ASSOCIATION STUDY DESIGNS FOR COMPLEX DISEASES

Lon R. Cardon and John I. Bell

Assessing the association between DNA variants and disease has been used widely to identify regions of the genome and candidate genes that contribute to disease. However, there are numerous examples of associations that cannot be replicated, which has led to scepticism about the utility of the approach for common conditions. With the discovery of massive numbers of genetic markers and the development of better tools for genotyping, association studies will inevitably proliferate. Now is the time to consider critically the design of such studies, to avoid the mistakes of the past and to maximize their potential to identify new components of disease.

POWER
The probability of correctly rejecting the null hypothesis when it is truly false. For association studies, the power can be considered as the probability of correctly detecting a genuine association.

University of Oxford, Nuffield Department of Clinical Medicine, Headington, Oxford OX3 9DU, UK. Correspondence to: J.I.B. e-mail: john.bell@ndm.ox.ac.uk

Genetic association studies assess correlations between genetic variants and trait differences on a population scale. On the genetic side, until recently there were relatively few (100s to 10,000s) DNA variants available to study, but advances in the past two years alone have identified nearly two million such polymorphisms[1,2]. On the trait side, the phenotypes assessed in association studies include disease status, such as osteoporosis, asthma or diabetes; continuously varying measures, such as bone density, bronchial responsiveness or glucose levels; and response to environmental stimuli, such as drug efficacy or side effect. This breadth of genetic information and depth of phenotypic measure holds considerable promise for identifying genotype–disease correlations, and is one of the main thrusts of the drive towards personalized medicine[3].

Unfortunately, the literature is teeming with reports of associations that either cannot be replicated or for which corroboration by linkage has been impossible to find[4–6]. Explanations for this lack of reproducibility are well-rehearsed, and typically include poor study design, incorrect assumptions about the underlying genetic architecture and simple overinterpretation of data. The common errors encountered in association studies of complex diseases are summarized in BOX 1. Despite these known limitations, the power of association analysis to detect genetic contributions to complex disease can be much greater than that of linkage studies[7]. The opportunity to use this power of association strategies has led to efforts to develop association methodologies that attenuate some of the most widely perceived limitations of the approach.

The requirement for linkage disequilibrium mapping and association methodologies in common disease has arisen because of the recognized limitations of existing linkage strategies in these disorders[8]. Although powerful for detecting genetic loci in single gene disorders, linkage analysis of common, multifactorial diseases has been limited by the lack of clear genetic segregation of any DNA variants in multigenerational family material, and by the modest contribution to disease made by individual genetic variants. Even where linkage has been identified reproducibly, it has seldomly led to the resolution of linkage regions to less than a few megabases in most common diseases[6]. A strategy for refining this linkage information, and for searching for genetic variants of small effect, is therefore essential, and association studies are seen to address this need[9–11]. Despite their recognized limitations, association studies represent an essential step in advancing the field to the definition of disease-mediating genetic variants.

Recently, Risch[7] reviewed the statistical framework of association studies, comparing the statistical power to detect multifactorial genetic effects in linkage and association designs. His POWER calculations highlighted some advantages of the association design over the linkage strategy in terms of sample size. Still, there are a number

GENETIC DRIFT
The random fluctuation in allele frequencies as genes are transmitted from one generation to the next.

POPULATION ADMIXTURE
A population in which multiple subgroups are included. Admixture often refers to intermarriage/reproduction from different groups of individuals, but most simply is used to denote a population of subgroups having different allele frequencies (see population stratification).

Box 1 | Common errors in association studies

- Small sample size
- Subgroup analysis and multiple testing
- Random error
- Poorly matched control group
- Failure to attempt study replication
- Failure to detect linkage disequilibrium with adjacent loci
- Overinterpreting results and positive publication bias
- Unwarranted 'candidate gene' declaration after identifying association in arbitrary genetic region

of potential pitfalls and limitations of association studies[6], many of which depend on the particular design, study aims and analytical framework adopted[12]. Here we review some of the most commonly used association designs and applications, and discuss some of their limitations and possible solutions.

Linkage disequilibrium

Genetic variants and trait scores can become associated by several mechanisms[6]; that of greatest present interest is linkage disequilibrium (LD), also known as gametic phase disequilibrium or allelic association (BOX 2). The history of LD dates back to 1909, to the original observations of Weinberg, who documented that alleles at two adjacent loci asymptotically approach random association in a population (reviewed in REF. 13). The presence of LD between alleles mainly reflects the recombination history in the population of that haplotype. Therefore, recently acquired mutations, or those in founder or isolated populations with limited chromosome diversity, are likely to show extensive LD that might extend over long distances[14–18]. However, as a result of a number of contributing factors, including regional variability in recombination patterns, GENETIC DRIFT, mutation age, ethnic diversity and recent POPULATION ADMIXTURE, local chromosomal composition and the pattern of mating within a population, patterns of LD can vary significantly within and between different populations[19–25].

Perhaps the first formal use of LD mapping was in 1947 by Fisher[26], who used it to establish the order of loci on the basis of allele frequencies for the blood antigen, rhesus factor. The strategy was later applied widely in the major histocompatability complex region on chromosome 6 when serologically defined protein polymorphisms in the human leucocyte antigen (HLA) system were identified. These studies led to the association of the HLA region with a range of immunologically mediated diseases and, more recently, infectious diseases. The association studies were facilitated by the extensive LD that exists in this region over distances of greater than 3 cM. Many of the rules currently applied to association strategies arose from characterization of the HLA with serological tools, before the advent of robust DNA-based genotyping. Work in the HLA region demonstrated the power of association studies to define areas contributing to disease, but also the difficulty that can arise in attempting to clarify the role of individual DNA variants in regions of strong and extreme LD.

Ironically, the HLA work also indirectly led to considerable scepticism about association studies that persists unabated today. Because the HLA variants were the first to be widely available, association studies of a broad spectrum of phenotypes were undertaken, occasionally in the absence of careful study design, assessments of statistical power or reasoned hypotheses of HLA involvement. An inevitable consequence of this HLA overexposure was a large number of unreplicable findings[27].

For several reasons, there is great enthusiasm at present about the promise of association studies for uncovering the genetic components of complex disease: dense single-nucleotide polymorphism (SNP) maps across the genome[1,2]; elegant, high-throughput genotyping technologies[28]; simultaneous comparison of groups of loci; statistical measures for assessing genome-wide significance; and the phenotypic insight that might accompany comparative genomic studies among different human subgroups. Although these areas are indeed exciting, complex trait studies must be careful to avoid recreating the outcomes of the 'blind' association strategy used 20 years ago with HLA. This requires careful attention to the sample collected and to the study design employed, which are probably the most crucial elements of any association study.

Sampling strategies for association studies

The case–control study has been the most widely applied strategy of association studies for characterizing the genetic contributions to disease. The advantages of this approach are that cases are readily obtained and can be efficiently genotyped and compared with control populations. Despite its ease, however, this approach has been the most prone to identifying DNA variants that prove to be 'spuriously associated' with disease. As in the case of HLA, the spurious nature of the association is usually defined by the failure to replicate the original association data in subsequent studies, and by the inability to provide evidence for genetic linkage[10]. A particular difficulty with this methodology is the choice of control populations, which are often, by necessity, retrospectively defined; that is, identified, matched and ascertained after collection of the disease group.

The selection of controls is crucial because any systematic allele frequency differences between cases and controls can appear as disease association, even if they only reflect the results of evolutionary or migratory history, gender differences, mating practices or other independent processes. When allele frequency differences are coupled with differences in disease frequency (as is the aim of case–control studies), the resulting evidence for association might be statistically highly significant, but unrelated to the actual influence of the allele under investigation. From the perspective of a disease gene, such studies are spurious in that the association is due to the structure of the population studied rather than to LD (BOX 3). Linkage studies of such DNA variants would not detect significant results, because no familial segregation would be apparent in the population subgroups.

PROSPECTIVE COHORT
Longitudinal study of individuals initially assessed for exposure to certain risk factors and then followed over time to evaluate the progression towards specific outcomes (often disease).

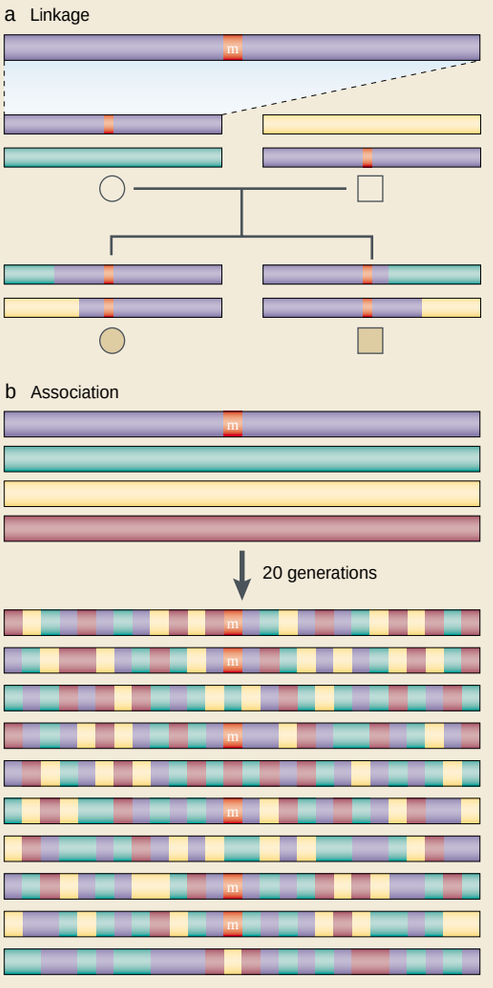
Control ascertainment can be improved by using a PROSPECTIVE COHORT study. This requires a substantial collection of individuals to be selected before the onset of disease and to be matched with individuals followed under the same experimental protocol. In this way, there is no bias for the selection of a control population. This approach, however, requires significantly more resources and patience to allow sufficient numbers of cases to emerge before association studies can proceed. It has the additional advantage that prospective collection of environmental information can be achieved. Given that most studies to date have not incorporated environmental effects, nor the likely influence of genotype–environment interactions, this will be of increasing importance for the analysis of common disease[29].

These study designs require significant numbers of cases to be adequately powered to study disease genetics. Because of the modest effects likely to be contributed by disease genes, the presence of LOCUS HETEROGENEITY, the tendency to analyse subgroups of the patient population and the characterization of allele frequencies at markers distant and with uncertain linkage disequilibrium from the functional variants, it is easy to understand why such studies have historically yielded ambiguous results[8]. These studies often incorporated 100 or fewer cases and a similar number of controls, and it is now thought to be likely that such studies are possible only if done with much larger patient samples[30,31]. Indeed, studies involving up to thousands of individuals have provided accurate estimates of risk associated with particular alleles[32–34]. Of course, particular attention must be paid to sample sizes when more than one variable is studied simultaneously, such as multiple imperfectly correlated traits, intergenic interactions (epistasis) or gene–environment interactions. Such studies require even larger patient populations to ensure that individual subgroups retain adequate power to detect significant associations with narrow confidence intervals.

The crude definition of most diseases being studied using this methodology will inevitably progress towards a refinement of disease definition once more is known about the molecular basis of disease subtypes. This will lead to opportunities to limit association studies to subgroups defined by specific phenotypes not previously recognized to be significant, on the basis of the presence of particular alleles at certain loci, or of epigenetic factors such as parent-of-origin affects. These opportunities create an iteratively developing knowledge of genetic associations, in which weak associations are improved by establishing hypotheses on the basis of subgroups and then testing new cohorts. This approach will be necessary, given the extent of locus heterogeneity for many common diseases, but might be statistically hazardous unless considered carefully with regard to study design and data analysis.

Repeated subgroup analysis is a valid route to generate hypotheses, but these hypotheses must be subsequently tested in additional patient populations. A common error in association studies is to analyse the patient population repeatedly using different clinical phenotypes or genotypically defined subgroups. This repeated testing might be implicit, with subgroups considered and abandoned because of a lack of significance. Multiple testing generates subgroups that are small, providing less robust results, and also creates a substantial risk of associations being described by chance (statistical Type I error). Such analysis might be necessary given the crude phenotypic definition of disease. Ultimately, this will drive the need for large patient populations to test

Box 2 | Linkage versus association



At a fundamental level, genetic association and linkage analysis rely on similar principles and assumptions[87]. Both rely on the co-inheritance of adjacent DNA variants, with linkage capitalizing on this by identifying haplotypes that are inherited intact over several generations (such as in families or pedigrees of known ancestry), and association relying on the retention of adjacent DNA variants over many generations (in historic ancestries). Thus, association studies can be regarded as very large linkage studies of unobserved, hypothetical pedigrees. In growing populations, such as humans, recombination is the primary force that eliminates linkage and association over generations[88]. When a functional mutation occurs ('m' in the figure) — perhaps one that contributes to disease — it does so on a haplotype of other pre-existing DNA variants. Because linkage focuses only on recent, usually observable ancestry, in whom there have been relatively few opportunities for recombination to occur, disease gene regions that are identified by linkage will often be large, and can encompass hundreds or even thousands of possible genes across many megabases of DNA (figure panel **a**). By contrast, association studies draw from historic recombination so disease-associated regions are (theoretically) extremely small in outbred random mating populations[89], encompassing only one gene or gene fragment (figure panel **b**). Through subsequent generations, as the disease mutation is transmitted, recombination will cause it to be separated from the specific alleles of its original haplotype. Particular DNA variants can remain together on ancestral haplotypes for many generations. This type of non-random association of alleles is known as linkage disequilibrium. It is linkage disequilibrium that provides the genetic basis for most association strategies.

LOCUS HETEROGENEITY
The appearance of phenotypically similar characteristics resulting from mutations at different genetic loci. Differences in effect size or in replication between studies and samples are often ascribed to different loci leading to the same disease.

and validate hypotheses sequentially in different populations. Sample sizes of a thousand or more will not prove to be excessive for these sorts of iterative studies.

The merits of large sample sizes have recently been shown in studies on the role of polymorphisms around the angiotensin-converting enzyme (*ACE*) locus and its contribution to the risk of cardiovascular disease. Early publications on this disease association in the ECTIM (Étude Cas-Temoin de l'Infarctus du Myocarde) study, which involved 610 men who survived myocardial infarctions and 733 controls, suggested that the *ACE* locus had a role in the risk of particular subgroups to cardiovascular disease[35]. Confidence intervals for this initial study were large. Subsequent studies often involved even fewer patients, and those published seemed to produce variable results. The positive nature of many of these smaller studies might have reflected publication bias towards positive results. When the hypothesis was tested in a very large case–control study (4,629 cases and 5,934 controls), the evidence of an association between *ACE* and an increased risk of cardiovascular disease was much diminished (RISK RATIO 1.1, confidence interval 1.00–1.21)[33]. This example demonstrates that, for modest genetic effects and where genotypic subgroups are being identified, sample sizes involving 1,000 to 10,000s of individuals might be required to generate robust data. An additional example of this outcome has been recently reported for late-onset diabetes[34]. This principle, if applied widely, would considerably reduce the number of unreplicated results obtained using association methodology.

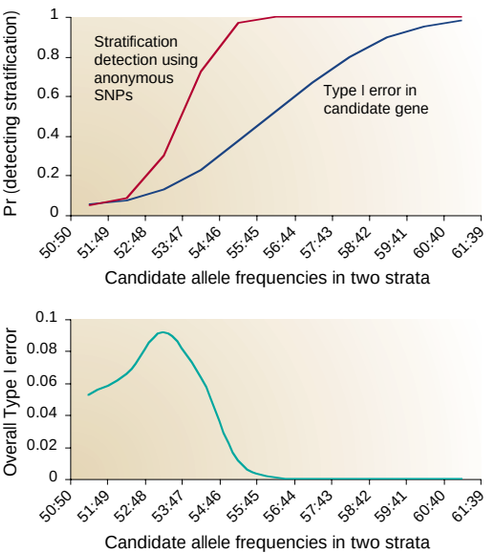
Technically, case–control studies have one main advantage. DNA samples from cases and controls can be pooled and genotyped together to determine differences in allele frequency. This methodology, pioneered in the HLA system[36] and recently extended in many contexts[30,37–40], should provide a powerful mechanism for testing efficiently large numbers of alleles across cases and controls. Pooling in complex diseases requires very high accuracy, however, because genuine differences in allele frequencies between cases and controls might be quite small. It is not yet clear that the measurement of signal intensity for each allele within pooled DNA samples can reach the necessary tolerance levels for these studies, as even small levels of experimental error in pooling assays (1–2%) might be too high for such applications. Moreover, pooling does not lend itself to direct haplotype assessment, which is likely to be essential for complex trait mapping.

What is the optimal control population?

Family-based controls. Nearly all human genetic association methods developed in the past decade have focused on the need to account for population stratification. To reduce the effect of population stratification on association studies, various approaches have been developed that use controls selected from the families of affected probands. Falk and Rubinstein[41] were among the earliest developers of this methodology, and used it to construct the haplotype relative risk test. They classified parental alleles into those transmitted to affected children and those not transmitted. By using those alleles not transmitted as controls, they were able to evaluate the risk of disease arising from particular allelic markers.

At present, the most popular extension of this approach, known as the transmission disequilibrium test (TDT[42–44]), focuses only on heterozygous parental genotypes, thereby providing a joint test of linkage and association that eliminates the effects of stratification when applied to single probands and parents. Various further extensions have been developed for multiallelic

Box 3 | Population stratification



Fear of false-positive outcomes arising from POPULATION STRATIFICATION has virtually dictated the progress in human association study design and analysis methodology over the past decade. In large part, complex disease association studies have moved from the traditional case–control model to family-based controls, in which alleles that are not transmitted from parents serve as proxies for control samples. Recently, however, some investigators have begun to explore the use of unlinked genetic markers to detect stratification and even correct for it when it is present[90–94]. At the core of these methods is the idea that population substructure, operationally defined by differences in allele frequencies between subsets of a given sample, should be detectable by evaluating allele frequency patterns at a number of anonymous markers. This information can then be used to alter the interpretation or analysis of the candidate gene(s) of interest.

The top graph illustrates the balance between the false-positive rate at a candidate locus and the rate of detecting stratification using anonymous unlinked markers (single nucleotide polymorphisms (SNPs)). Two hundred cases and two hundred controls were simulated with allele frequencies at a candidate locus differing by a prescribed amount (shown on the *x* axis), and 40 unlinked markers were simulated on the basis of these frequencies under the model of Bacanu *et al.*[94]. The *y* axis indicates the probability (Pr) of detecting evidence for population stratification at the 95% significance level. Clearly, TYPE I ERROR at the candidate locus (an apparent difference in allele frequencies between cases and controls) increases rapidly as the allele frequency difference widens (blue line); however, the ability to detect stratification increases even faster (red line). At first glance, this could lead one to question the ongoing concerns about stratification: genotyping 40–100 markers in available samples is not a difficult endeavour. Nevertheless, consideration of the overall false-positive rate (that is, a false-positive outcome at a candidate locus without detecting stratification using additional markers[90]), suggests that substantial inflation of Type I error can still occur (bottom graph, green line). In the hypothetical study shown in the figure, a nominal false-positive rate of 5% is inflated by a factor of two. Moreover, the maximal inflation occurs when the cases and controls show only small differences in allele frequency. Unfortunately, these small differences might reflect precisely the magnitude of effect we seek to detect at complex disease loci. Actual inflation levels can be better or worse than this, depending on the nominal significance rate adopted, the sample size studied and the number and spacing of background markers genotyped. Thus, population stratification might yet remain a thorn in the side of complex trait association studies.

POPULATION STRATIFICATION
The presence of multiple subgroups with different allele frequencies within a population. The different underlying allele frequencies in sampled subgroups might be independent of the disease within each group, and they can lead to erroneous conclusions of linkage disequilibrium or disease relevance.

TYPE I ERROR
The probability of rejecting the null hypothesis when it is true. For association studies, Type I errors are manifest as false-positive reports of phenotype–genotype correlation.

RISK RATIO
A measure of association effect reflecting the probability of disease in people with a particular allele or genotype versus the probability of disease in those who do not have the particular genotype.

markers[45], multiple siblings in a family[46], missing parental data[47,48] and quantitative traits[49–52]. This test has been popularized by applications to many diseases, but perhaps most often in positional cloning attempts in juvenile onset diabetes[42,53–58]. A detailed mathematical review of the TDT has recently been presented[59].

An unfortunate side effect of the TDT is that it effectively throws away some genotype information, owing to its reliance on heterozygous parents. This creates a loss of statistical power to detect genuine allelic association. In its original form, the TDT also requires parental genotypes, which are not easily accessible for late-onset disorders. Consequently, although more robust in the presence of population stratification, the family-based methodologies can be more difficult or even impossible to implement, and might require significantly more patients and family collections than case–control studies[60].

Matched control populations. If family-based selection is not practical (for example, in pharmacogenetics applications, some late-onset diseases or large prospective cohorts), then selecting an appropriate control population is essential to enable the optimal design of case–control studies. Ideally, a control sample should reflect the ethnic and genetic composition of the case sample. For this reason, 'off-the-shelf' control samples used for a range of disease association studies are likely to prove inappropriate, as controls should be carefully matched with the disease sample. This matching can be notoriously difficult to achieve, however, and several different control populations might have to be used in any given study. Prospective studies can provide more robust control populations but are substantially more costly because of their necessary scale.

Another approach for selecting a control population is to collect several control populations reflecting the various substructures that might exist in the case population. Such a control set might include populations that are both geographically dispersed and ethnically defined to establish the range of allele frequencies that might be encountered at a single locus. Such a control panel might also provide age-related allele frequencies, as the population structure will vary with age as a result of historical patterns of migration, mating, selection and other forces. This enables allele frequencies from the defined case population to be tested against a matched control panel, as well as evaluated against a range of allele frequencies from subpopulations that might bias the outcome. Any distortion of allele frequency in the case groups that lies outside the range of control frequencies would reflect apparently real effects.

Ethnic diversity leading to population stratification can bear significantly on the power of an association study. The evolutionary history of haplotypes and linkage disequilibrium patterns will vary significantly in different ethnic populations. Disease populations that contain an ethnic predominance must be matched to appropriate controls. It has been argued that the choice of particular ethnic groups might also facilitate association studies as these might arise from limited numbers of founders and, owing to a relatively short evolutionary history, would provide less disease heterogeneity and larger regions of linkage disequilibrium, facilitating association strategies. Sardinia, Finland and Iceland have all been selected for large studies that are based around such 'founder effects'. Interestingly, evidence for more extensive patterns of linkage disequilibrium has been difficult to demonstrate when tested objectively in these populations compared with more outbred populations[16–18]. Whether there is disease heterogeneity in these latter populations remains uncertain.

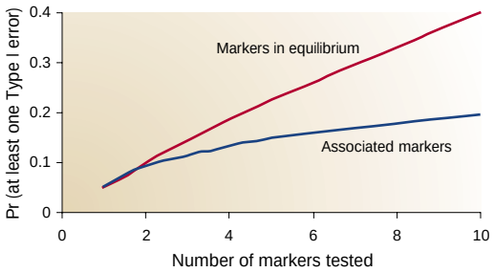
Applications of association methodology

Regional linkage disequilibrium mapping. One application of association methodology has been to try to identify the precise disease-causing DNA variant(s) in a region that is known to be linked and associated with a disease. Within a targeted region, two association strategies are common: a positional candidate approach, in which specific genes or variants are examined on the basis of proposed relationships with the phenotype; and a positional cloning approach, in which markers are selected for evaluation purely on the basis of their proximity to one another on a chromosome. In either case, for common diseases these regions are often localized by replicated linkage data, which typically refine the initial region down to 1–10 cM. Within such a region, the identification of a causative DNA variant might prove to be a challenging task.

Such strategies were originally applied to identify genes involved in fully penetrant gene disorders, such as cystic fibrosis and Huntington disease, in which haplotypes in linkage disequilibrium within a region of linkage allowed disease genes to be identified[61–64]. This has proved to be more difficult to achieve in complex traits and has relied on the characterization of large numbers of polymorphisms within a region — a task requiring extensive sequencing and the development of genotyping assays. By examining allelic variants across a region of known linkage, it should be possible to see variations in the strength of associations between these markers and a phenotype or disease, enabling the trait-associated locus and causative or susceptibility mutation to be mapped[65–68]. Conventional case–control strategies can be applied in these circumstances, as can TDT approaches. The application of haplotypes to this analysis can provide stronger evidence of a region being associated with the disease phenotype than can the use of individual markers[69].

These studies are most often limited by poor understanding of the patterns of linkage disequilibrium within large regions of linkage and by the limited number of polymorphisms that are available within such regions. For example, it has been widely assumed that linkage disequilibrium declines as markers become more distant from disease polymorphisms. This proves to be a significant oversimplification of the real situation. Consistent with the complex evolutionary history of any set of haplotypes, islands of linkage disequilibrium can be separated by regions where little linkage disequilibrium exists. A marker with strong disease association surrounded by markers with no association cannot be

Box 4 | Multiple testing in association studies



The increasing numbers of association scans implies that many thousands of statistical comparisons will be conducted on the same patient samples. In addition, many more association tests might be conducted in follow-up to the initial findings. Clearly, the classic nominal significance-threshold framework is inappropriate for such studies; if k is the nominal significance rate for one marker and m independent markers are tested, then false-positive results will be obtained at a frequency of $1 - (1 - k)^m$. For example, if $k = 0.05$ and as few as 100 markers are tested for association, there is a greater than 99% chance that one or more of the markers will appear significantly related to disease. This is obviously an unacceptable rate. The most appropriate mechanism for preventing such erroneous conclusions has been a matter of debate[10,95], and although the details of this issue might seem to be of an esoteric statistical nature, it is important to realize that inappropriate correction for multiple testing inevitably leads to undesirable outcomes: either increased false-positive levels (owing to a weak correction), or decreased statistical power to detect effects (owing to an overly stringent correction).

In the figure, ten highly associated markers in a published dataset were evaluated for association with randomly assigned cases and controls (founders in the study of Keavney *et al.*[69]). The graph shows the probability (Pr) of the detection of a Type I error plotted against the numbers of markers tested. Each of the markers was successively tested for significance and the number of test statistics exceeding the $k = 0.05$ threshold were tabulated. The expected increase in false-positive rate, assuming that the markers are entirely independent, is shown as a red line. The actual increase in false-positive rate is shown as a blue line. Clearly there is an increase in false positives, but it is not nearly as high as predicted under an independence model. Bonferroni correction (a popular correction that assumes independent markers) of these data would markedly overcorrect for the inflated false-positive rate and thereby throw away valid information in the sample.

It is unlikely that any single statistical model can account for the variability in LD across the genome. Consequently, the most prudent method might be one involving permutations of each individual dataset. These can be computationally time consuming, but at least should provide an accurate view of extreme events in the context of the genomic background.

interpreted as being a disease-mediating variant, as more distant markers in linkage disequilibrium might exist outside those where no association occurs[13,19,68,70,71]. In this case, the burden of aetiological allele identification could require assessment of many or all polymorphisms in a region, with inferences drawn using the relative strength and pattern of association observed.

Even when convincing evidence is obtained, the problems associated with very tight linkage disequilibrium in regions that have been identified as being both linked and associated with disease are often understated. Thus, in contrast to intermittent, weak evidence for association, an alternative problem is that many alleles in a gene or genes might be strongly associated[16], thereby precluding resolution of individual effects. This problem is exemplified in associations between circulating levels of ACE and variants in the *ACE* gene, in which a number of alleles exhibit strong, almost uniform, LD with the phenotype[69,72,73]. Such problems can reduce the value of isolated populations, despite strong linkage disequilibrium across predominant ancestral haplotypes. Given strong evidence for LD, it can still prove difficult or impossible to identify causative mutations in disease genes themselves. This can be further complicated by the clustering of loci that share similar functions within a single genomic region of strong LD, as is seen in the HLA region.

One approach to breaking down such regions of linkage disequilibrium is to characterize the disease phenotype in diverse populations that might share substantially different ancestries for the genomic region of interest. Such 'transracial mapping' has allowed the dissection of strongly conserved regions with extensive linkage disequilibrium, and is likely to prove even more useful as linkage-disequilibrium mapping becomes used more widely[74,75]. This is essentially a domain of comparative genomics, but in which comparisons within species form the basis of interest. Whatever the regional association strategy adopted, definitive evidence of a role for a mutation or gene will require functional studies of the polymorphisms, as aetiological effects cannot be proved from association alone.

Whole-genome association. In the 1970s and early 1980s, the availability of DNA-based polymorphisms led to the application of LD to the hunt for causative genes in diseases such as cystic fibrosis and Huntington disease. Later in the 1980s, the possibility existed to use sets of restriction fragment length polymorphisms to provide LD maps of regions of the genome or eventually to obtain association data across the whole genome. Limitations imposed by marker numbers, typing efficiencies and extent of LD prevented this from being applied systematically, but, with more recent marker development and improvement in typing technology, such approaches now seem more realistic. The availability of large sets of SNPs throughout the genome will permit LD mapping to be an increasingly important component for regional and whole-genome association strategies in mapping the genes of human disease.

With the recent compilation of the draft human genome sequence, genome-wide studies of association will soon be a reality. Because of the nature of genetic resolution in outbred families described above, linkage studies do not benefit directly from the genome sequence, other than in the provision of a broader array of informative genetic markers that can be tailored to the properties of a particular family collection. Association studies, however, are likely to benefit greatly from the genome sequence data[11] or at least from the follow-on identification of sequence variants[76]. Sequencing the human genome might alter the course (again) of the positional cloning paradigm, as genome-wide association studies are likely to define new candidate genes that require some form of verification. Linkage studies of the candidate region would provide compelling evidence for genuine LD.

Regardless of the role of linkage to complement genome-wide association studies, all of the potential problems with association studies outlined above remain crucial, and, indeed, will be magnified according to the density of the scan employed[24,77,78]. Given a high

Box 5 | Proposed guidelines for association studies

- Replication of allelic association should be common practice. Such replication should eliminate the need for subgroup analysis in at least one association population. Linkage data and functional assays provide supportive and desirable evidence for valid association.
- Optimal control population selection should limit the impact of population substructures. Where possible, several control populations should be used and individually selected to maximize similarity to the disease population. In general, prospective studies provide better controls than case–control studies.
- Strategies that use family-based controls and identify linkage and association can be valuable, but are inefficient in their use of available information. Positive outcomes provide meaningful data, but a lack of statistical significance is meaningless.
- The power of a study is influenced by a host of factors, such as the effect size, local patterns of LD, allele frequency differences between marker and trait loci, and allelic and genetic heterogeneity. Sample sizes should be chosen to assume suboptimal conditions, such as weak effects, rare alleles and incomplete LD.
- Multiple testing, either with multiple markers or independent phenotypes, will produce false positives under nominal significance thresholds. All tests considered or conducted should be reported in association results, even if nonsignificant. The multiple-testing effect will be exacerbated by publication bias.
- Selection of controls sets the range of the allele frequencies in a population, particularly those related to ethnic diversity. Consideration of different sampling panels might help refine regions of LD and provide insight into LD patterns and history.

density of markers, many sampling concerns shift from statistical power to the inflation of false-positive rates caused by testing very large numbers of markers. The markers available for study will almost always outnumber the actual size of the sample examined: more than 1.5 million markers are already available in the public domain[1,2], but few human sample collections match this size. This situation of many data points and few observations highlights the need to develop methods that take into account the interdependence of genomic data. This problem resembles that recently faced by biologists working on microarray expression data. Development of such methods, coupled with realistic calculations of false-positive rates for highly correlated genomic data, are vital for the success of genome-wide association studies (BOX 4).

Limitations of association strategies

Despite the potential for improving our ability to detect disease genes, the inconsistency of association data is still a strong feature of this approach. The reasons underlying this problem need careful consideration before the technique is applied more widely.

The most common explanation for 'spurious' association (that is, association without linkage) is population stratification. Despite this commonly being used as an explanation for non-replicable associations there are few actual examples to support this assumption[7], suggesting that this problem has been overemphasized in genetic studies and that other factors are likely to be more important. For example, the importance of overinterpreting marginal findings and of publication bias has been underemphasized. There have been large numbers of association studies conducted that, coupled with a possible publication bias for positive results, might account for much of the apparent unreliability of this strategy. Attempts to reduce the effects of population stratification will have no effect on this type of error. Consequently, methodological development, which at present focuses on family designs to reduce stratification detection, should be refocused on issues such as multiple testing, multi-locus association, background linkage disequilibrium levels and large-scale study design.

A further source of study error involves the relatively small effect of many of the genetic factors that contribute to disease. As with many environmental risks, relative risks arising from genetic variation might be small. Such small effects might also be exacerbated by locus heterogeneity in disease studies, although this is often used as an excuse to explain marginally significant or non-replicable results. By undertaking association studies with modest sample sizes, association studies that reach publication tend to overestimate the size of the genetic effect. These small studies are prone to random error and often provide wide confidence intervals for any significant conclusions[33]. Coupled with frequent post-hoc subgroup analysis of heterogeneity, as popularized in the linkage domain[67,79] to separate linked and unlinked samples at specific marker loci, it is clear why this approach can lead to unreproducible or spurious outcomes.

Much can be learned from the experience of the epidemiology community, which has been applying this methodology over many years in its attempts to understand the relationship between environment and disease, particularly where the effects are small and the disease is heterogeneous[80]. These lessons are widely applicable to genetic association studies and are widely encompassed in the unified field of genetic epidemiology[81,82].

Despite the concern over false-positive findings, a lack of detecting genuine effects might be more commonplace and therefore more important, owing to several confounding limitations. One such limitation is the variability in LD throughout the genome. Markers in a region might show variable patterns of LD, reflecting their history and selection[83]. Consequently, markers close to a functional DNA variant might show less or more LD than markers further away; it is currently not possible to predict which will be strongest with another adjacent functional variant[84]. As a result, associations with some markers might not be identified in a region containing a disease-mediating DNA variant, whereas associations at adjacent markers are convincingly detected. This can lead to inconsistencies between studies. Even where a marker in LD is used, the LD is seldom complete, diluting the genetic effect being assayed. Characterization of background levels of LD, including correlations with genomic features such as G+C levels, repetitive elements, predicted or known recombination hotspots, and many others[85], would be of great use in this regard.

An additional limitation concerns interpretation. Given the relative paucity of current understanding of the mechanisms of action of complex trait loci, a plausible biological argument can be constructed for virtually any associated allele. In the absence of other information, for example linkage evidence, association replica-

tion or functional assays, such arguments are extremely difficult or impossible to prove or disprove. This is particularly difficult when the allele or gene becomes a 'candidate' after association with disease is initially detected. This difficulty places strong demands on the framework of statistical inference used to characterize the initial association. It also emphasizes the need for further understanding of the background levels of LD. At present, it is essentially impossible to consider the veracity of a new report of complex disease association in the light of past knowledge because so few 'real' associations are known.

A further complexity arises if many mutations in a gene can create the same phenotype. Each such variant will have its own genetic ancestral heritage, with different ancestral haplotypes and non-random allelic association. Such effects could substantially reduce the power to detect an association between a phenotype and any specific allele[4,6].

All these limitations have created problems in establishing robust criteria for undertaking association studies. Nevertheless, a few simple guidelines might help (BOX 5), as association strategies will inevitably command more attention in the next few years.

The human genome project is providing access to very large numbers of SNPs[1,2], recent engineering advancements are yielding high-throughout genotyping technologies[28], and population-based case–control or prospective cohorts are being initiated[86]. All of this will yield vast amounts of association data, ranging from whole-genome association studies to large sets of candidate gene association or local LD mapping data. The primary challenges will be to appreciate the complexity and subtlety of multifactorial trait loci, to develop and apply study design and analysis methods that increase the likelihood of detecting real effects while minimizing statistical false positives, and to recognize the limitations inherent in any single study design or application. When properly applied and interpreted, it is likely that association will continue to provide an essential component of the expanding arsenal needed to dissect and characterize the genetic basis of common disease.

Links

DATABASE LINKS rhesus factor | HLA | ACE | cystic fibrosis | Huntington disease

FURTHER INFORMATION TDT and statistical genetics software | Cystic fibrosis foundation | Hereditary disease foundation

1. Mullikin, J. C. *et al.* An SNP map of human chromosome 22. *Nature* **407**, 516–520 (2000).
2. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
3. Drews, J. & Ryser, S. The role of innovation in drug development. *Nature Biotechnol.* **15**, 1318–1319 (1997).
4. Terwilliger, J. D. & Weiss, K. M. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* **9**, 578–594 (1998).
5. Gambaro, G., Anglani, F. & D'Angelo, A. Association studies of genetic polymorphisms and complex disease. *Lancet* **355**, 308–111 (2000).
6. Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157 (2000).
This paper is essential reading for anyone undertaking association studies of common characters. The primary aim is to elucidate the difficulties in identifying genetic loci that contribute to complex traits. The literature cited covers some necessary population genetics material.
7. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
An excellent summary of current statistical procedures and their comparative strengths and weaknesses for complex trait mapping. Very useful for comparing linkage and association and for distinguishing familial influences on discrete versus quantitative traits.
8. Schork, N. J., Cardon, L. R. & Xu, X. The future of genetic epidemiology. *Trends Genet.* **14**, 266–272 (1998).
9. Collins, F. Positional cloning moves from perditional to traditional. *Nature Genet.* **9**, 347–350 (1995).
10. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
11. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
12. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
13. Xiong, M. & Guo, S. W. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *Am. J. Hum. Genet.* **60**, 1513–1531 (1997).
14. Freimer, N. B. *et al.* Genetic mapping using haplotype, association and linkage methods suggests a locus for severe bipolar disorder (BPI) at 18q22-q23. *Nature Genet.* **12**, 436–441 (1996).
15. Hastbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
This is becoming a classic paper on using disequilibrium/haplotype data to identify disease loci. The trait studied does not reflect the common disease framework of current widespread interest, but the procedures used offer a useful model from which to start.
16. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
One of a series of key papers by these authors who compare disequilibrium measures, evaluate real data patterns to infer genome-wide marker spacing requirements, and combine population genetics principles with those of disease-gene mapping to characterize allelic association.
17. Eaves, I. A. *et al.* The genetically isolated populations of finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genet.* **25**, 320–323 (2000).
18. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
19. Nickerson, D. A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
20. Clark, A. G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
21. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
22. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
23. Templeton, A. R. *et al.* Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66**, 69–83 (2000).
24. Ott, J. Predicting the range of linkage disequilibrium. *Proc. Natl Acad. Sci.* USA **97**, 2–3 (2000).
25. Chapman, N. H. & Thompson, E. A. Linkage disequilibrium mapping: the role of population history, size, and structure. *Adv. Genet.* **42**, 413–437 (2001).
26. Fisher, R. A. The rhesus factor: a study in scientific method. *Am. Sci.* **35**, 95–103 (1947).
27. Tiwari, J. L. & Terasaki, P. I. *HLA and Disease Associations* (Springer, New York, 1985).
28. Lander, E. S. Array of hope. *Nature Genet.* **21**, 3–4 (1999).
29. Risch, N. & Teng, J. Design and analysis of linkage disequilibrium studies for complex human diseases. *Am. J. Hum. Genet.* **61**, 1707 (1997).
30. Risch, N. & Teng, J. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* **8**, 1273–1288 (1998).
31. Teng, J. & Risch, N. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* **9**, 234–241 (1999).
32. Keavney, B. Genetic association studies in complex diseases. *J. Hum. Hypertens.* **14**, 361–367 (2000).
33. Keavney, B. *et al.* Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls. International Studies of Infarct Survival (ISIS) Collaborators. *Lancet* **355**, 434–442 (2000).
The need for association studies to involve thousands of patients is clearly shown by comparing the results of a number of typical, small studies with that of a large-scale, well-controlled design. Reference 33 offers a similar example for non-insulin-dependent diabetes mellitus.
34. Altshuler, D. *et al.* The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
35. Cambien, F. *et al.* Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction. *Nature* **359**, 641–644 (1992).
36. Arnheim, N., Strange, C. & Erlich, H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proc. Natl Acad. Sci. USA* **82**, 6970–6974 (1985).
37. Barcellos, L. F. *et al.* Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* **61**, 734–747 (1997).
38. Daniels, J. *et al.* A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *Am. J. Hum. Genet.* **62**, 1189–1197 (1998).
39. Shaw, S. H., Carrasquillo, M. M., Kashuk, C., Puffenberger, E. G. & Chakravarti, A. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* **8**, 111–123 (1998).

40. Kirov, G., Williams, N., Sham, P., Craddock, N. & Owen, M. J. Pooled genotyping of microsatellite markers in parent-offspring trios. *Genome Res.* **10**, 105–115 (2000).
41. Falk, C. T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**, 227–233 (1987).
42. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *Am. J. Hum. Genet.* **52**, 506–516 (1993).
The TDT test and its immediate predecessors changed the way human genetic studies were conducted throughout the past decade. This is the original paper describing the method.
43. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. The transmission/disequilibrium test detects cosegregation and linkage. *Am. J. Hum. Genet.* **54**, 559–560 (1994).
44. Spielman, R. S. & Ewens, W. J. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* **59**, 983–989 (1996).
45. Sham, P. C. & Curtis, D. An extended transmission/disequilibrium test (TDT) for multiallelic marker loci. *Ann. Hum. Genet.* **59**, 323–326 (1995).
46. Spielman, R. S. & Ewens, W. J. A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450–458 (1998).
47. Curtis, D. Use of siblings as controls in case–control association studies. *Ann. Hum. Genet.* **61**, 319–333 (1997).
48. Martin, E. R., Kaplan, N. L. & Weir, B. S. Tests for linkage and association in nuclear families. *Am. J. Hum. Genet.* **61**, 439–448 (1997).
49. Allison, D. B. Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**, 676–690 (1997).
50. Rabinowitz, D. A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* **47**, 342–350 (1997).
51. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
52. Martin, E. R., Monks, S. A., Warren, L. L. & Kaplan, N. L. A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Am. J. Hum. Genet.* **67**, 146–154 (2000).
53. Pritchard, L. E. *et al.* Analysis of the CD3 gene region and type 1 diabetes: application of fluorescence-based technology to linkage disequilibrium mapping. *Hum. Mol. Genet.* **4**, 197–202 (1995).
54. Bennett, S. T. & Todd, J. A. Human type 1 diabetes and the insulin gene: Principles of mapping polygenes. *Annu. Rev. Genet.* **30**, 343–370 (1996).
55. Bennett, S. T. *et al.* Insulin VNTR allele-specific effect in type 1 diabetes depends on identity of untransmitted paternal allele. The IMDIAB Group. *Nature Genet.* **17**, 350–352 (1997).
56. Merriman, T. R. *et al.* Transmission of haplotypes of microsatellite markers rather than single marker alleles in the mapping of a putative type 1 diabetes susceptibility gene (IDDM6). *Hum. Mol. Genet.* **7**, 517–524 (1998).
57. Eaves, I. A. *et al.* Transmission ratio distortion at the INS-IGF2 VNTR. *Nature Genet.* **22**, 324–325 (1999).
58. Lernmark, A. & Ott, J. Sometimes it's hot, sometimes it's not. *Nature Genet.* **19**, 213–214 (1998).
59. Goring, H. H. & Terwilliger, J. D. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am. J. Hum. Genet.* **66**, 1310–1327 (2000).
60. Morton, N.E. & Collins, A. Tests and estimates of allelic association in complex inheritance. *Proc. Natl Acad. Sci. USA* 95, 11389-93 (1998).
61. Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).
62. Rommens, J. M. *et al.* Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059–1065 (1989).
63. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
64. Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
65. Martin, E. R. *et al.* SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* **67**, 383–394 (2000).
66. Martin, E. R. *et al.* Analysis of association at single nucleotide polymorphisms in the APOE region. *Genomics* **63**, 7–12 (2000).
67. Horikawa, Y. *et al.* Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175 (2000).
68. Roses, A. D. Pharmacogenetics and the practice of medicine. *Nature* **405**, 857–865 (2000).
69. Keavney, B. *et al.* Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum. Mol. Genet.* **7**, 1745–1751 (1998).
The *ACE* locus and ACE phenotype is a model quantitative system. Despite the unusually clear haplotype relationships in this gene and population, the study clearly demonstrates the difficulty in distinguishing which specific variants are responsible for phenotypic variability.
70. Moffatt, M. F., Traherne, J. A., Abecasis, G. R. & Cookson, W. O. Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. *Hum. Mol. Genet.* **9**, 1011–1019 (2000).
71. Abecasis, G. R. *et al.* Patterns of linkage disequilibrium from three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
72. Farrall, M. *et al.* Fine-mapping of an ancestral recombination breakpoint in DCP1. *Nature Genet.* **23**, 270–271 (1999).
73. Abecasis, G. R., Cookson, W. O. & Cardon, L. R. Pedigree tests of transmission disequilibrium. *Eur. J. Hum. Genet.* **8**, 545–551 (2000).
74. Todd, J. A. *et al.* Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. *Nature* **338**, 587–589 (1989).
75. Mijovic, C. H., Barnett, A. H. & Todd, J. A. Genetics of diabetes. Trans-racial gene mapping studies. *Baillieres Clin. Endocrinol. Metab.* **5**, 321–340 (1991).
76. Cardon, L. R. & Watkins, H. Waiting for the working draft from the human genome project: A huge achievement, but not of immediate medical use. *Br. Med. J.* **320**, 1221–1222 (2000).
77. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
Mathematical population genetics modelling is used to simulate background levels of linkage disequilibrium in the genome, indicating that very fine-scale maps are required for disease gene association mapping. Although hotly contested and not always supported by empirical reports, this paper clearly outlines the issues and importance of disequilibrium levels in the genome.
78. Collins, A. & Morton, N. E. Mapping a disease locus by allelic association. *Proc. Natl Acad. Sci. USA* **95**, 1741–1745 (1998).
79. Cox, N. J. *et al.* Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genet.* **21**, 213–215 (1999).
80. Risch, N. Evolving methods in genetic epidemiology. 2. Genetic linkage from an epidemiologic perspective. *Epidemiol. Rev.* **19**, 24–32 (1997).
81. Potter, J. D. At the interfaces of epidemiology, genetics and genomics. *Nature Rev. Genet.* **2**, 142–147 (2001).
82. Khoury, M. J., Beaty, T. H. & Cohen, B. H. *Fundamentals of Genetic Epidemiology* (Oxford Univ. Press, Oxford, 1993).
83. Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722 (1999).
84. Goddard, K. A., Hopkins, P. J., Hall, J. M. & Witte, J. S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216–234 (2000).
85. Majewski, J. & Ott, J. GT repeats are associated with recombination on human chromosome 22. *Genome Res.* **10**, 1108–1114 (2000).
86. Abbott, A. Manhattan versus Reykjavik. *Nature* **406**, 340–342 (2000).
87. Borecki, I. B. & Suarez, B. K. Linkage and association: basic concepts. *Adv. Genet.* **42**, 45–66 (2001).
88. Slatkin, M. Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336 (1994).
89. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer Associates, Sunderland, MA, 1997).
90. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
This paper describes the use of unlinked genetic markers to detect population stratification, with minimal mathematical complexity. The key issues of marker spacing and informativeness are evaluated in detail. Reference 94 should be read in follow-up of this paper to see how stratification can be accounted for when it is present.
91. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
92. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
93. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
94. Bacanu, S. A., Devlin, B. & Roeder, K. The power of genomic control. *Am. J. Hum. Genet.* **66**, 1933–1944 (2000).
95. Witte, J. S., Elston, R. C. & Schork, N. J. Genetic dissection of complex traits. *Nature Genet.* **12**, 355–358 (1996).

Acknowledgements

This work was supported by the Wellcome Trust and in part by a grant from the NIH (to L.R.C.). We wish to thank Dr Joe Terwilliger for critical review of this manuscript.