeking associations requires the use of a dense array of closely spaced markers. It has only recently become feasible to conduct genomewide association studies.
- Identifying susceptibility factors, either by linkage studies or by association, has proved unexpectedly difficult. Only recently have studies become sufficiently powerful to identify susceptibility factors reliably.
- Association studies can only identify factors that are present on chromosome segments that are shared by many individuals in the study group. Thus, the ability of association studies to identify susceptibility factors for common complex diseases depends on the common disease–common variant hypothesis, which supposes that most susceptibility factors are ancient variants found on shared ancestral chromosome segments.
- An alternative hypothesis, the mutation–selection hypothesis, supposes most factors to be the result of a highly heterogeneous set of individually rare recent mutations.
- If the mutation–selection hypothesis is true, identifying susceptibility factors will require large-scale resequencing of individual genomes from cases and controls. New sequencing technologies make this feasible.

The main genetic contribution to morbidity and mortality is through the genetic component of common complex diseases. These diseases have no one single cause, but result from the cumulative effects of a variety of genetic and environmental factors, often different in different affected individuals. Identifying the genes concerned is a central task for medical research. However, this is a much more formidable task than identifying the mutations that cause monogenic diseases. For Mendelian diseases, given a sufficient collection of affected families, linkage analysis as described in Chapter 14 can usually localize the causative gene to a small chromosomal segment containing only a handful of candidate genes. Similar studies in complex diseases have been much less successful.

A meta-analysis by Altmüller and colleagues in 2001 reviewed 101 linkage studies in 31 complex diseases. The result was sobering. Candidate regions defined in different linkage studies of the same disease rarely coincided. There were some real successes, reviewed by Lohmueller and colleagues (see Further Reading) but, despite huge efforts by leading research teams, overall the studies had made only limited progress in localizing susceptibility genes. Recently this has changed. A combination of new technology (high-resolution SNP chips) and new understanding of the structure of human genomes (the HapMap project) is finally allowing susceptibility factors to be reliably identified. In this chapter, we describe the ways in which this difficult problem has been approached.

15.1 FAMILY STUDIES OF COMPLEX DISEASES

Before work can begin on uncovering the genetic factors involved in common complex diseases, it is necessary to establish the criteria by which people are to be labeled as affected or unaffected. With Mendelian syndromes it is usually fairly obvious which features of a patient form part of the syndrome and which are coincidental. Different features may have different penetrances, but basically the components of the syndrome are those that co-segregate in a Mendelian pattern. Things are much less clear cut for non-Mendelian conditions. Even physically obvious conditions such as the major birth defects are very variable in severity—where is the line to be drawn? Should we lump together or split apart the various types of congenital heart malformation for example?

Psychiatric and behavioral conditions are especially difficult. A diagnostic label can be valid, in the sense that two independent investigators will agree whether or not it applies to a given patient, without being biologically meaningful. Great efforts have been made to establish valid diagnostic categories. These are codified in the successive versions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) published by the American Psychiatric Association. Adhering to such conventions helps make different studies comparable, but it does not guarantee that the right genetic question is being asked.

Having established clear diagnostic criteria for a complex disease, it is necessary to show whether or not genetics has a role in the etiology of the condition. People who share more of their DNA should be more likely to share the phenotype under investigation. The obvious way to approach this is to show that the character runs in families. This involves family, twin, or adoption studies. The difficulty of distinguishing the effects of shared family environment from those of heredity has often made such studies controversial, especially for psychiatric conditions.

The risk ratio (λ) is a measure of familial clustering

Nobody would dispute the involvement of genes in a character that consistently gives Mendelian pedigree patterns or that is associated with a chromosomal abnormality. However, with non-Mendelian characters, whether continuous (quantitative) or discontinuous (dichotomous), there is no such reality check. It is necessary to prove claims of genetic determination. The obvious way to approach this is to show that the character runs in families.

The degree of family clustering of a disease can be expressed by the **risk ratio** (λ_R), the risk to a relative (R) of an affected proband compared with the risk in the general population. A risk ratio of 1 implies no additional risk above that of the general population. Separate values can be calculated for each type of relative, for example λ_S for sibs. The mathematical properties of λ_R are derived in the 1990

TABLE 15.1 RISK OF SCHIZOPHRENIA AMONG RELATIVES OF SCHIZOPHRENICS: POOLED RESULTS OF SEVERAL STUDIES

Relative	No. at risk	Risk (%)	λ
Parents	8020	5.6	7
Sibs	9920.7	10.1	12.6
Sibs, one parent affected	623.5	16.7	20.8
Offspring	1577.3	12.8	16
Offspring, both parents affected	134	46.3	58
Half-sib	499.5	4.2	5.2
Uncles, aunts, nephews, nieces	6386.5	2.8	3.5
Grandchildren	739.5	3.7	4.6
Cousins	1600.5	2.4	3

Numbers at risk are corrected to allow for the fact that some at-risk relatives were below or only just within the age of risk for schizophrenia (say, 15–35 years). λ values are calculated assuming a population incidence of 0.8%. [Data from McGuffin P, Shanks MF & Hodgson RJ (eds) (1984) The Scientific Principles of Psychopathology. Grune & Stratton.]

papers by Risch (see Further Reading). As an example, Table 15.1 shows pooled data from several studies of schizophrenia. Family clustering is evident from the raised λ values, for example a sevenfold increased risk for somebody, one of whose parents is schizophrenic. As expected, λ values drop back toward 1 for more distant relationships such as nephews, nieces, and cousins.

Shared family environment is an alternative explanation for familial clustering

Geneticists must never forget that humans give their children their environment as well as their genes. Many characters run in families because of the shared family environment—whether one's native language is English or Chinese, for example. One always has to ask whether shared environment might be the explanation for familial clustering of a character. This is especially important for behavioral attributes such as IQ or schizophrenia, which depend at least partly on upbringing. Even for physical characters or birth defects it cannot be ignored: a family might share an unusual diet or some traditional medicine that could cause developmental defects. Among the Fore people of New Guinea, a degenerative brain disease, kuru, ran in families because, as part of funerary rituals, close relatives ate infectious brain material from deceased affected people. Something more than a familial tendency is necessary to prove that a non-Mendelian character is under genetic control. These reservations are not always as clearly stated in the medical literature as perhaps they should be. Table 15.5 on p. 472 shows what can happen if shared family environment is ignored.

Twin studies suffer from many limitations

Francis Galton, the brilliant but eccentric cousin of Charles Darwin, who laid so much of the foundation of quantitative genetics, pointed out the value of twins for human genetics. Monozygotic (MZ) twins are genetically identical clones and should always be **concordant** (both the same) for any genetically determined character. This is true regardless of the mode of inheritance or number of genes involved; the only exceptions are for characters dependent on post-zygotic somatic genetic changes (the pattern of X-inactivation in females, the repertoire of functional immunoglobulin and T-cell receptor genes, and random post-zygotic somatic mutations leading to mosaicism). Dizygotic (DZ) twins share half their genes on average, the same as any pair of sibs.

Concordance can be calculated in two ways. Pairwise concordance counts the number of pairs of twins, ascertained through an affected proband, in which

TABLE 15.2 TWIN STUDIES IN SCHIZOPHRENIA

Study	Country	Concordant pairs	
		MZ	DZ
Kringlen et al. (1968)	Norway	14/50 (0.28)	6/94 (0.06)
Fischer et al. (1969)	Denmark	5/21 (0.23)	4/41 (0.10)
Tienari et al. (1975)	Finland	3/20 (0.15)	3/42 (0.07)
Farmer et al. (1987)	UK	6/17 (0.35)	1/20 (0.05)
Onstad et al. (1991)	Norway	8/24 (0.33)	1/28 (0.04)

The numbers show the total number of twin pairs ascertained and the number that were concordant (both twins diagnosed as schizophrenic). Diagnostic and inclusion criteria varied between studies; despite the heterogeneity there is a clear tendency for more monozygotic (MZ) than dizygotic (DZ) pairs to be concordant. [For references, see Onstad S, Skre I, Torgersen S & Kringlen E (1991) *Acta Psychiatr. Scand.* 83, 395–401.]

both twins have the condition (concordant, +/+) and the number of pairs in which only one twin is affected (discordant, +/–). Probandwise concordance is obtained by counting a pair twice if both were probands, and thus gives higher values for the concordance. For example, in the study by Onstad et al. cited in **Table 15.2**, in 7 of the 24 MZ twin pairs, both twins were independently ascertained as affected probands. Thus, the probandwise concordance was 0.48, calculated as [(8 + 7)/(24 + 7)], in comparison with the pairwise concordance of 0.33 (8/24). Probandwise concordances are thought to be more comparable with other measures of family clustering.

Genetic characters should show a higher concordance in MZ than DZ twins, and many characters do. However, a higher concordance in MZ twins than in DZ twins does not prove a genetic effect. For a start, half of DZ twins are of different sexes, whereas all MZ twins are the same sex. Even if the comparison is restricted to same-sex DZ twins (as it is in the studies shown in Table 15.2), at least for behavioral traits the argument can be made that MZ twins are more likely to look very similar, to be dressed and treated the same, and thus to share more of their environment than DZ twins.

Separated monozygotic twins

Monozygotic twins separated at birth and brought up in entirely separate environments seem to provide an ideal experiment for separating the effects of shared genes and shared environment. Francis Crick once made the tongue-in-cheek suggestion that one of each pair of twins born should be donated to science for this purpose. Such separations happened in the past more often than one might expect—the birth of twins was sometimes the last straw for an overburdened mother. Fascinating television programs can be made about twins reunited after 40 years of separation, who discover they have similar jobs, wear similar clothes, and like the same music. As research material, however, separated twins have many drawbacks:

- Any research is necessarily based on small numbers of arguably exceptional people.
- The separation was often not total—often the twins were separated some time after birth, and were brought up by relatives.
- There is a bias of ascertainment—everybody wants to know about strikingly similar separated twins, but separated twins who are very different are not newsworthy.
- Research on separated twins cannot distinguish intrauterine environmental causes from genetic causes. This may be important, for example in studies of sexual orientation (the gay gene), in which some people have suggested that maternal hormones may affect the fetus *in utero* so as to influence its future sexual orientation.

Thus, for all their anecdotal fascination, separated twins have contributed relatively little to human genetic research.

Adoption studies are the gold standard for disentangling genetic and environmental factors

If separating twins is an impractical way of disentangling heredity from family environment, studying adopted people is much more promising. Two study designs are possible:

- Find adopted people who suffer from a particular disease known to run in families, and ask whether it runs in their biological family or their adoptive family.
- Find affected parents whose children have been adopted away from the family, and ask whether being adopted saved the children from the family disease.

A celebrated but controversial study by Rosenthal & Kety (see Further Reading) used the first of these designs to test for genetic factors in schizophrenia. The diagnostic criteria used in this study have been criticized, and there have also been claims (disputed) that not all diagnoses were made truly blind. However, an independent re-analysis using DSM-III diagnostic criteria reached substantially the same conclusion: it was the genes rather than the environmental influence of a schizophrenic parent that increased the risk for the offspring. **Table 15.3** shows the results of a later extension of this study.

The main obstacle in adoption studies is lack of information about the biological family, frequently made worse by the undesirability of approaching them with questions. Efficient adoption registers exist in only a few countries. A secondary problem is selective placement, in which the adoption agency, in the interests of the child, chooses a family likely to resemble the biological family. Adoption studies are unquestionably the gold standard for checking how far a character is genetically determined, but because they are so difficult, they have in the main been performed only for psychiatric conditions, for which the nature–nurture arguments are particularly contentious.

15.2 SEGREGATION ANALYSIS

Pure Mendelian and pure polygenic characters represent the opposite ends of a continuum. In between are *oligogenic traits* governed by a few major susceptibility loci, possibly operating against a polygenic background, and possibly subject to major environmental influences. **Segregation analysis** is a statistical tool for analyzing the inheritance of any character. It can provide evidence for or against a major susceptibility locus and can at least partly define its properties. The results can help guide future linkage or association studies.

Complex segregation analysis estimates the most likely mix of genetic factors in pooled family data

Analyzing data on the relatives of a large collection of people affected by a familial but non-Mendelian disease is not a simple task. There could be both genetic and environmental factors at work; the genetic factors could be polygenic, oligogenic, or monogenic (Mendelian) with any mode of inheritance, or any mixture

TABLE 15.3 AN ADOPTION STUDY IN SCHIZOPHRENIA

Case types	Schizophrenia cases among biological relatives	Schizophrenia cases among adoptive relatives
Index cases (47 chronic schizophrenic adoptees)	44/279 (15.8%)	2/111 (1.8%)
Control adoptees (matched for age, sex, social status of adoptive family, and number of years in institutional care before adoption)	5/234 (2.1%)	2/117 (1.7%)

The study involved 14,427 adopted persons aged 20–40 years in Denmark; 47 of them were diagnosed as chronic schizophrenic. The 47 were matched with 47 non-schizophrenic control subjects from the same set of adoptees. [Data from Kety SS, Wender PH, Jacobsen B et al. (1994) *Arch. Gen. Psychiatry* 51, 442–455.]

TABLE 15.4 COMPLEX SEGREGATION ANALYSIS OF HIRSCHSPRUNG DISEASE

Model	*d*	*t*	*q*	*H*	*z*	*x*	χ^2	*p*
Mixed	1.00	7.51	9.6×10^{-6}		0.01	0.15		
Sporadic							334	$<10^{-5}$
Polygenic				1.00	1.00		78	$<10^{-5}$
Major recessive locus	0.00	8.22	3.8×10^{-3}				35	$<10^{-5}$
Major dominant locus	1.00	7.56	1.2×10^{-5}			0.19	2.8	0.42

Data are for families ascertained through a proband with long-segment Hirschsprung disease (OMIM 142623). Parameters that can be varied are as follows: *d*, the degree of dominance of any major disease allele; *t*, the difference in liability between people homozygous for the low-susceptibility and the high-susceptibility alleles of a major susceptibility gene, measured in units of standard deviation of liability; *q*, the gene frequency of any major disease allele; *H*, the proportion of total variance in liability that is due to polygenic inheritance, in adults; *z*, the ratio of heritability in children to heritability in adults; *x*, the proportion of cases due to new mutation. The values shown are those that best account for the family data using the stated model. The χ^2 statistic is a standard test that compares the performance of each model with the mixed model, in which a mix of all mechanisms is allowed. A single major locus encoding dominant susceptibility explains the data as well as the mixed model. [Data from Badner JA, Sieber WK, Garver KL & Chakravarti A (1990) *Am. J. Hum. Genet.* 46, 568–580.]

of these, and the environmental factors may include both familial and non-familial variables. In complex segregation analysis, a whole range of possible inheritance patterns, gene frequencies, penetrances, and so on, are modeled by computer analysis to find the mix of scenarios that gives the greatest overall likelihood for the observed data. Factors in this mixed model are then omitted or constrained to identify the minimum that must be included so as to avoid a significant loss of explanatory power. As with lod score analysis (see Chapter 14), the question asked is how much more likely the observations are when one hypothesis is compared with another.

In the example of Table 15.4, the ability of specific models (sporadic, polygenic, recessive, dominant) to explain the data was compared with the likelihood calculated by a general mixed model, in which the computer program could freely optimize the mixture of single-gene, polygenic, and random environmental causes. All models were constrained by the overall epidemiology, sex ratios, and probabilities of ascertainment estimated from the collected data. A single-locus dominant model is not significantly worse than the mixed model at explaining the data ($\chi^2 = 2.8$, $p = 0.42$). In contrast, models assuming no genetic factors (sporadic), pure polygenic inheritance, or pure recessive inheritance perform very badly. On the argument that simple explanations are preferable to complicated explanations, the analysis suggests the existence of a major dominant susceptibility to Hirschsprung disease. One such factor has now been identified, the *RET* (rearranged during transfection) oncogene (OMIM 164761).

However clever the segregation analysis program is, it can do no more than maximize the likelihood across the parameters it was given. If a major factor is omitted, the result can be misleading. This was well illustrated by the data in Table 15.5. McGuffin & Huckle asked their classes of medical students which of their relatives had attended medical school. When they fed the results through a

TABLE 15.5 A RECESSIVE GENE FOR ATTENDING MEDICAL SCHOOL?

Model	*d*	*t*	*q*	*H*	χ^2	*p*
Mixed	0.087	4.04	0.089	0.008		
Sporadic					163	$<10^{-5}$
Polygenic				0.845	14.4	<0.005
Major recessive locus	0.00	7.62	0.88		0.11	n.s.

The data are taken from a survey of medical students and their families. The meaning of the symbols is explained in the foornote to Table 15.4. Affected is defined as attending medical school. The analysis seems to support recessive inheritance, because this accounts for the data equally well as the unrestricted model (but see the text). n.s., not significant. [Data from McGuffin P & Huckle P (1990) *Am. J. Hum. Genet.* 46, 994–999.]

segregation analysis program, it came up with results apparently favoring the existence of a recessive gene for attending medical school. Though amusing, this was not done as a joke, nor to discredit segregation analysis. The authors did not allow the segregation analysis program to consider the likely true mechanism, namely a shared family environment. The program's next best alternative was mathematically valid but biologically unrealistic. The serious point that McGuffin & Huckle were making was that there are many pitfalls in segregation analysis of human behavioral traits, and incautious analyses can generate spurious genetic effects.

Although complex segregation analysis can provide a valuable framework for detailed genetic studies of a condition, in recent years its use in research has declined. This is partly because most of the major complex diseases have already been investigated. Another key reason is that segregation analysis is not able to decompose the heterogeneity that is characteristic of complex traits. Segregation analysis necessarily provides a top-down, bird's-eye view of a condition. Given the extreme genetic heterogeneity of most if not all multifactorial diseases, the value of such a view can be questioned. Provided there is evidence that genetic factors are involved somewhere in the etiology, it may be more productive to dive in and use the tools of molecular genetics to hunt for the factors directly, rather than worry about their overall statistical properties. Thus, the focus has moved on, first to linkage analysis and more recently to association studies.

15.3 LINKAGE ANALYSIS OF COMPLEX CHARACTERS

Linkage studies for complex diseases use rather different methods from those described in Chapter 14. Rather than test a detailed hypothesis about recombination fractions, the analysis seeks to identify chromosomal segments shared by affected family members, without having to specify exactly how any susceptibility factors carried on those shared segments contribute to the disease. These methods are very robust for detecting susceptibility factors that are neither necessary nor sufficient for the disease to develop.

Standard lod score analysis is usually inappropriate for non-Mendelian characters

Standard lod score analysis is called **parametric** because it requires a precise genetic model, detailing a series of parameters: the mode of inheritance, gene frequencies, and information about the penetrance of each genotype. As long as a valid model is available, parametric linkage provides a wonderfully powerful method for scanning the genome in 5–10 Mb segments to locate a disease gene. For Mendelian characters, specifying an adequate model should be no great problem. Non-Mendelian conditions, however, are much less tractable. Although complex segregation analysis can provide parameters for the overall genetic susceptibility, the parameters for any one susceptibility factor are impossible to guess in advance. Thus, the type of linkage analysis that was described in Chapter 14 cannot be used for complex diseases.

Near-Mendelian families

One approach to this problem is to look for a subset of families in which the condition segregates in a near-Mendelian manner. Segregation analysis is used to define the parameters of a genetic model in those families, which are then used in a standard (parametric) linkage analysis. Such families may arise in three ways:

- Any complex disease is likely to be heterogeneous, so the family collection may well include some with Mendelian conditions phenotypically indistinguishable from the non-Mendelian majority.
- The near-Mendelian families may represent cases in which, by chance, many determinants of the disease are already present in most people, so that the balance is tipped by the Mendelian segregation of just one of the many susceptibility factors.
- The near-Mendelian pattern may be spurious—just chance aggregations of affected people within one family.

In the first case, identifying the Mendelian subset is intrinsically valuable, but it does not necessarily cast any light on the causes of the non-Mendelian disease. That was the case with breast cancer and Alzheimer disease. In breast cancer this led to the discovery of the *BRCA1* and *BRCA2* genes, as described in Chapter 16 (see Case 7, p. 526). Mutations in the *PSEN1*, *PSEN2*, and *APP* genes cause rare Mendelian forms of early-onset Alzheimer disease. However, the common non-Mendelian forms of breast cancer and Alzheimer disease do not usually involve any of these genes. In the second case, the loci mapped are also susceptibility factors for the common non-Mendelian disease—Hirschsprung disease is an example. Finally, an early study of schizophrenia exemplified the third case, producing a lod score of 6 that is now generally agreed to have been spurious. This debacle was enough to persuade most investigators to switch to alternative methods of analysis.

Non-parametric linkage analysis does not require a genetic model

Model-free or **non-parametric** methods of linkage analysis look for alleles or chromosomal segments that are shared by affected individuals more often than random Mendelian segregation would predict. Some of the basic ideas underlying these approaches were set out in 1990 in the three papers by Risch mentioned previously.

Identity by descent versus identity by state

It is important to distinguish segments *identical by descent* (IBD) from those *identical by state* (IBS). Alleles that are IBD are *demonstrably* copies of the same ancestral (usually parental) allele. IBS alleles look identical, and may indeed be copies, but their common ancestry is not demonstrable. This may be because of a lack of information about any common ancestor, or because the genotypes do not permit unambiguous tracing of the ancestral origin of the alleles. **Figure 15.1** illustrates the difference. In non-parametric analysis, IBD alleles are treated mathematically in terms of the Mendelian probability of inheritance from the defined common ancestor; for IBS alleles, the population frequency is used. For very rare alleles, two independent origins are unlikely, so IBS generally implies IBD. With common alleles, no such inference can be made. Multiallele microsatellites are more efficient than two-allele markers such as SNPs for defining IBD, and multilocus multiallele haplotypes are better still, because any one haplotype is likely to be rare. Shared segment analysis can be conducted with either IBS or IBD data, provided that the appropriate analysis is used. IBD is the more powerful of the two, but it requires samples from more relatives because it is necessary to work out the likelihood that sharing is IBD rather than IBS.

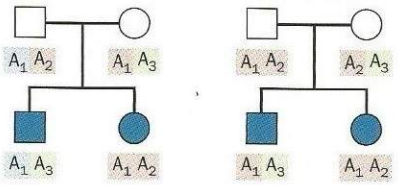

Figure 15.1 Identity by state and identity by descent. Both sib pairs share allele A_1. The first sib pair have two independent copies of A_1 (colored red and blue), indicating identity by state but not by descent. The second sib pair share copies of the same paternal A_1 allele (red), showing identity by descent. The difference is only apparent if the parental genotypes are known.

Affected sib pair analysis

Picking a chromosomal segment at random, pairs of sibs are expected to share 0, 1, or 2 parental haplotypes with frequencies of ¼, ½, and ¼, respectively (ratios of 1:2:1). However, if both sibs are affected by a genetic disease, then any segment of chromosome that carries the disease locus is likely to be shared. For a fully penetrant Mendelian dominant disease, affected sibs would always share the parental haplotype that carried the disease allele; if the disease is recessive they would always share *both* the relevant parental haplotypes (**Figure 15.2**). If a susceptibility factor is neither necessary nor sufficient for disease, then not all affected sib pairs will share the relevant chromosomal segment, but there will still be a statistical tendency to share that segment more often than just by chance. This allows a simple form of linkage analysis. **Affected sib pairs (ASPs)** are typed for markers,

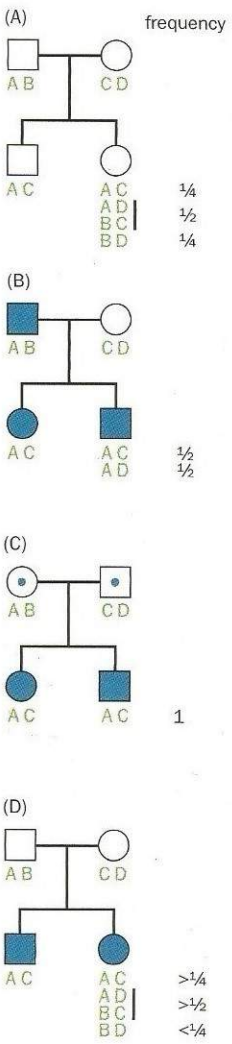

Figure 15.2 Affected sib pair analysis. (A) By random segregation, sib pairs share 2 (both *AC*), 1 (*AC* and either *AD* or *BC*), or 0 (*AC* and *BD*) parental haplotypes ¼, ½, and ¼ of the time, respectively. (B) Pairs of sibs who are both affected by a Mendelian dominant condition must share the segment that carries the disease allele, and they may or may not (a 50:50 chance) share a haplotype from the unaffected parent. (C) Pairs of sibs who are both affected by a Mendelian recessive condition necessarily share the same two parental haplotypes for the relevant chromosomal segment. (D) For complex conditions, haplotype sharing above that expected to occur by chance (as in panel A) identifies chromosomal segments containing susceptibility genes.

and chromosomal regions are sought where the sharing is above the random 1:2:1 ratios of sharing 2, 1, or 0 haplotypes IBD. If the sib pairs are tested only for IBS, the expected sharing on the null hypothesis must be calculated as a function of the gene frequencies.

Shared segment analysis can be performed using any set of affected relatives and without making any assumptions about the genetics of the disease. Affected sib pairs are especially favored because they are relatively easy to collect. Multilocus analysis is preferable to single-locus analysis because it more efficiently extracts the information about IBD sharing across the chromosomal region. The Mapmaker/Sibs computer program is widely used to analyze multipoint ASP data. Programs such as Genehunter extend shared segment analysis to other relationships. The programs calculate the extent to which affected relatives share alleles IBD. The result across all affected pedigree members is compared with the null hypothesis of simple Mendelian segregation (markers will segregate according to Mendelian ratios unless the segregation among affected people is distorted by linkage or association). The comparison can be used to compute a **non-parametric lod (NPL) score**. Theoretical arguments by Lander & Kruglyak suggest genomewide lod score thresholds of 3.6 for IBD testing of affected sib pairs, and 4.0 for IBS testing.

Linkage analysis of complex diseases has several weaknesses

One drawback of ASP analysis is that the candidate regions it identifies are large and likely to contain many genes. Few recombinants separate sibs, so shared parental chromosome segments are large. Identifying the actual disease gene or susceptibility factor in such a large region will be challenging. Crucially, complex disease analysis has no process analogous to the end-game of Mendelian mapping, in which closer and closer markers are tested until there are no more recombinants. Moreover, sib pairs share many segments by chance. Nevertheless, because of its simplicity and robustness, ASP mapping has been one of the main tools for seeking genes conferring susceptibility to common complex diseases.

Any individual susceptibility factor is neither necessary nor sufficient for a person to develop a complex disease. This means that any genetic hypothesis being tested is necessarily much looser than the hypotheses involved in Mendelian linkage analysis. In addition, it is often supposed that many susceptibility factors are of ancient origin and common in the population, unlike the variants that cause Mendelian diseases. This means that, in extended families, even if two affected relatives both have a certain factor, they may have inherited it through two different recent ancestors, maybe in association with different alleles of the marker. As a result of these problems, lod scores are typically modest, and only occasionally reach the threshold of significance.

Significance thresholds

For a genomewide study (whether of a Mendelian character or a complex one), the threshold of significance is a lod score at which the probability of finding a false positive anywhere in the genome is 0.05. In Chapter 14, we noted the distinction between pointwise (or nominal) and genomewide significance:

- The **pointwise *p* value** of a linkage statistic is the probability of exceeding the observed value at a specified position in the genome, assuming the null hypothesis of no linkage at that location.
- The **genomewide *p* value** is the probability that the observed value will be exceeded anywhere in the genome, assuming the null hypothesis of no linkage for each individual location.

Lander & Kruglyak proposed the terminology shown in **Table 15.6** for reporting the strength of linkage data. This is now widely accepted for use in reports. Note that a p value of 10^{-3} is *not* equivalent to a lod score of 3.0—the two measures are not the same:

- A lod score of 3 means that the data are 10^3 times more likely on the given linkage hypothesis than on the null hypothesis. This is a measure of likelihood, or the relative probability of the data on two competing hypotheses.
- A p value of 10^{-3} means that the stated lod score will be exceeded only once in 10^3 times, given the null hypothesis. This is a statement about the probability

TABLE 15.6 CRITERIA FOR REPORTING LINKAGE IN GENOMEWIDE STUDIES

Category of linkage	Expected number of occurrences by chance in a genomewide scan	Range of approximate *p* values	Range of approximate lod scores
Suggestive	1	7×10^{-4} to 3×10^{-5}	2.2–3.5
Significant	0.05	2×10^{-5} to 4×10^{-7}	3.6–5.3
Highly significant	0.001	$\leq 3 \times 10^{-7}$	≥5.4
Confirmed	0.01 in a search of a candidate region that gave significant linkage in a previous independent study		

Criteria suggested by Lander E & Kruglyak L (1995) *Nat. Genet.* 11, 241–247. The figures for *p* values and lod scores are for ASP studies as given by Altmüller J, Palmer LJ, Fischer G et al. (2001) *Am. J. Hum. Genet.* 69, 936–950.

of the data on the given hypothesis. For example, as explained in Chapter 14, in two-point analysis of a Mendelian condition, a lod score of 3 corresponds to an absolute probability of 0.05.

Complex disease studies often use simulation to estimate their significance thresholds. In a typical *permutation test,* 1000 replicates of the family collection are generated by a computer program with random marker genotypes, but based on correct allele frequencies, recombination fractions, and so on. A genomewide search is conducted in each simulated data set and the maximum lod score is noted. The genomewide threshold of significance is set at a score that is exceeded in less than 5% of the replicates. This is taken to be a lod score that has no more than a 5% probability of having arisen purely by chance in the real data set.

Striking lucky

The actual lod score for any particular factor depends on how it happened to segregate in the particular families studied. But there are many susceptibility factors, and a genomewide scan tests for all of them. If there are a dozen or more susceptibility loci, but the studies are only marginally powerful enough to detect the effect of any of them, then there is a large and unavoidable element of chance. In one set of families, a certain factor may happen to segregate in a way that gives a significant lod score, whereas in an independent set of families used in a replication study, that factor may not happen to show such a favorable pattern (**Figure 15.3**). In fact, targeted replication requires a study design with a much higher statistical power than the original study. Thus, failure to replicate the conclusion of a study does not necessarily mean that the original report was a false positive. Nevertheless, for a long time the paucity of well-replicated findings cast a shadow over linkage studies of complex disease.

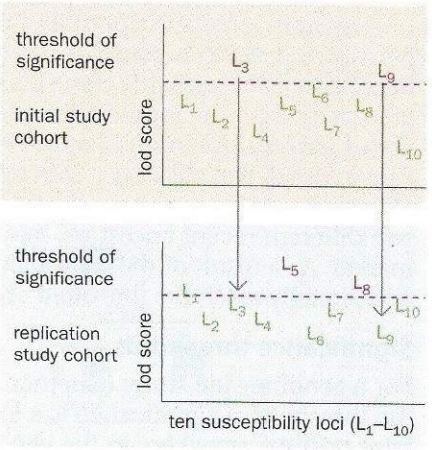

Figure 15.3 The role of chance in detecting weak effects. In a genomewide linkage study of a disease, alleles at each of ten loci (L_1–L_{10}) confer susceptibility to the disease under study, but each has such a weak effect that a linkage study of a given collection of families is not predicted to give a significant lod score. By chance, however, genotypes at two of the loci, L_3 and L_9, happen to give a significant lod score in this collection of families. In a replication study using an independent sample, family members do not happen to have this chance favorable set of genotypes for L_3 and L_9, and therefore the previous significant lod scores are not confirmed. Instead, other loci, L_5 and L_8, happen to have genotypes that give significant lod scores.

An example: linkage analysis in schizophrenia

As an example of these problems, **Table 15.7** gives a rough summary of major findings from the 10 genomewide linkage studies of schizophrenia that formed part of the meta-analysis by Altmüller et al. that was cited in the introduction to this chapter. That paper should be consulted for references to the 10 studies, and the original papers should be consulted for more details—the studies often used sophisticated multi-stage designs and produced large amounts of interesting information. Nevertheless, the conclusion is clear: suggestive linkages found in one study were almost never confirmed in any of the others, and overall there was no evidence that any true susceptibility loci had been identified.

Figure 15.4 shows the results of one of these studies in more detail. Because it is unclear what definition of affected might be the most biologically relevant, these authors analyzed their data using three alternative criteria for affected status (for details see the figure legend). This is an entirely reasonable strategy, but it does introduce extra degrees of freedom into the analysis, and hence requires a more stringent threshold of significance to be used. In this case, no suggestive or significant linkages were found.

TABLE 15.7 POSITIVE FINDINGS FROM 10 GENOMEWIDE LINKAGE STUDIES OF SCHIZOPHRENIA

Study	Sample	No. of individuals genotyped	Significance level	Chromosomal region reported
Coon et al. (1994)	families (mixed)	126[a]	suggestive	4q, 22q
Moises et al. (1995)	families (mixed)	213[a]	none	
Blouin et al. (1998)	families (mixed)	363	significant	13q
			suggestive	8p
Faraone et al. (1998)	families (European American)	146	suggestive	10p
Kaufman et al. (1998)	families (African American)	98	none	
Levinson et al. (1998)	families (mixed)	269	none	
Shaw et al. (1998)	ASPs (European descent)	171[a]	none	
Hovatta et al. (1999)	families (Finnish isolate)	20 families	suggestive	1q
Williams et al. (1999)	ASPs (UK or Irish Caucasian)	196 ASPs	suggestive	4p, 18q, Xcen
Ekelund et al. (2000)	ASPs (Finnish)	134 ASPs	suggestive	1q, 7q

[a]Number of affected individuals. ASPs, affected sib pairs. Significance levels follow the Lander–Kruglyak criteria shown in Table 15.6. These studies formed part of the meta-analysis by Altmüller J, Palmer LJ, Fischer G et al. (2001) *Am. J. Hum. Genet.* 69, 936–950. Full references are given in that paper.

15.4 ASSOCIATION STUDIES AND LINKAGE DISEQUILIBRIUM

The low success rate of linkage studies for complex traits in the 1990s suggested that many, if not most, of the susceptibility factors must be relatively weak, highly heterogeneous, or both. More statistical power would be needed to detect them reliably. Some researchers responded to the challenge by starting studies of much larger samples. However, a seminal paper by Risch & Merikangas in 1996 showed that power to detect weak effects should be achieved more easily by studying associations rather than through linkage analysis. This prompted a general move to association studies. Rather than studying affected relatives, association studies seek populationwide associations between a particular condition and a particular allele or haplotype somewhere in the genome. In this section, we consider how populationwide associations between a specific susceptibility allele and a disease arise, and how they might be detected.

Association is not a specifically genetic phenomenon; it is simply a statistical statement about the co-occurrence of alleles or phenotypes. Allele *A* is associated with disease D if people who have D also have *A* significantly more often (or maybe less often) than would be predicted from the individual frequencies of D and *A* in the population. For example, *HLA-DR4* is found in 36% of the general UK population but in 78% of people with rheumatoid arthritis (RA). Thus *HLA-DR4* is *associated* with RA. The strength of the association is measured by the **relative risk**. This is the amount by which being DR4-positive multiplies the baseline risk of RA. It would be calculated by ascertaining the incidence of RA in *DR4*-positive and *DR4*-negative people. A relative risk of 1 means that being *DR4*-positive confers no additional risk of RA. An alternative measure that is often used is the **odds ratio**. This is explained in Box 19.3 (p. 609). The odds ratio has the advantage that it can be calculated directly from the results of a case-control study, without the need for any information about the population incidence. Again, an odds ratio of 1 means that the factor has no effect on risk.

Associations have many possible causes

A population association can have many possible causes, not all of which are genetic.

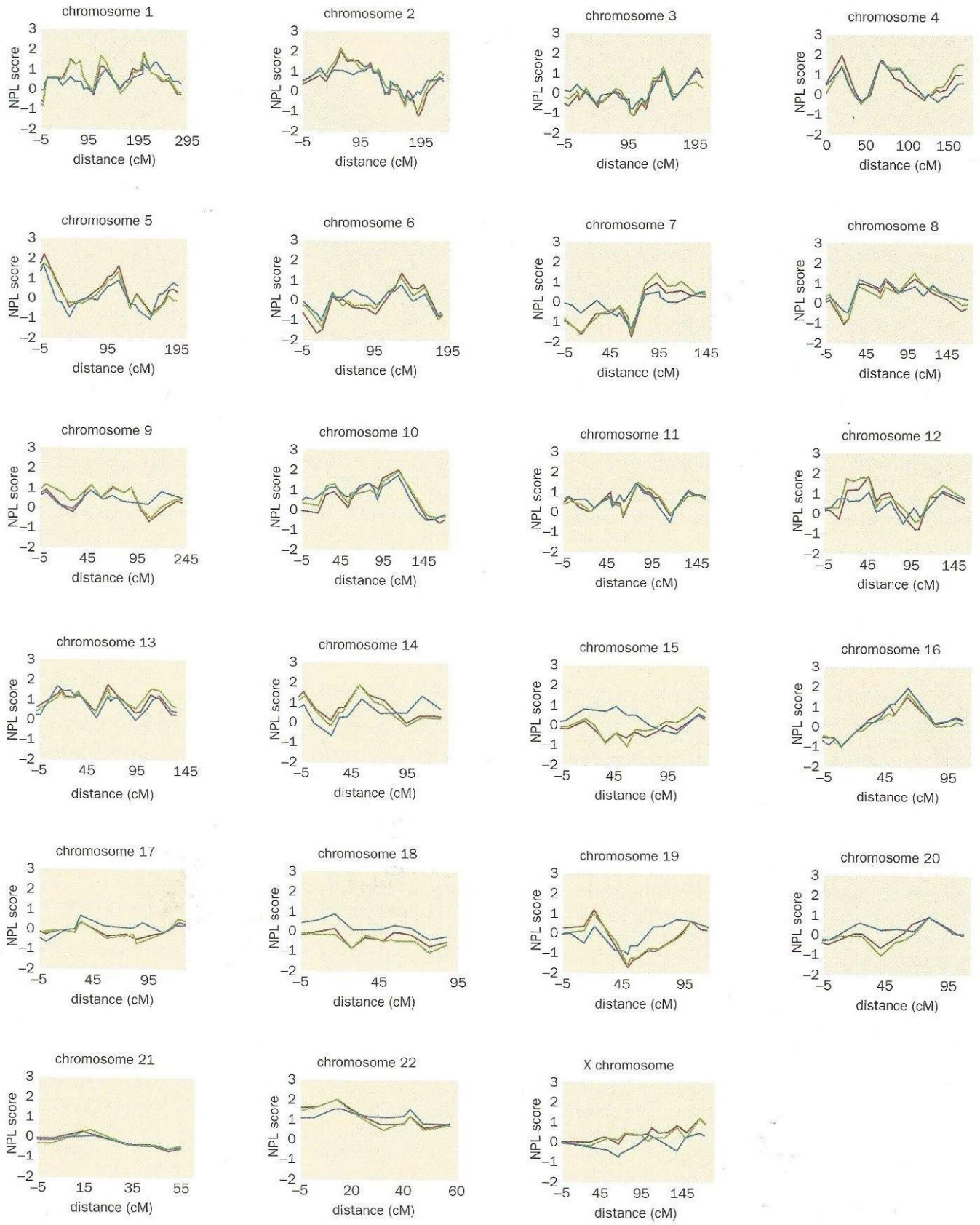

Figure 15.4 Results of a genomewide association study of schizophrenia. A total of 171 affected individuals from 70 multiply affected sibships were genotyped for 338 microsatellite markers spaced across the genome. The blue line shows results when only people with schizophrenia (DSM-III criteria) were counted as affected. For the red line, individuals with DSM schizoaffective disorder were also included. The green line uses a broad definition of affected (schizophrenia, schizoaffective disorder, paranoid or schizotypal personality disorder, delusional disorder, or brief reactive psychosis). In the event, no significant or suggestive lod scores were observed. NPL, nonparametric lod score. [From Shaw SH, Kelly M, Smith AB et al. (1998) *Am. J. Med. Genet. (Neuropsychiatr. Genet.)* 81, 364–378. With permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.]

- **Direct causation:** having allele *A* makes you susceptible to disease D. Possession of *A* may be neither necessary nor sufficient for somebody to develop D, but it increases the likelihood.
- **An epistatic effect:** people who have disease D might be more likely to survive and have children if they also have allele *A*.
- **Population stratification:** the population contains several genetically distinct subsets, and both the disease and allele *A* happen to be particularly frequent in one subset. Lander & Schork give the example of the association in the population of the San Francisco Bay area between carrying the *A1* allele at the HLA locus and being able to eat with chopsticks. *HLA*A1* is more frequent among Chinese than among Caucasians.
- **Type I error:** association studies normally test a large number of markers for association with a disease. Even without any true effect, 5% of results will be significant at the $p = 0.05$ level and 1% at the $p = 0.01$ level. These are false positives, or type I errors. The raw p values need correcting for the number of questions asked. In the past, researchers often applied inadequate corrections. Even after adequate correction, there will remain a certain probability of a false positive result.
- **Linkage disequilibrium (LD):** the disease-associated allele *A* marks an ancestral chromosome segment that carries a sequence variant causing susceptibility to the disease, as described below. Most disease association studies aim to discover associations caused by linkage disequilibrium; it is then an additional step to identify the actual causative sequence variant.

Association is quite distinct from linkage, except where the family and the population merge

In principle, linkage and association are totally different phenomena. Linkage is a relation between *loci* (physical sites on the chromosome), but association is a relation between specific *alleles* and/or *phenotypes*. Linkage is a specifically genetic relationship, whereas association is simply a statistical observation that might have various causes. However, where the family and the population merge, so do linkage and association.

Linkage does not of itself produce any association in the general population. For example, the *STR45* microsatellite locus is linked to the dystrophin locus. Nevertheless, the distribution of *STR45* alleles among a set of unrelated Duchenne muscular dystrophy patients (OMIM 310200), all of whom carry dystrophin mutations, is just the same as in the general population. The mutations arose independently, on a set of chromosomes whose distribution of *STR45* alleles was typical of the general population. However, *within a family* where a particular dystrophin mutation is segregating, we would expect affected people to share the *same* allele of *STR45*, because the loci are tightly linked. Thus, linkage creates associations within families, but not between unrelated people. But how far does a family extend?

All humans are related, if we go back far enough. A rough calculation suggests that, in the UK, two unrelated people would typically have common ancestors not more than 22 generations ago. If fully outbred, they would have $2^{22} = 4$ million ancestors each at that time. Assuming a generational time of 25 years, 22 generations is about 500 years, and in the year 1500 the population of Britain was only a little over 4 million (**Figure 15.5**). It is nevertheless useful to draw a distinction between people who know they are part of one family and those who simply have unknown common ancestors. We will use the word **unrelated** to describe people who do not have a *demonstrable* common ancestor, normally in the past four or so generations.

Association studies depend on linkage disequilibrium

If two supposedly unrelated people with disease D have actually inherited their disease susceptibility from a distant common ancestor, they may well tend also to share particular ancestral alleles at loci closely linked to the susceptibility locus. Thus, in so far as a population is one extended family, population-level associations should exist between alleles of ancestral disease susceptibility genes

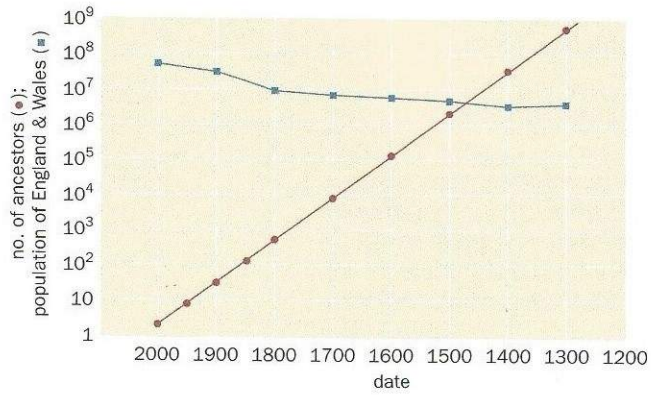

Figure 15.5 Merging into the gene pool. The number of a person's ancestors (red circles) compared with the population of England and Wales (blue squares). This model assumes that there are no consanguineous marriages and that there is a 25-year generation time. On this model two unrelated present-day people would share all their ancestors in 1470. In reality, of course, the population is not fully outbred, and the two people would have strongly overlapping but not identical pools of ancestors.

and closely linked markers. Particular combinations of alleles at closely linked loci occur more often (or less often) than the individual allele frequencies would predict. This phenomenon is called **linkage disequilibrium** (**LD**). Strictly, it would be better called allelic association, but the use of the term LD is firmly established in human genetics.

Linkage disequilibrium can only point to susceptibility factors that have been inherited from ancient common ancestors. A set of unrelated Duchenne muscular dystrophy patients would not be expected to share an ancient disease allele. As mentioned in Chapter 3, there is a rapid turnover of mutant dystrophin alleles, a high rate of loss through natural selection being balanced by a high rate of fresh mutations. Unrelated patients are likely to carry unrelated mutations. Linkage disequilibrium can only cause populationwide disease associations if the disease allele has persisted for many generations. The hope behind disease association studies is that many susceptibility factors are in themselves quite benign, at least in the environments in which our genomes evolved (or alternatively, any deleterious effect in relation to one disease might be balanced by an advantageous effect for a different condition). It is only when they get into combination with other factors that they cause susceptibility. Moreover, many common diseases affect mainly older people well past childbearing age, by which time natural selection has far less relevance for the survival of their genes.

The size of shared ancestral chromosome segments

Suppose that two unrelated people each inherit a disease susceptibility allele from their common ancestor. During the many generations and many meioses that separate them from their common ancestor, repeated recombination will have reduced the shared chromosomal segment to a very small region (**Figure 15.6**). Only alleles at loci tightly linked to the disease susceptibility locus will still be shared. For a marker showing recombination fraction θ with the shared disease locus, a proportion θ of ancestral chromosomes will lose the association each generation, and a proportion $(1 - \theta)$ will retain it. After n meioses, a fraction $(1 - \theta)^n$ of chromosomes will retain the association. The half-life of LD between loci 1 cM ($\theta = 0.01$) and 2 cM ($\theta = 0.02$) apart is 69 and 34 meioses respectively, since $(0.99)^{69} \approx (0.98)^{34} \approx 0.5$.

We calculated above that the ancestry of two unrelated British people merges completely 22 generations back. That calculation was grossly simplified because it assumed that the entire British population has been one freely interbreeding unit over the past 500 years. However, it provides a first crude estimate that the ancestral segments shared between two unrelated people from a large population might be of the order of a few megabases long (using the rule of thumb that 1cM = 1 Mb). For a population association, we require ancestral segments that are shared not just by two particular individuals but by a significant proportion of all descendants of that ancestor. The locations of crossovers will be different in each lineage leading down from the common ancestor. Thus, segments that are shared by enough people to produce population associations will, on average, be much shorter than the segments shared by any two particular individuals (in reality this will be a few kilobases—see Figure 15.6). More sophisticated calculations use a Poisson distribution of recombination events and incorporate

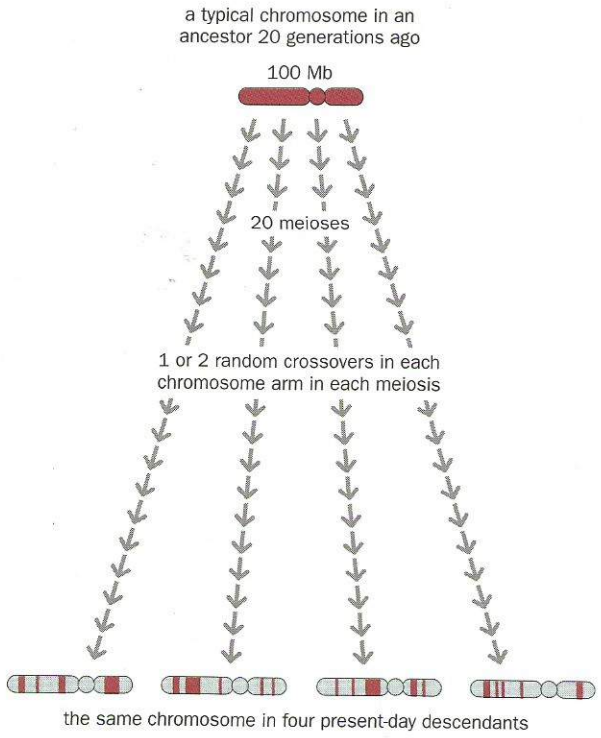

Figure 15.6 The size of shared ancestral chromosome segments. A typical chromosome is shown in a common ancestor, 20 generations ago, of four present-day individuals. There will be one or two random crossovers in each chromosome arm in each of the 20 meioses linking each present-day person to their common ancestor. Only a small proportion of the sequence of the ancestor's chromosome will be inherited by descendants after 20 generations (red segments). The ancestral segments that are shared by a significant proportion of all descendants are very small, typically 5–15 kb.

assumptions about population structure and history to estimate the size distribution of shared ancestral segments. However, the wide stochastic variance and reliance on unknowable details of population history make even the most elaborate calculations unreliable. What we need is data, and recently increasing quantities of real data have become available.

Studying linkage disequilibrium

LD is a statistic about populations, not individuals. To study it, a sample of individuals from the chosen population is genotyped for a series of linked polymorphic markers. SNPs are the markers of choice, for three reasons:

- They are sufficiently abundant that they can be used to check very short chromosome segments for disequilibrium.
- In comparison with microsatellites, they have a far lower rate of mutation. This is important when the aim is to identify ancient conserved chromosomal segments.
- SNPs are easy to genotype on a large scale, up to 1 million at once, spread across the genome. This is very hard with other polymorphisms.

Unless genotyping is done on individual chromosomes isolated by laboratory manipulation (an expensive option), the raw data will consist of the genotypes at each individual locus, rather than haplotypes of alleles across multiple linked loci. The raw genotype data must be *phased*—that is, converted into haplotypes—before any useful analysis can be done. This can be done by genotyping other family members to infer haplotypes; alternatively, computer programs can be used to convert genotypes into haplotypes. These programs use a maximum-likelihood procedure: they guess possible haplotypes, and keep trying until they find the guess that best explains the whole data set with the minimum number of plausibly related haplotypes (haplotypes within a population should be related to one another by a minimum number of recombinations).

Various statistics have been used as measures of LD (**Box 15.1**).

The HapMap project is the definitive study of linkage disequilibrium across the human genome

Early studies of specific chromosome regions showed that LD is common but unpredictable. The data were in striking contradiction to the naive expectation that LD between any two loci would fall off as a smooth function of their physical

BOX 15.1 MEASURES OF LINKAGE DISEQUILIBRIUM

If two loci have alleles A,a and B,b with frequencies p_A, p_a, p_B, and p_b, there are four possible haplotypes: AB, Ab, aB, and ab. Suppose that the frequencies of the four haplotypes are p_{AB}, p_{Ab}, p_{aB}, and p_{ab}. If there is no LD, $p_{AB} = p_A p_B$ and so on, because the haplotype will be constructed just by random assortment of the constituent alleles. The degree of departure, D, from this random association can be measured by $D = p_{AB}p_{ab} - p_{Ab}p_{aB}$. As a measure of LD, D suffers from the property that its maximum absolute value depends on the gene frequencies at the two loci, as well as on the extent of disequilibrium. Among preferred measures are:

- $D' = (p_{AB} - p_A p_B)/D_{max}$, where D_{max} is the maximum value of $|p_{AB} - p_A p_B|$ possible with the given allele frequencies; the vertical lines indicate the absolute value or modulus of the expression.
- $\Delta^2 = (p_{AB} - p_A p_B)^2/(p_A p_a p_B p_b)$.

D' is the most widely used. It varies between 0 (no LD) and ±1 (complete association) and is less dependent than D on the allele frequencies. As a rule of thumb, $D' > 0.33$ is often taken as the threshold level of LD above which associations will be apparent in the usual size of data set. The proliferation of alternative measures suggests that none is ideal.

distance apart. Closely spaced SNPs often showed little or no LD, whereas sometimes there was strong LD between more widely separated SNPs. **Box 15.2** shows how the patterns of disequilibrium can be represented graphically, and also illustrates just how complex and irregular the patterns can be. These irregularities reflect the combined effects of several factors:

- Recombination, the force that breaks up ancestral segments, occurs mainly at a limited number of discrete hotspots (see Chapter 14). SNPs that are close together but separated by a hotspot show little or no LD, whereas, conversely, LD may be strong across even a large chromosomal region if it is devoid of hotspots.
- Gene conversion (see Box 14.1) may replace a small internal part of a conserved segment, producing a localized breakdown of LD, whereas markers either side of the replaced segment continue to show LD with each other.
- Population history is important. The older a population is, the shorter are the conserved segments. LD is more extensive and of longer range in populations derived from recent founders, as often occurs with small, genetically isolated populations. LD will have a shorter range in populations that have remained constant in size than in populations that have undergone a recent expansion. Superimposed on this regularity are many stochastic effects. Chromosome segments may carry a mixture of marker alleles that are IBD and alleles that are only IBS (through independent mutations, back mutation, and so on), but LD statistics do not distinguish between these two cases.

Efficient disease association studies depend on a detailed knowledge of the patterns of LD across the genome. It is important to know how big and how diverse the ancestral chromosome segments are, so that SNPs can be chosen to test each conserved ancestral segment. The International HapMap Project was established to provide this detailed knowledge. A consortium of academic institutions and pharmaceutical companies typed several million SNPs in 269 individuals drawn from four human populations: 30 white American parent–child trios from Utah (CEU), 30 Yoruba parent–child trios from Ibadan, Nigeria (YRI), 45 individual Han Chinese from Beijing (CHB), and 44 individual Japanese from Tokyo (JPT). All were ostensibly healthy and, although not a formal population sample, were hopefully typical of the population from which they were drawn. The Phase I report in 2005 (see Further Reading) summarized results from 1,007,329 SNPs; Phase II has added a further 3 million. The same individuals have been used in several other studies, for example in surveys of copy-number variants and of gene expression levels, adding value and depth to the HapMap data.

The CEU and YRI samples, which each consisted of parent–child trios, could be phased directly. For each trio, the genotype of the child could be used to infer phase in the parental genotypes. Thus, each set of 30 trios yielded 120 phased parental haplotypes. For the CHB and JPT samples, computer programs were used to convert genotypes into haplotypes.

The key findings can be summarized as follows (**Table 15.8**):

- All four populations show a similar structure of blocks of strong linkage disequilibrium, with little or no LD between markers in adjacent blocks. Block boundaries are generally at similar positions across the four populations.

BOX 15.2 GRAPHICAL REPRESENTATIONS OF LINKAGE DISEQUILIBRIUM

In HapMap and similar studies, hundreds of SNPs are genotyped across a fairly small chromosomal region, and the extent of linkage disequilibrium (LD) is calculated for every pair of markers. To help the human mind grasp patterns in the data, the results are usually displayed graphically. The markers, in chromosomal order, are set out on both the horizontal and vertical axes of a grid. Each axis may be a physical map, with markers spaced according to their actual physical distance apart, or markers may simply be listed in order along each axis. The linkage disequilibrium between each pair of markers is indicated by color coding of the point on the graph at which the vertical and horizontal coordinates of the two markers intersect. **Figure 1** and **Figure 2** show two examples.

In Figure 1 a physical map has been used, and the computer has interpolated colors to produce a continuous distribution of colors. The distribution is necessarily symmetrical about the diagonal because coordinate (*m*,*n*) contains the same data as coordinate (*n*,*m*). In Figure 2 this redundancy is removed by showing only one half of the square. Tilted over so that what was the diagonal is now the horizontal axis, this axis can be envisaged as a map of the chromosome, with the colored triangles giving an impression of the size of haplotype blocks. In this case, the actual physical distance between markers is not represented; the axes simply list them in order.

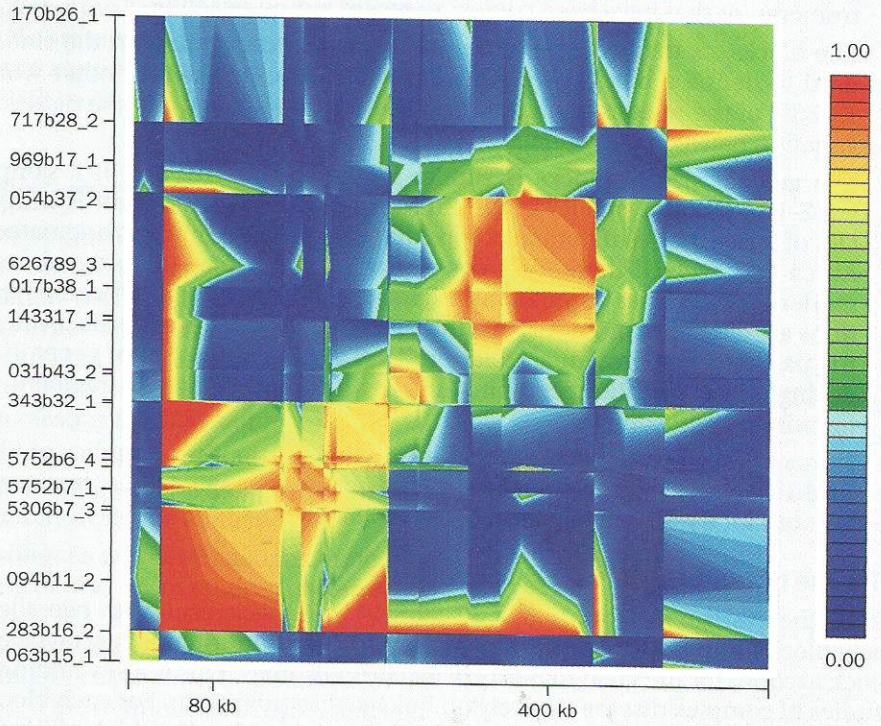

Figure 1 The pattern of linkage disequilibrium across a 500 kb section of chromosome 13, as displayed by the GOLD program. (Courtesy of William Cookson, University of Oxford.)

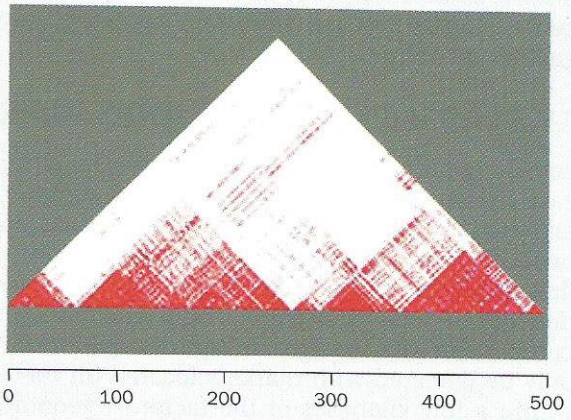

Figure 2 The pattern of linkage disequilibrium at 7q31.33 in the HapMap CEU sample. Key: red, $D' = 1$, lod score ≥ 2; pink, $D' < 1$, lod score ≥ 2; blue, $D' = 1$, lod score < 2; white, $D' < 1$, lod score < 2. [From The International HapMap Consortium (2005) *Nature* 437, 1299–1320. With permission from Macmillan Publishers Ltd.]

TABLE 15.8 HAPLOTYPE BLOCK STRUCTURES IN FOUR HUMAN POPULATIONS AS REPORTED BY PHASE I OF THE HAPMAP PROJECT

Parameter	YRI	CEU	CHB+JPT
Average number of SNPs per block	30.3	70.1	54.4
Average length per block (kb)	7.3	16.3	13.2
Percentage of genome spanned by blocks	67	87	81
Average number of haplotypes per block	5.57	4.66	4.01
Percentage of chromosomes accounted for by these haplotypes	94	93	95

Only haplotypes having a frequency of 0.05 or greater in the relevant population are reported here. The CHB and JPT samples have been combined because they show very similar patterns. A different statistical method of defining blocks gave somewhat different detailed figures but a similar overall pattern. YRI, Yoruba from Ibadan, Nigeria; CEU, white Americans from Utah; CHB, Han Chinese from Beijing; JPT, Japanese from Tokyo. [Data from The International HapMap Consortium (2005) *Nature* 437, 1299–1320.]

- Blocks vary in size, but are typically 5–15 kb. Much longer blocks can be found in chromosomal regions where there is very little recombination, such as centromeres, or that have been subject to recent strong selection.
- The blocks, as defined here, cover only 67–87% (depending on the statistic used to define a block) of the genomic regions examined; in other words, 13–33% of the regions analyzed show a more chaotic and diverse pattern of variation between individuals.
- The general size of blocks is similar in the CEU, CHB, and JPT samples (13.2–16.3 kb), but smaller in the YRI samples (7.3 kb). This accords with the Out of Africa model of human origins that suggests that humans originated in Africa and evolved there for a considerable time before the progenitors of modern non-African populations migrated out. Sub-Saharan African populations are older and more variable than all others. Their founders lie further in the past, and the present structure reflects a greater number of generations during which meiotic recombination has had the chance to fragment the founder chromosomes.
- Human genetic variability is much more limited than the number of SNPs might suggest. In each population studied, on average 4.0–5.6 haplotypes account for 93–95% of copies of any given block.

The use of tag-SNPs

Given the average in Table 15.8 of 30–70 SNPs per block, each with two alleles, each block has 2^{30}–2^{70} possible haplotypes. In reality, four to six haplotypes per block account for the great majority of all chromosomes. This is a key finding for studies of complex disease that rely on linkage disequilibrium. For each block, a small number of SNPs (**tag-SNPs**) can be defined that identify which of the four to six common alternative blocks a chromosome carries (**Figure 15.7**). Thus, a disease association study no longer needs to genotype all 30–70 SNPs in a block; genotyping as few as two to four tag-SNPs may serve to capture most of the genetic variability.

15.5 ASSOCIATION STUDIES IN PRACTICE

Searching for population associations is an attractive option for identifying disease susceptibility genes. Association studies are easier to conduct than linkage analysis, because neither multi-case families nor special family structures are needed. Under some circumstances, association can also be more powerful than linkage for detecting weak susceptibility alleles (see below). However, association depends on linkage disequilibrium, which is a very short-range phenomenon in comparison with linkage. Linkage disequilibrium with a susceptibility factor can only be detected with markers located on the same haplotype block—that is, within a few kilobases of the factor. A genomewide association study would therefore require samples to be genotyped for hundreds of thousands of markers (in comparison with the few hundred needed for a genomewide linkage scan).

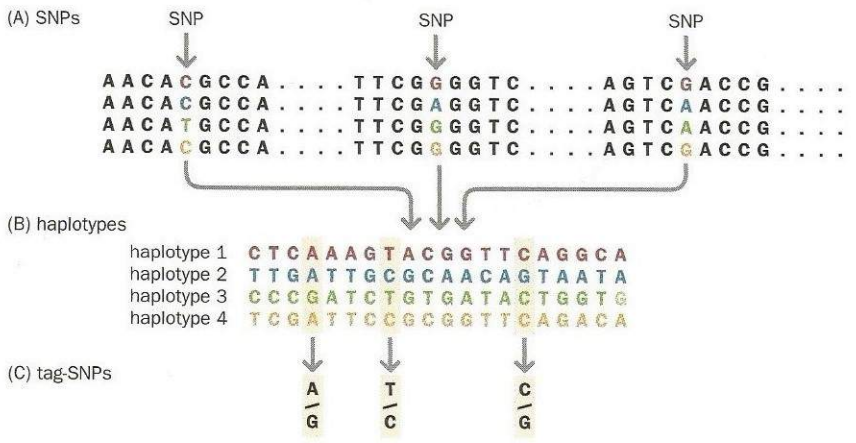

Figure 15.7 SNPs, haplotypes, and tag-SNPs. (A) A short segment of four individual copies of the same chromosome shows three biallelic SNPs. (B) Haplotypes from a larger region on these four chromosomes containing 20 SNPs, showing which allele of each SNP the chromosome carries. Although there are 2^{20} possible combinations of 20 biallelic SNPs, a population survey shows that most copies of this chromosome have one of these four haplotypes. (C) Genotyping just three of the 20 SNPs serves to identify each of these four haplotypes. [From The International HapMap Consortium (2003) *Nature* 426, 789–796. With permission from Macmillan Publishers Ltd.]

Until recently, association studies had to be focused on small candidate chromosomal regions, and reported associations were seldom replicated in follow-up studies. New technical developments, and the availability from the HapMap project of a genomewide catalog of ancestral chromosome segments, have led to a new generation of genomewide association studies that are at last producing robust, replicable results. The background to these studies is discussed in the rest of this section.

Early studies suffered from several systematic weaknesses

Disease association studies in human genetics have a very long but distinctly checkered history. In the 1960s, long before the identification of common DNA variants made human linkage analysis generally useful, many studies looked for associations between particular HLA tissue types and various diseases. Mostly these were autoimmune diseases, in which an HLA effect was *a priori* plausible, but many other diseases were studied, and many associations were reported. Few were replicated. There were three main reasons for this poor record:

- **Inadequate matching of controls.** These were all case-control studies, and often insufficient attention was paid to matching cases and controls. This is in fact a major worry, even in the most carefully designed studies.
- **Insufficient correction for multiple testing.** As we mentioned in connection with genomewide linkage scans, in any set of tests 5% of random observations will be significant at the $p = 0.05$ level and 1% at the $p = 0.01$ level. Each time the data are checked for another possible association, there is another chance of a type I error. The overall threshold of significance must be adjusted for the number of independent questions asked. If the frequencies in cases and controls of alleles at three HLA loci (HLA-A, HLA-B, and HLA-DR) are compared, it is not sufficient to say that this represents three independent questions. Each allele that was checked was a separate, although not fully independent, question. A full (Bonferroni) correction divides the threshold p value by N, the total number of questions asked. To maintain an overall 5% chance of a false positive result, the threshold p value for a single question is 0.05, for 10 questions 0.005, and for 1,000,000 questions (typical of studies using high-density SNP arrays) 5×10^{-8}. For large values of N this is overconservative, and the preferred threshold of significance is $p' = 1 - (1 - p)^N$. The report from the Wellcome Trust Case-Control Consortium (see Further Reading) has a good discussion of this problem.
- **Striking lucky in underpowered studies.** We noted that chance is a factor causing low rates of replication in linkage studies of complex disease (see Figure 15.3). The same applies to association studies. An underpowered study may occasionally get lucky, but this luck is unlikely to be repeated in a similarly powered study. Targeted replication studies need much more power than the initial trawl.

The transmission disequilibrium test avoids the problem of matching controls

In any association study, the choice of the control group is crucial. One way of avoiding the matching problem altogether is to use internal controls—that is, the control data come from the same people as the case data. The **transmission disequilibrium test** (**TDT**) implements this idea. The TDT starts with couples who have one or more affected offspring. It is irrelevant whether or not either parent is affected. Suppose we wish to test the hypothesis that allele I of marker M is associated with the disease. We identify cases in which a parent is heterozygous for I and any other allele of the marker (denoted X—**Box 15.3**). I is probably not necessary for the disease to develop, so the affected offspring will not necessarily have inherited that allele; however, if I has any role in susceptibility, we would expect the affected children in this cohort to have inherited I more often than the parent's other allele. We therefore proceed as follows:

- Affected probands are ascertained and DNA is obtained from the proband and both parents.

BOX 15.3 THE TRANSMISSION DISEQUILIBRIUM TEST

The transmission disequilibrium test (TDT) statistic, which is based on the standard χ^2 statistic, is

$$(a - b)^2/(a + b)$$

where a is the number of times that a heterozygous parent transmits *I* to the affected offspring, and b is the number of times that the other allele is transmitted. The mathematically inclined can find the justification for this procedure in the paper by Spielman, McGinnis, and Ewens (see Further Reading).

Figure 1 shows the transmission disequilibrium test applied to type 1 diabetes. Ninety-four families were investigated for an association between type 1 diabetes and a particular allele at a repetitive sequence upstream of the insulin gene. Among the 94 families were 57 parents who were heterozygous for the allele under investigation (denoted by *I*) and some other allele (denoted collectively by *X*). These 57 parents transmitted 124 alleles to diabetic offspring (some had more than one affected child). Of these, 78 were allele *I* and 46 allele *X*.

The TDT statistic had a value of $(78 - 46)^2/(78 + 46) = 32^2/124 = 8.26$, corresponding to $p = 0.004$. Thus, the data demonstrated an association between the *I* allele and type 1 diabetes.

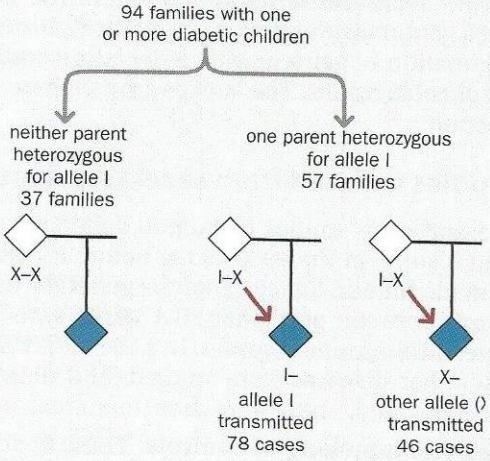

Figure 1 An example of TDT data. Only one parent is shown in each case. [Data reported in Spielman RS, McGinnis RE & Ewens WJ (1993) *Am. J. Hum. Genet.* 52, 506–516. With permission from Elsevier.]

- The probands and their parents are genotyped for marker M.
- Only data from those parents who are heterozygous for marker allele *I* are considered. It does not matter what their other allele is, provided it is not *I*.

Box 15.3 shows the TDT statistic applied to data from families with one or more diabetic children.

The result is unaffected by population stratification because the non-inherited parental allele serves as an internal control. An extended TDT test (ETDT) has been developed to handle data from multiallelic markers such as microsatellites. The TDT can be used when only one parent is available, but this may bias the result. When there is no parent available (a common problem with late-onset diseases) an alternative variant, sib-TDT, looks at differences in marker allele frequencies between affected and unaffected sibs.

There has been some argument about whether the TDT is a test of linkage or association. Because it asks questions about alleles and not loci, it is fundamentally a test of association—but association in the presence of linkage. The associated allele may itself be a susceptibility factor, or it may be in linkage disequilibrium with a susceptibility allele at a nearby locus. The TDT cannot detect linkage if there is no linkage disequilibrium—this is a point to remember when considering schemes to use the TDT for whole-genome scans.

Association can be more powerful than linkage studies for detecting weak susceptibility alleles

In 1996, Risch & Merikangas (see Further Reading) compared the power of linkage (affected sib pair, or ASP) and association (TDT) testing to identify a marker tightly linked to a disease susceptibility locus. They calculated the number of ASPs or TDT trios (affected child and both parents) required to distinguish a genetic effect from the null hypothesis, with a given power and significance level. **Box 15.4** illustrates their method (consult the original paper for more detail), and Table 1 in Box 15.4 shows typical results of applying their formulae. The conclusion is clear. ASP analysis would require unfeasibly large samples to detect susceptibility loci that confer a relative risk of less than about 3, whereas TDT might detect alleles giving a relative risk below 2 with manageable sample sizes. Note, however, that the calculation incorporates various assumptions. In particular, it assumes a single ancestral susceptibility allele at the disease locus. Any allelic heterogeneity would rapidly degrade the performance of an association test such as TDT, while not affecting the power of a linkage test.

Although the specific calculations used by Risch & Merikangas applied to the TDT, the implication is more general: association tests are more powerful than linkage tests to detect weak susceptibility factors. Moreover, it is easier to collect

1000 isolated cases and their parents than it is to collect 1000 affected sib pairs, at least for early-onset diseases, for which both parents are likely to be still alive. The paper helped trigger a widespread move away from linkage studies and toward studies of association.

Case-control designs are a feasible alternative to the TDT for association studies

As an alternative to TDT, conventional case-control studies are now coming back into favor. Case-control studies have advantages, provided that the matching of cases and controls is not problematic. They need 50% fewer samples than TDT (two rather than three per comparison) and are more feasible for late-onset diseases, for which parents are seldom available.

How closely the controls need to be matched depends on the population. In the UK, the large Wellcome Trust Case-Control Consortium study excluded individuals who reported non-European or non-Caucasian ancestry, but having done that, dividing the UK population into 12 geographic regions identified only a small number of chromosomal locations where gene frequencies showed significant differences between the 12 regions. By comparing allele frequencies at a range of unlinked loci in the cases and controls, the data can be checked for stratification. However, caution must remain: a recent study by Campbell and colleagues (see Further Reading) showed how subtle population stratification explained an association between tall stature and persistence of intestinal lactase (which allows adults to digest milk) among European Americans. The frequency

BOX 15.4 SAMPLE SIZES NEEDED TO FIND A DISEASE SUSCEPTIBILITY LOCUS BY A GENOMEWIDE SCAN

Risch & Merikangas (1996) calculated the sample sizes needed to find a disease susceptibility locus by a genomewide scan by using either affected sib pairs (ASPs) or the transmission disequilibrium test (TDT). This Box summarizes their formulae and equations, but the original paper (see Further Reading) should be consulted for the derivations and for details.

A standard piece of statistics tells us that the sample size M required to distinguish a genetic effect from the null hypothesis with power $(1 - \beta)$ and significance level α is given by $(Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2$, where Z refers to the standard normal deviate. The mean μ and variance σ^2 are calculated as functions of the susceptibility allele frequency (p) and the relative risk γ conferred by one copy of the susceptibility allele. The model assumes that the relative risk for a person carrying two susceptibility alleles is γ^2, that the marker used is always informative, and that there is no recombination with the susceptibility locus.

For ASP, the expected allele sharing at the susceptibility locus is given by $Y = (1 + w)/(2 + w)$, where $w = [pq(\gamma - 1)^2]/(p\gamma + q)$. $\mu = 2Y - 1$ and $\sigma^2 = 4Y(1 - Y)$. The genomewide threshold of significance (probability 0.05 of a false positive anywhere in the genome; testing for sharing IBD) requires a lod score of 3.6, corresponding to $\alpha = 3 \times 10^{-5}$, and $Z_\alpha = 4.014$. For 80% power to detect an effect, $1 - \beta = 0.2$ and $Z_{1-\beta} = -0.84$.

For the TDT, the probability that a parent will be heterozygous for the allele in question is $h = pq(\gamma + 1)/(p\gamma + q)$. $P(\mathrm{tr}A)$, the probability that such a heterozygous parent will transmit the high-risk allele to the affected child, is $\gamma/(1 + \gamma)$. $\mu = \sqrt{h}(\gamma - 1)/(\gamma + 1)$, and $\sigma^2 = 1 - [h(\gamma - 1)^2/(\gamma + 1)^2]$. As discussed above, for a genomewide screen involving 1,000,000 tests, $\alpha = 5 \times 10^{-8}$, $Z_\alpha = 5.33$, and, as before, $Z_{1-\beta} = -0.84$.

In Table 1, the Z_α, $Z_{1-\beta}$, μ, and σ^2 values are used to calculate sample sizes by substitution in the formula $M = (Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2$. For the TDT, the answer is halved because each parent–child trio allows two tests, one on each parent.

Table 1 Comparison of the power of linkage and association studies

Susceptibility factor		ASP analysis		TDT analysis	
γ	p	Y	No. of pairs	$P(\mathrm{tr}A)$	No. of trios
5	0.01	0.534	2530	0.830	747
	0.1	0.634	161	0.830	108
	0.5	0.591	355	0.830	83
3	0.01	0.509	33,797	0.750	1960
	0.1	0.556	953	0.750	251
	0.5	0.556	953	0.750	150
2	0.1	0.518	9167	0.667	696
	0.5	0.526	4254	0.667	340
1.5	0.1	0.505	115,537	0.600	2219
	0.5	0.510	30,660	0.600	950
1.2	0.1	0.501	3,951,997	0.545	11,868
	0.5	0.502	696,099	0.545	4606

The table shows the number of affected sib pairs (ASPs) or parent–child trios needed for 80% power to detect significant linkage or association in a genomewide search. γ is the relative risk for individuals of genotype *Aa* compared with *aa*; p is the frequency of the susceptibility allele, *A*. For affected sib pair analysis, Y is the expected allele sharing; for the trios analyzed by the transmission disequilibrium test, $P(\mathrm{tr}A)$ is the probability that a parent heterozygous for allele *A* will transmit *A* to an affected child. For low values of γ, unfeasibly large numbers of affected sib pairs are needed to detect an effect. [Data from Risch N & Merikangas K (1996) *Science* 273, 1516–1517.]

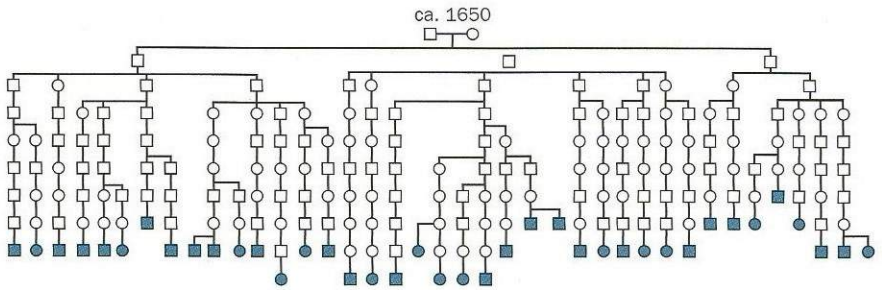

Figure 15.8 Patterns of recent common ancestry in an isolated Finnish population. The people in the youngest generation were recruited because they each had multiple offspring affected by schizophrenia. Note, however, that the ancestry shown is highly selective; multiple ancestors have been removed from each generation for clarity. Ten generations ago, in 1650, an outbred person has 1024 ancestors, but only two are shown here. [From Hovatta I, Varilo T, Suvisaari J et al. (1999) *Am. J. Hum. Genet.* 65, 1114–1124. With permission from Elsevier.]

of each of these two characters varies considerably between populations, and rematching of individuals on the basis of European ancestry greatly decreased the association.

Special populations can offer advantages in association studies

Populations derived from a small number of relatively recent founders are expected to show limited haplotype diversity and more extensive linkage disequilibrium than older populations. The HapMap data summarized in Table 15.8 illustrate this. In comparison with the Yoruba subjects from Nigeria, the non-African study subjects come from populations that were established more recently, and on average their haplotype blocks are longer and less diverse. Ancestral disease-bearing haplotypes should be easier to identify in populations that derive from a small number of founders and have remained relatively isolated since that time. This belief lies behind the DeCode project in Iceland, and similar projects in Quebec and elsewhere. **Figure 15.8** shows an example from an isolated Finnish population. The 39 shaded individuals each have two or more offspring affected by schizophrenia, and were ascertained as part of the study by Hovatta et al. listed in Table 15.7. All could be traced back to at least one common ancestral couple 7–10 generations back. Because the common ancestors are relatively recent (dating to about 1650), any segments of their chromosomes that are shared by several of the 39 shaded individuals are likely to be quite large in comparison with ancestral segments in an older population.

Populations derived by recent admixture are a second group that might offer special advantages. For example, consistent long-range linkage disequilibrium has been documented in the Lemba, a Bantu–Semitic hybrid population in Africa. If the two source populations had widely different incidences of a common disease, the mixed population could be used to map the determinants rather efficiently. This is the human analog of an interspecific mouse cross, although of course all humans are the same species. In reality, the availability of large numbers of potential subjects with good medical records may be more important than population structure—certainly, that is the thinking behind the BioBank project in the UK, which seeks to collect medical and lifestyle data and DNA from 500,000 British people aged 45–69 years and follow their health prospectively.

A new generation of genomewide association (GWA) studies has finally broken the logjam in complex disease research

It would be wrong to paint a wholly negative picture of the search for complex disease susceptibility factors over the decade up to 2005. Although the frequency of successful replication has been low, it has not been negligible. In 2003, Lohmueller and colleagues made a meta-analysis of 301 publications on 25 frequently studied associations in 11 different diseases. They concluded that at least 8 of the 25 associations had been adequately replicated. **Table 15.9** lists some associations that were well established before the current generation of GWA studies.

It remains true that the successes listed in Table 15.9 have been hard won. The cost, in money and person-hours, of identifying and thoroughly validating a common disease association has been orders of magnitude higher than the cost of identifying a Mendelian disease locus. But at least we have now put the history of irreproducible results behind us. Several key developments have made this possible:

TABLE 15.9 SOME CONFIRMED DISEASE ASSOCIATIONS

Disease	Gene containing associated variant	Odds ratio	Frequency of risk allele or haplotype
Type 1 diabetes	*IF1H1*	1.17[a]	0.65
	CTLA4	1.31[b]	0.53
Type 2 diabetes	*ABCC8*	2.23[b]	0.02
	PPARG	1.21[b]	0.85
	SLC2A1	1.82[b]	0.30
	TCF7L2	1.49[a]	0.26
Age-related macular degeneration	*CFH*	2.45[a]	0.46
Systemic lupus erythematosus	*IRF5*	1.78[a]	0.12
Schizophrenia	*DRD3*	1.13[b]	0.01
	HTR2A	1.07[b]	0.39

Different studies of the same disease and candidate gene may report associations with different markers, or with multilocus haplotypes. [a]These odds ratios and all allele or haplotype frequencies are representative examples of data from one out of several reports on each association. [b]These odds ratios are from the meta-analysis by Lohmueller KE, Pearce CL, Pike M et al. (2003) *Nat. Genet.* 33, 177–182.

- Consortia have been established to perform case-control studies with 1000 or more subjects in each arm—and funding bodies have recognized that work on this scale is necessary. In the USA a private–public partnership, the Genetic Association Information partnership (GAIN), and in the UK the Wellcome Trust Case-Control Consortium, are conducting or coordinating GWA studies on large cohorts of patients with several different complex diseases. Similar efforts are underway in other countries.
- Efficient whole-genome amplification methods (using ϕ29 DNA polymerase) have been developed that allow very extensive analysis of even small DNA samples.
- Massively parallel microarray or bead-based methods are available, by which a sample can be genotyped for 500,000 or more SNPs in parallel.
- The HapMap project has provided data that permit a rational choice of tag-SNPs to capture a significant proportion of all the genetic variation in a population.

A good flavor of the new clutch of studies is provided by the report from the Wellcome Trust Case-Control Consortium (WTCCC), published in June 2007. Large consortia of British researchers assembled 2000 well-phenotyped cases of each of seven diseases: bipolar disorder (manic-depressive psychosis), coronary artery disease, Crohn disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. In addition, 3000 healthy controls were collected. All samples were typed for more than 500,000 SNPs. The estimated power of the study to detect risk factors was 43% for a factor conferring a relative risk of 1.3, and 80% for a relative risk of 1.5 (risks are expressed on a per-allele basis using a multiplicative model—that is, if allele *A* confers a relative risk of x, the risks for genotypes *aa*, *Aa*, and *AA* are $1{:}x{:}x^2$).

Across the seven patient groups, 25 independent disease-association signals were identified with p values beyond the threshold of significance, calculated as $p < 5 \times 10^{-7}$ (**Figure 15.9**). The key question is: How many of these could be replicated? Previous studies in the seven diseases were considered to have identified 15 variants with strong, replicated evidence of association with one or other of the seven diseases. Thirteen of the 15 were unambiguously identified in the

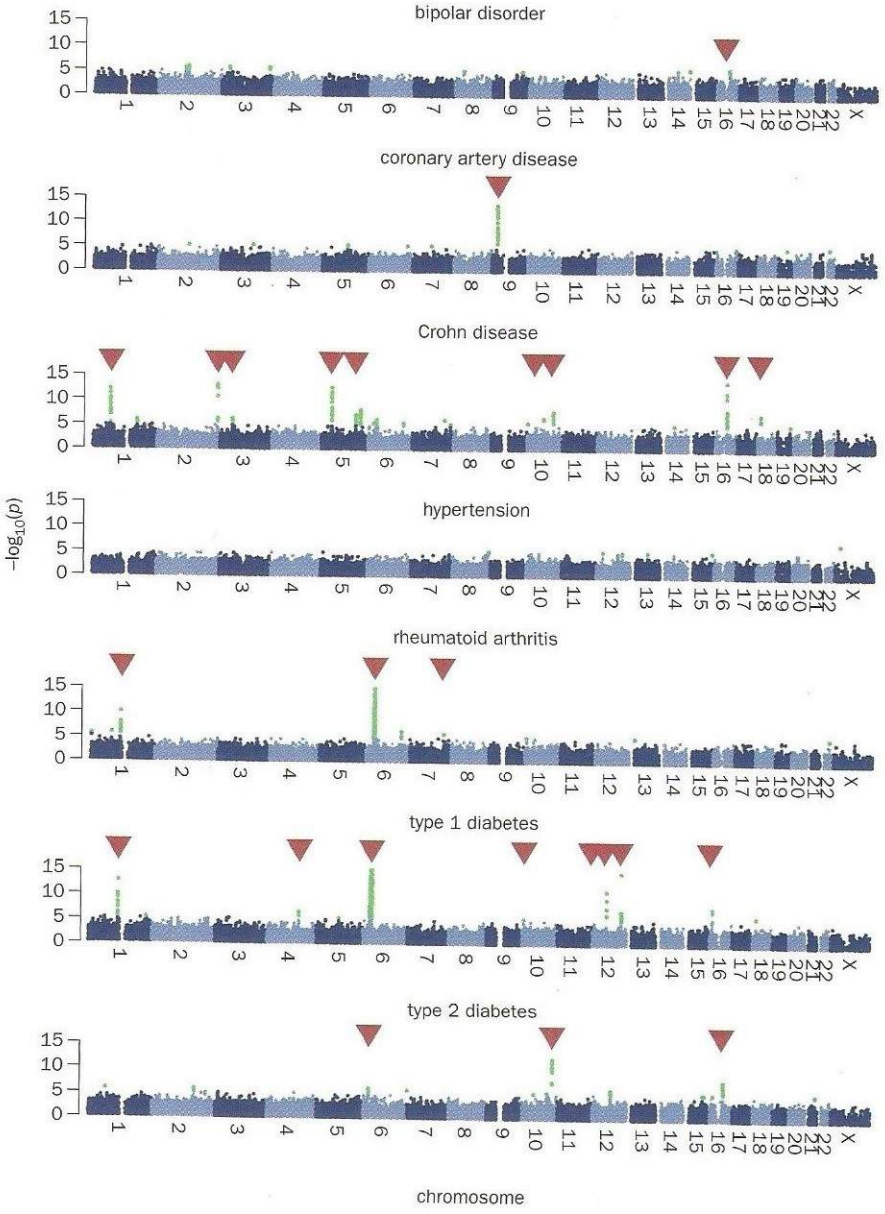

Figure 15.9 Results of the Wellcome Trust Case-Control Consortium (WTCCC) genomewide association study. For each of the seven diseases studied, the distribution of p values (as $-\log_{10}p$) for the association of each SNP with the disease is shown, at the appropriate chromosomal position. Most of the 469,557 SNPs that passed all the quality control checks showed weak or no association with the respective disease (blue dots, merged together). Those showing stronger evidence of association ($p < 10^{-5}$) are marked with green dots. The 25 most strongly associated SNPs or clusters of SNPs ($p < 5 \times 10^{-7}$, the threshold of significance in this study) are marked with red triangles. [Adapted from The Wellcome Trust Case-Control Consortium (2007) *Nature* 447, 661–676. With permission from Macmillan Publishers Ltd.]

WTCCC study, and one of the remaining two showed positive signals that did not quite reach significance. Almost all of the novel susceptibility factors have subsequently been confirmed in independent follow-up studies.

The WTCCC study is not the only one to have exploited the new tools and knowledge noted above. Between them, these studies clearly show that the era of reliable GWA studies has finally arrived.

The size of the relative risk

Considering the 25 SNPs showing the strongest evidence of association in the WTCCC study, the odds ratios (risk for a heterozygous carrier of the risk allele compared with a homozygous non-carrier) varied between 1.09 and 5.49. However, only the previously well-known associations of certain HLA types with autoimmune disease (type 1 diabetes and rheumatoid arthritis) produced high odds ratios. For all other cases, the highest ratio observed was 2.08, and in only five cases did it even exceed 1.5. As already mentioned, the study was predicted to have 80% power to detect a factor with an odds ratio of 1.5. The clear conclusion is that there are no individually strong ancient susceptibility factors for any of these typical common complex diseases.

A common story is emerging from this and other similar studies: the ancient susceptibility factors can be identified by current methods—but the relative risks they confer are quite small.

15.6 THE LIMITATIONS OF ASSOCIATION STUDIES

The new GWA studies on large cohorts of subjects have unprecedented power to detect susceptibility factors conferring relative risks as low as 1.2, and a proven track record of success. The question remains, however, whether these studies will unravel the complete genetic architecture of complex disease susceptibility.

One complicating factor is the existence of large-scale copy-number variations (CNVs) in the genomes of normal healthy persons. As described in Chapter 13, these CNVs actually account for a greater number of variable nucleotides than do all the SNPs. Individually, they are less numerous than SNPs, but each one covers many nucleotides. Many involve variable copy numbers of one or more genes, and in some cases (e.g. salivary amylase) there are demonstrable phenotypic effects of differing gene dosage. It is entirely likely that some CNVs will be susceptibility factors for common diseases. The question therefore arises: How well would current SNP association studies pick up associations between a disease and a CNV?

This question was addressed in the study by Redon et al. described in Chapter 13 of CNV in the HapMap sample. Their data showed that the HapMap Phase I SNPs would act as reasonably good tags for about 51% of the biallelic CNV in the non-African subjects, but only about 22% in the Africans. In other words, in 51% or 22% of cases, respectively, a particular SNP allele was strongly associated with a particular form of an insertion–deletion CNV, so that the tag-SNP would be associated with any disease risk caused by variation at the CNV. Multiallelic CNVs were not well tagged by any SNPs. Thus, older SNP association studies would be quite poor at detecting any disease susceptibility caused by a CNV. However, this is a passing phase. Newer generations of SNP genotyping arrays include assays for common CNVs.

The common disease–common variant hypothesis proposes that susceptibility factors have ancient origins

A more fundamental problem lies in the basic premise of testing for associations: the assumption that susceptibility factors are ancient common variants. This is the **common disease–common variant hypothesis**. But this hypothesis is controversial. Evidence both for and against it has been presented.

The fact that GWA studies are identifying associations that can be replicated tells us that some susceptibility factors are indeed ancient common variants. But the modest size of their effects leaves open the question of how much of the total disease susceptibility they will explain. Studies of breast cancer provide interesting figures. Breast cancer is about twice as common in first-degree relatives of affected women as in the general population. All known susceptibility factors identified before 2007 (*BRCA1*, *BRCA2*, *BRIP1*, *PALB2*, *ATM*, *CHEK2*, and *CASP8*) collectively accounted for less than 25% of the familial tendency of breast cancer. A whole-genome association study by Easton and colleagues (described in Chapter 16) identified five new factors. Despite the very large scale of the study, involving more than 20,000 patients, the new factors together accounted for only a further 3.6% of the excess familial risk.

Indeed, the underlying premise of the common disease–common variant hypothesis seems quite a tall order. The variants in question must have been able to persist through thousands of generations of natural selection to be associated with ancient haplotype blocks and remain in the population at high frequency. Yet, at the same time they must be sufficiently pathogenic, at any rate against certain relatively common genetic backgrounds and under certain relatively common environmental conditions, to be significant risk factors for serious diseases. It is hardly surprising that the common susceptibility variants that are now being identified through association studies have such weak effects. In defense, we can note that many of the diseases are of late onset, and so are relatively immune from natural selection and relatively unimportant in overall health terms until our current aging population made such diseases more prominent. Additionally, the precipitating environments may be features of modern life that have existed only recently. Genetically, we are all adapted to life as cavemen.

The mutation–selection hypothesis suggests that a heterogeneous collection of recent mutations accounts for most disease susceptibility

The alternative view is that common diseases are common because of mutation–selection balance. On this view, many or most susceptibility factors are deleterious enough to be removed by natural selection. These are replaced by new deleterious variants generated by recurrent random mutation. There is a simple relationship between the deleterious effect of a variant genotype and its likely persistence in a population:

- The variants that cause Mendelian diseases have very strong effects, so that most people with the susceptible genotype have the relevant disease. Such strongly disadvantageous variants, such as the dystrophin mutations that cause Duchenne muscular dystrophy, have a very rapid turnover.
- At the other end of the spectrum of pathogenicity, completely neutral variants can persist indefinitely in a large population (in a small population, sooner or later by random chance a variant will be either lost or fixed—that is, it will either be lost or entirely replace the alternative form).
- Very mildly deleterious variants may persist long enough to be present on ancient haplotype blocks, and so may be identifiable through association studies.
- Variants with a rather stronger effect will be removed from the population too fast to persist on common ancient haplotype blocks. The removal is balanced by *de novo* mutation creating fresh mildly deleterious variants.

The **mutation–selection hypothesis** supposes that this last class of variants make up the major susceptibility factors for common complex diseases. Calculations by Bodmer & Bonilla (see Further Reading) show that a susceptibility factor could have a penetrance as high as 10–20%—that is, that factor alone could have a 10–20% likelihood of causing disease, independently of other genetic or environmental factors—and yet would not give rise to a Mendelian pedigree pattern of disease. Typically, the physiological basis of disease susceptibility might be a partial loss of function in some complex pathway. This loss could be caused by any number of possible mutations in any of the genes involved in that pathway. There would be heterogeneity at the locus level (different affected people could have deleterious variants in different genes that are involved in the affected pathway) and very great heterogeneity at the allelic level (even if two people have variants in the same gene, the actual sequence change would most probably be different in each case). Each individual mutation may be rare, but deficiencies in the pathway may overall be quite common in the population.

Detecting such a heterogeneous set of rare variants, and proving a convincing role for them in pathogenesis of the disease, is quite a challenge. Each individual variant is directly pathogenic, but rare in the population. Association studies have little power to detect variants present in less than 5% of subjects, and so would be unable to pick up the postulated variants. Linkage analysis is not sensitive to allelic heterogeneity, and the individual effects might well be strong enough to show up in ASP analysis—if only all the sib pairs in the study sample had variants in the same gene. But the extensive locus heterogeneity would prevent success, because different affected sib pairs would often have variants at different unlinked loci.

Instead, it would be necessary to resequence candidate genes in large collections of cases and controls. Relevant genes would carry a higher overall frequency of rare, mildly deleterious variants in affected people than in controls. This excess would have to be picked out against a background of rare neutral variants that would be similarly diverse but equally common in the cases and controls. How easily this could be done would depend on the proportion of *de novo* mutations that are deleterious rather than neutral.

If the proportion of deleterious variants is low, they would probably be obscured among all the neutral variants. It could be difficult or impossible to find convincing differences between cases and controls. Inspection of the sequence would not usually reveal whether a change is neutral or very mildly deleterious. However, studies of naturally occurring missense changes in many different

genes suggest that maybe half of all such changes may be mildly deleterious (with a quarter being seriously deleterious and a quarter neutral). For silent changes and changes in noncoding sequences, the proportion of neutral variants is likely to be much higher.

Several preliminary resequencing studies have indeed found an excess of rare missense variants in candidate genes in disease cohorts, but such studies need to be conducted on a wider selection of genes and in more diseases to see whether the effect is general. The new massively parallel sequencing technologies allow this to be done, so that definitive tests of the mutation–selection hypothesis are now possible.

A complete account of genetic susceptibility will require contributions from both the common disease–common variant and mutation–selection hypotheses

The common disease–common variant and mutation–selection hypotheses should not be seen as mutually exclusive (**Table 15.10**). Both may very well be true. *A priori*, the common disease–common variant hypothesis is less plausible than the mutation–selection hypothesis. We know for sure that mutation and selection happen, and that there is a whole spectrum of mutations, ranging from lethal, through varying degrees of deleterious effect, to neutrality. It was not equally obvious that ancestral haplotypes carrying variants that predispose to disease should be able to withstand natural selection over immense time-scales and remain common in present-day populations. However, the successful identification of some common susceptibility factors shows that this has indeed happened, at least in some cases. There is no reason to suppose that all factors have to be of one sort. Different experimental designs can reveal different factors, and a complete account of genetic susceptibility will require contributions of both types.

Under either hypothesis, the problem of distinguishing pathogenic from neutral variants looms large. Factors identified by large-scale resequencing are expected to be directly pathogenic, but it will still be necessary to pick them out against a background of neutral variants. Meanwhile, linkage and association studies only flag the chromosomal location of the causative factor—to a relatively large region by linkage, or to a haplotype block by association. Thus, there is always a problem of identifying the actual causal variant. In Chapter 16 we consider how one approaches this problem, starting with Mendelian disease genes, then moving on to susceptibility factors. Finally, the implications of this work for predictive testing and personalized medicine are the subject of Chapter 19.

TABLE 15.10 COMPARISON OF THE COMMON DISEASE–COMMON VARIANT HYPOTHESIS AND THE MUTATION–SELECTION HYPOTHESIS

Parameter	Common disease–common variant hypothesis	Mutation–selection hypothesis
Frequencies of susceptibility alleles	high	low
Effect sizes of susceptibility alleles	small	moderate
Locus heterogeneity (number of susceptibility loci for a given disease)	high	could be low
Allelic heterogeneity (number of different susceptibility alleles at a locus)	low	high
Origin of susceptibility alleles	ancient common ancestor	relatively recent mutations
Technology to detect susceptibility factors	association studies	resequencing

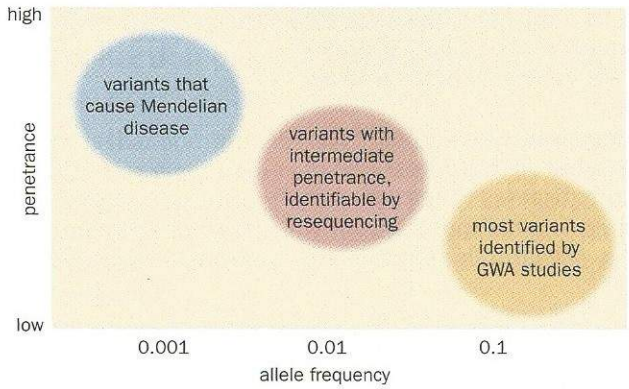

Figure 15.10 Frequency and effect size of pathogenic alleles. Variants causing Mendelian diseases have high penetrance but are individually rare. The variants conferring susceptibility to complex diseases that are identified through association studies may be common, but they have very weak effects (low penetrance). An intermediate class of individually rare susceptibility factors may exist that have stronger effects. These would not be expected to show populationwide associations with disease, and could be identified only by large-scale resequencing. [Adapted from McCarthy MI, Abecasis GR, Cardon LR et al. (2008) *Nat. Rev. Genet.* 9, 356–369. With permission from Macmillan Publishers Ltd.]

CONCLUSION

DNA sequence variants associated with disease can be divided into those that cause disease directly and those that merely modulate a person's susceptibility or resistance to environmental triggers of disease. These two types of sequence are not discrete, but lie on a continuum of disease penetrance. Variants that have a direct cause underlie Mendelian diseases, and are necessarily rare. As described in Chapter 14, they can be mapped by parametric linkage analysis to very small candidate chromosomal regions. Variants that confer susceptibility lie at the other end of the continuum—they may be common, but they have only very modest effects on susceptibility (**Figure 15.10**).

Parametric linkage analysis cannot be used to identify these low-penetrance factors, because it is not possible to specify the necessary detailed genetic model. Non-parametric linkage analysis, for example of affected sib pairs, has been widely used, but with limited success because of the low statistical power of the method. Large-scale successful identification has depended mainly on association studies. **Figure 15.11** summarizes the general protocol of genomewide association studies.

The HapMap data allow suitable SNPs to be identified for genomewide association studies and placed on high-density SNP arrays, and funding agencies have recognized the need for very large-scale collaborative studies. As a result, the trickle of validated associations reported before 2005 has turned into a flood. Association studies of feasible size (a few thousand cases and controls) can detect only factors that are common in populations, typically with allele frequencies of 0.05 or more. Factors that are common and associated with tag-SNPs are likely to be present on shared ancestral chromosome segments that have existed in the population for thousands of years. The common disease–common variant hypothesis proposes that such factors are largely responsible for the genetic susceptibility to complex diseases. It would be unusual for such an ancient variant to have a strongly pathogenic effect: natural selection should have long ago removed such variants from the population. It is therefore not surprising that almost all the susceptibility factors identified through association studies have very small individual effects.

The alternative mutation–selection hypothesis proposes that much susceptibility is due to individually rare variants that are likely to be of fairly recent origin. The hypothesis predicts extensive allelic heterogeneity, but maybe more limited locus heterogeneity—that is, many different mutations at a possibly limited number of loci. Such variants would be too rare to be identified by association studies, but may be detected by large-scale resequencing of candidate regions (or whole genomes) in cases and controls. Because the variants may not have survived in the population for very long periods, they may have stronger individual effects than the ancient common variants discussed above.

Most of the hundreds of well-validated susceptibility factors identified through association studies are probably not directly responsible for disease susceptibility, but rather are in linkage disequilibrium with the truly functional variants. Identifying the true causative variants would have many benefits. Chapter 16 discusses the various strategies that have been used to identify variants that function to cause or contribute to disease.

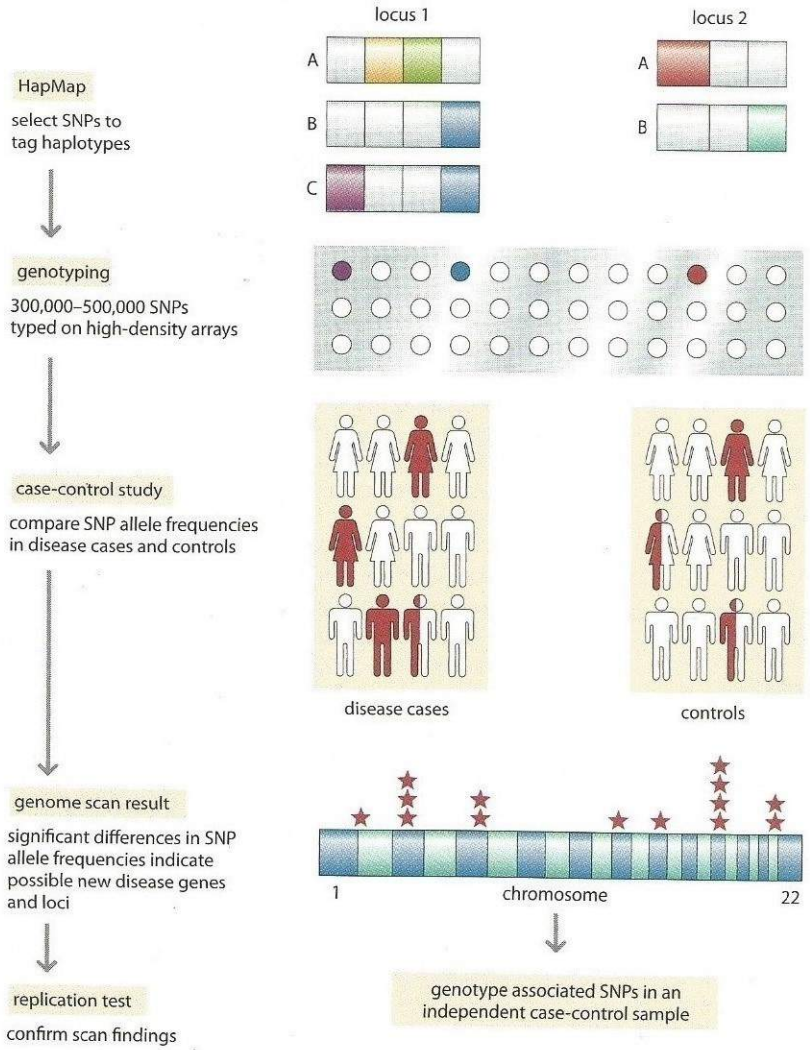

Figure 15.11 A genomewide association scan. Common haplotypes at each location in the genome are defined by SNPs (color versus white). Information from the HapMap project is used to select a subset of SNPs that will serve to identify (tag) each haplotype—purple and blue for locus 1, and either red or blue for locus 2. Disease cases and controls are genotyped with microarrays. SNPs that are associated with disease at an appropriate statistical threshold (stars) are then genotyped in a second independent sample of cases and controls to establish which of the associations from the primary scan are robust. [From Mathew CG (2008) *Nat. Rev. Genet.* 9, 9–14. With permission from Macmillan Publishers Ltd.]

FURTHER READING

Family studies of complex disease

Burmeister M, McInnis MG & Zollner S (2008) Psychiatric genetics: progress amid controversy. *Nat. Rev. Genet.* 9, 527–540. [An overview of progress since the pioneering studies used as examples in this section.]

McGuffin P, Shanks MF & Hodgson RJ (eds) (1984) The Scientific Principles Of Psychopathology. Grune & Stratton.

Risch N (1990) Linkage strategies for genetically complex traits. 1. Multilocus models. 2. The power of affected relative pairs. 3. The effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.* 46, 222–228; 229–241; 242–253. [Three key papers establishing the statistical basis of familial clustering and shared segment analysis.]

Rosenthal D & Kety SS (1968) The Transmission Of Schizophrenia. Pergamon Press.

Segregation analysis

Badner JA, Sieber WK, Garver KL & Chakravarti A (1990) A genetic study of Hirschsprung disease. *Am. J. Hum. Genet.* 46, 568–580. [A good example of segregation analysis applied to a non-Mendelian condition.]

McGuffin P & Huckle P (1990) Simulation of Mendelism revisited: the recessive gene for attending medical school. *Am. J. Hum. Genet.* 46, 994–999. [A warning about some pitfalls of segregation analysis.]

Linkage analysis of complex characters

Altmüller J, Palmer LJ, Fischer G et al. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* 69, 936–950. [A sobering meta-analysis of 101 linkage studies in 31 complex diseases.]

Lander ES & Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* 11, 241–247. [Introducing the widely used categories of suggestive, significant, and highly significant results.]

Association studies and linkage disequilibrium

Cardon LR & Bell JI (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99. [A general non-mathematical discussion of designs for association studies.]

International HapMap Consortium (2003) The International HapMap Project. *Nature* 426, 789–796. [A description of the aims and methods of the project.]

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320. [The primary report of Phase I of the HapMap project; available for download from http://www.hapmap.org]

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861. [Report of Phase II.]

Jobling MA, Hurles ME & Tyler-Smith C (2004) Human Evolutionary Genetics: Origins, People and Disease. Garland Science. [A unique and excellent textbook that, among many other things, sets out the whole background to linkage disequilibrium.]

Lohmueller KE, Pearce CL, Pike M et al. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33, 177–182. [Showing that some early association studies did indeed produce replicable results.]

Risch N & Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517. [The power calculations in this paper helped trigger the move from linkage to association studies; see also *Science* 275, 1327–1330 (1997) for discussion.]

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485. [A useful general introduction to LD, including definitions, measures, and origins.]

Association studies in practice

Altshuler D & Daly M (2007) Guilt beyond a reasonable doubt. *Nat. Genet.* 39, 813–815. [A brief review of the achievements of the first wave of successful genomewide association studies.]

Campbell CD, Ogburn EL, Lunetta KL et al. (2005) Demonstrating stratification in a European American population. *Nat. Genet.* 37, 868–872. [A cautionary tale about the need to match cases and controls very carefully.]

Database of Genotypes and Phenotypes (dbGaP). http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/about.html [Planned to be a central resource for raw data relating genotypes and phenotypes, including results of association studies.]

Schaid DJ (1998) Transmission disequilibrium, family controls and great expectations. *Am. J. Hum. Genet.* 63, 935–941. [A review of the strengths and weaknesses of the TDT.]

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. [Describes the statistical basis for the transmission disequilibrium test, and shows an example of its power.]

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. [An excellent overview of how to perform GWA studies and what they might show.]

Wilson JF & Goldstein DB (2001) Consistent long-range linkage disequilibrium generated by admixture in a Bantu–Semitic hybrid population. *Am. J. Hum. Genet.* 67, 926–935. [The advantages of an admixed population for association studies.]

Wright AF, Carothers AD & Pirastu M (1999) Population choice in mapping genes for complex diseases. *Nat. Genet.* 23, 397–404. [Despite its age, a good review of the options.]

The limitations of association studies

Bodmer W & Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common disease. *Nat. Genet.* 40, 695–710. [An important contribution to the debate about the relative merits of the common disease–common variant and mutation–selection hypotheses.]

Helbig I, Mefford HC, Sharp AJ et al. (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat. Genet.* 41, 160–162. [One of several studies showing that the microdeletion acts as a typical low-penetrance susceptibility factor for several different neuropsychiatric disorders. This paper includes references to other studies.]

Jacobsson M, Scholz SW, Scheet P et al. (2008) Genotype, haplotype and copy number variation in worldwide human populations. *Nature* 451, 998–1003. [An extension of HapMap-type investigations to 485 individuals from 29 populations.]

Kryukov GV, Pennachio LA & Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* 80, 727–739. [Data and calculations suggesting that around 50% of rare coding-sequence variants are mildly deleterious, thus supporting the mutation–selection hypothesis.]

Pritchard JK & Cox NJ (2002) The allelic architecture of human disease genes: common disease–common variant or not? *Hum. Mol. Genet.* 11, 2417–2423. [Arguments against the common disease–common variant hypothesis.]

Reich DE & Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510. [Arguments in favor of the common disease–common variant hypothesis.]

Topol EJ & Frazer KA (2007) The resequencing imperative. *Nat. Genet.* 39, 439–440. [A brief review of the case for large-scale resequencing and results to date.]