GENOME-WIDE ASSOCIATION STUDIES FOR COMMON DISEASES AND COMPLEX TRAITS

Joel N. Hirschhorn[*‡§] *and Mark J. Daly*[*||]

Abstract | Genetic factors strongly affect susceptibility to common diseases and also influence disease-related quantitative traits. Identifying the relevant genes has been difficult, in part because each causal gene only makes a small contribution to overall heritability. Genetic association studies offer a potentially powerful approach for mapping causal genes with modest effects, but are limited because only a small number of genes can be studied at a time. Genome-wide association studies will soon become possible, and could open new frontiers in our understanding and treatment of disease. However, the execution and analysis of such studies will require great care.

*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA.
‡Departments of Genetics and Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA.
§Divisions of Genetics and Endocrinology and Program in Genomics, Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115, USA.
||Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA.
Correspondence to J.N.H.
e-mail: joel.hirschhorn@childrens.harvard.edu
doi:10.1038/nrg1521

We have recently seen the completion of the human genome sequence[1,2], the deposition of millions of SNPs into public databases[3], rapid improvements in SNP genotyping technology and the initiation of the International HapMap Project[4]. These advances have set the stage for genome-wide ASSOCIATION STUDIES, in which a dense set of SNPs across the genome is genotyped to survey the most common genetic variation for a role in disease or to identify the heritable QUANTITATIVE TRAITS that are risk factors for disease. As yet, no comprehensive, well-powered study has been published. Proponents of genome-wide association studies suggest that such studies will identify many variants that contribute to common disease, although the size of the resulting data sets will raise significant issues of analysis and interpretation. Other researchers have called into question the necessity and usefulness of this still expensive approach[5,6].

It is therefore crucial to understand the circumstances under which genome-wide association studies might be an appropriate approach; what constitutes a well-powered, reasonably comprehensive genome-wide association study; the limitations of what such studies will be able to discover; and how to interpret their results. Here, we review the rationale for genome-wide association studies and discuss issues of their power, efficiency, comprehensiveness, interpretation and analysis. We also outline different possible approaches to such studies and examine some of the technical and analytical issues that might hinder their success.

Why genome-wide association studies?

The many possible approaches to mapping the genes that underlie common disease and quantitative traits fall broadly into two categories: CANDIDATE-GENE studies, which use either association or resequencing approaches, and genome-wide studies, which include both LINKAGE MAPPING and genome-wide association studies. The main approaches and their advantages and disadvantages are summarized in TABLE 1. In this review, we discuss these approaches and present arguments as to why genome-wide association studies might be advantageous for identifying the genetic variants associated with common disease. One fundamentally different approach, ADMIXTURE MAPPING, is not discussed here but has been described elsewhere[7–10].

Limitations of linkage studies. Genome-wide linkage analysis is the method traditionally used to identify disease genes, and has been tremendously successful for mapping genes that underlie monogenic 'Mendelian' diseases[11]. For linkage analysis to succeed, markers that flank the disease gene must segregate with the disease in families. Variants that cause monogenic disorders are

ASSOCIATION STUDY
A genetic variant is genotyped in a population for which phenotypic information is available (such as disease occurrence, or a range of different trait values). If a correlation is observed between genotype and phenotype, there is said to be an association between the variant and the disease or trait.

QUANTITATIVE TRAIT
A biological trait that shows continuous variation (such as height) rather than falling into distinct categories (such as diabetic or healthy). The genetic basis of these traits generally involves the effects of multiple genes and gene–environment interactions. Examples of quantitative traits that contribute to disease are body mass index, blood pressure and blood lipid levels.

CANDIDATE GENE
A gene for which there is evidence of its possible role in the trait or disease that is under study.

LINKAGE MAPPING
Where genes are mapped by typing genetic markers in families to identify regions that are associated with disease or trait values within pedigrees more often than are expected by chance. Such linked regions are more likely to contain a causal genetic variant.

often rare (probably because negative selection reduces the frequencies of variants that cause diseases characterized by early-onset morbidity and mortality[12,13]), so each segregating disease allele will be found in the same 10–20 cM chromosomal background within each family. Furthermore, because Mendelian diseases are, by definition, caused by highly PENETRANT variants, markers within 10–20 cM of the disease-causing alleles will co-segregate with disease status.

Genome-wide linkage analysis has also been carried out for many common diseases and quantitative traits, for which the aforementioned characteristics of Mendelian diseases might not apply. In some cases, genomic regions that show significant linkage to the disease have been identified, leading to the discovery of variants that contribute to susceptibility to diseases such as inflammatory bowel disease (IBD)[14–17], schizophrenia[18] and type 1 diabetes[19]. However, for most common diseases, linkage analysis has achieved only limited success[20], and the genes discovered usually explain only a small fraction of the overall HERITABILITY of the disease. For example, variants known to affect the risk of IBD together explain an excess risk to siblings of just over two-fold, compared with a total excess risk of ~30-fold[21], indicating that many other causal genes are yet to be discovered. The lack of success so far can be attributed to various factors. These include the low heritability of most complex traits; the inability of the standard set of microsatellite markers — which are spaced 10 cM apart — to extract complete information about inheritance[22–25]; the imprecise definition of phenotypes[26]; and inadequately powered study designs[6]. Linkage analysis using dense marker sets[22–25], larger sample sizes[27] and larger pedigrees[6] might well be more productive. However, even if statistically significant evidence of linkage is obtained, extensive candidate gene studies (see below) are still required to progress from a broad region of linkage — usually exceeding 10 cM (~10 million bases) — to the causal gene or genes within this region.

Importantly, linkage analysis is also much less powerful for identifying common genetic variants that have modest effects on disease[28–31]. Most common diseases and clinically important quantitative traits have complex architectures (reviewed in REF. 32 in the same issue of this journal), for which the phenotype is determined by the sum total of, and/or interactions between, multiple genetic and environmental factors[32]. Therefore, any individual genetic variant will generally have a relatively small effect on disease risk. The typical frequencies of variants that underlie common disease are largely unknown, but common variants (with frequencies of >1%) have been proposed to influence disease susceptibility[1,13,33,34]. As most of the sequence differences between any two chromosomes are accounted for by common variants[35,36], it is plausible that common variants might contribute to those common diseases in which susceptibility alleles might not be under intense negative selection[13]. Indeed, there are now several examples of common variants that contribute to common disease, most of which increase the risk of disease by two-fold or less when examined in large populations[16,18,37–46]. It is clear that some of these well-established disease-susceptibility alleles could never be detected by linkage analysis. For example, the Pro12Ala variant in the peroxisome proliferative activated receptor-γ gene (*PPARG*), which affects the risk of type 2 diabetes, would only be detected using linkage studies of over one million affected sib pairs[41]. Consistent with this calculation, none of the numerous genome-wide linkage studies of type 2 diabetes identified the *PPARG* region as a region of significant linkage[47,48]. Because it is reasonable to suppose that common alleles, as well as rare alleles, will contribute to common

Table 1 | **Approaches to identifying variants underlying complex traits and common diseases**

Potential advantages	Association*	Resequencing*	Linkage‡	Admixture‡	Missense SNPs‡	Association‡	Resequencing‡
No prior information regarding gene function required	–	–	+	+	+	+	+
Localization to small genomic region	+	+	–	–	+	+	+
Inexpensive	+	–	+	+	+/–	–	Prohibitive
Families not required	+	+	–	+	+	+	+
No assumptions necessary regarding type of variant involved	+	–	+	+	–	+	+
Not susceptible to effects of stratification§	–/+	–/+	+	+	–/+	–/+	–/+
No requirement for variation of allele frequency among populations	+	+	+	–	+	+	+
Sufficient power to detect common alleles (MAFs>5%) of modest effect	+	–	–/+	+	+	+	+
Ability to detect rare alleles (MAFs<1%)	–	+	+	–	–	–	+
Reasonable track record for common diseases	+	–/+	+/–	N/A	N/A	N/A	N/A
Tools for analysis available	+	+	+	+	+	+/–	–

*Candidate-gene studies. ‡Genome-wide studies. §Association and resequencing studies are immune to stratification if they use family-based designs. Symbols indicate whether the potential advantage in the left column applies completely (+), partially (+/–), weakly (–/+) or not at all (–). MAF, minor allele frequency; N/A, not yet attempted.

disease and quantitative traits, and because linkage analysis has poor power for detecting common alleles that have low penetrance, a strategic complement to linkage analysis is desirable.

Candidate-gene resequencing studies. So far, candidate-gene studies have been the only practical alternative to linkage analysis. In these hypothesis-based studies, genes are selected for further study, either by their location in a region of linkage, or on the basis of other evidence that they might affect disease risk (reviewed in REFS 30,31).

The most comprehensive analysis of candidate genes is obtained by resequencing the entire gene in patients and controls, and searching for a variant or set of variants that is enriched or depleted in disease cases. However, because such studies are still laborious and expensive, they have been largely limited to the coding regions of one or a few candidate genes, such as the small, single-exon melanocortin-4 receptor gene (*MC4R*), variants of which explain a small fraction of cases of severe, early-onset obesity[49]. In addition, properly interpreting the results might be challenging, particularly when considering rare non-coding variants (see REF. 50 for a more complete discussion of these issues). Recently, however, Cohen and colleagues have successfully applied the resequencing approach to high-priority candidate genes in which severe loss-of-function variants cause Mendelian disorders of lipid metabolism; they found that these genes also harbour less severe but still relatively rare missense variants that are associated with high, but not extreme, levels of high-density lipoprotein[51].

Association studies. Association studies using common allelic variants are cheaper and simpler than the complete resequencing of candidate genes, and have been proposed as a powerful means of identifying the common variants that underlie complex traits[28,30,31,52]. In their simplest form, association studies compare the frequency of alleles or genotypes of a particular variant between disease cases and controls. Alternative approaches include using family-based controls to avoid the potential problem of population stratification, which we discuss below.

Candidate-gene association studies have identified many of the genes that are known to contribute to susceptibility to common disease[30,31,38,53]. Such studies are greatly facilitated by using indirect LINKAGE-DISEQUILIBRIUM (LD)-based methods (described in detail below). However, candidate-gene studies rely on having predicted the identity of the correct gene or genes, usually on the basis of biological hypotheses or the location of the candidate within a previously determined region of linkage. Even if these hypotheses are broad (for example, involving the testing of all genes in the insulin-signalling pathway), they will, at best, identify only a fraction of genetic risk factors, even for diseases in which the pathophysiology is relatively well understood. When the fundamental physiological defects of a disease are unknown, the candidate-gene approach will clearly be inadequate to fully explain the genetic basis of the disease.

The genome-wide association approach. We define a genome-wide association approach as an association study that surveys most of the genome for causal genetic variants. Because no assumptions are made about the genomic location of the causal variants, this approach could exploit the strengths of association studies without having to guess the identity of the causal genes. The genome-wide association approach therefore represents an unbiased yet fairly comprehensive option that can be attempted even in the absence of convincing evidence regarding the function or location of the causal genes.

Due to the expense and labour involved, the extension of candidate-gene studies to a genome-wide approach has not, until now, been feasible. However, recent advances have moved genome-wide association studies from the futuristic to the realistic. In the sections that follow, we briefly review these advances and discuss different strategies for undertaking genome-wide association studies.

What has made these studies possible?

Genome-wide association studies require knowledge about common genetic variation and the ability to genotype a sufficiently comprehensive set of variants in a large patient sample. The dbSNP database now contains nearly 9 million SNPs, including most of the ~11 million SNPs with minor allele frequencies of 1% or greater that are estimated to exist in the human genome[54]. Importantly, genotyping technology has considerably improved and become cheaper in recent years. One recent review of SNP genotyping technology cited 'large-scale' studies that involved nearly a hundred thousand genotypes[55]; the lowest prices were generally around 0.50 US$ per genotype. By contrast, the HapMap project (discussed in more detail below) plans to include information on ~300 million genotypes, and the United States National Institutes of Health recently solicited proposals to determine an additional 600 million genotypes at 0.01 US$ per genotype. This cost is approaching that required to make genome-wide association studies feasible; once this falls to 0.001 US$ per genotype, such studies could become routine (for example, 500,000 genotypes could be typed for only 500 US$ per individual). Some of the high-throughput genotyping technologies that have been commercially developed are listed in TABLE 2.

Another crucial advance towards enabling efficient genome-wide studies is the determination of LD patterns on a genome-wide scale through the HapMap project[4], which, as discussed below, will be particularly useful for methods that use markers selected on the basis of LD.

Markers for genome-wide association studies

LD-based markers. To be useful, markers tested for association must either be the causal allele or highly correlated (in LD) with the causal allele[56,57]. Most of the genome falls into segments of strong LD, within which variants are strongly correlated with each other, and most chromosomes carry one of only a few common HAPLOTYPES[58–60]. Recently, several large genomic regions (of ~500 kb) have been comprehensively examined as

ADMIXTURE MAPPING
Predicting the recent ancestry of chromosomal segments across the genome to identify regions for which recent ancestry in a particular population correlates with disease or trait values. Such regions are more likely to contain causal variants that are more common in the ancestral population.

PENETRANCE
The proportion of individuals with a specific genotype who manifest the genotype at the phenotypic level. For example, if all individuals with a specific disease genotype show the disease phenotype, then the genotype is said to be 'completely penetrant'.

HERITABILITY
The proportion of the variation in a given characteristic or state that can be attributed to (additive) genetic factors.

LINKAGE DISEQUILIBRIUM
Correlation between nearby variants such that the alleles at neighbouring markers (observed on the same chromosome) are associated within a population more often than if they were unlinked.

HAPLOTYPE
A sequential set of genetic markers that are present on the same chromosome.

Table 2 | **Selected commercially available high-throughput genotyping platforms**

Company	Method of allele discrimination	Method of detection	Number of assays detected simultaneously
Third Wave	PCR, cleavase	Fluorescence; plate reader	1 (multiplexed 100-fold at PCR stage only)
Sequenom	PCR, primer extension	Mass spectrometry	7–12
ABI	PCR, primer extension	Fluorescence; gel electrophoresis	48
Illumina	Oligo ligation, generic PCR	Fluorescence; tags on beads	1,536
Parallele	Gap closure, generic PCR	Fluorescence; tags on array	10,000
Affymetrix	Generic PCR, hybridization	Fluorescence; hybridization to array	10,000–100,000
Perlegen	PCR, hybridization	Fluorescence; hybridization to array	100,000+

part of the Encyclopedia of DNA Elements (ENCODE) project. This project involved the resequencing of 96 chromosomes to ascertain all common variants, and the genotyping of all SNPs that are either in the dbSNP database or that were identified by resequencing. These studies strongly confirm the patterns of long segments of strong LD that were seen in earlier studies (S. Gabriel, D. Altshuler and M.J.D., personal communication; data available on the HapMap website).

These studies have shown that most of the roughly 11 million common SNPs in the genome have groups of neighbours that are all nearly perfectly correlated with each other — the genotype of one SNP perfectly predicts those of correlated neighbouring SNPs. One SNP can thereby serve as a proxy for many others in an association screen. Once the patterns of LD are known for a given region, a few TAG SNPs can be chosen such that, individually or in multimarker combinations (haplotypes), they capture most of the common variation within the region[60,61] (FIG. 1). A proportionally higher density of variants must be typed to comprehensively survey the fraction of the genome that shows low LD.

On the basis of previous studies[58–60,62] and initial HapMap data (P. DeBakker, D. Altshuler and M. J. Daly, personal communication), a few hundred thousand well-chosen SNPs should be adequate to provide information about most of the common variation in the genome; a larger number of tag SNPs is likely to be required in African populations (and those with very recent origins in Africa), because these populations generally contain more variation and less LD[60,63]. The precise number of tag SNPs needed is yet to be determined, and will depend on the methods used to select SNPs, the degree of long-range LD between blocks and the efficiency with which SNPs in regions of low LD can be tagged[52,64]. Various algorithms have been proposed for selecting tag SNPs[61,65–70]; the optimal method will depend partly on which of the many methods for searching for associations is employed (using haplotypes, single markers, multiple markers and so on). A full review of the statistical methods available for selecting tag SNPs and for finding associations with disease is beyond the scope of this manuscript; however, Wang and colleagues discuss some aspects of these in the same issue of this journal[32].

The missense approach. As a consequence of the high proportion of missense mutations among the alleles that underlie Mendelian disorders, Botstein and Risch have proposed that association studies should be focused on missense SNPs[71]. Because a typical gene contains one or two missense SNPs[35,36,72], this strategy would require the genotyping of only 30,000–60,000 SNPs. Of course, identifying all common missense SNPs would require a substantial effort — for example, the bidirectional resequencing of 300,000 exons in 48 individuals would entail nearly 30 million reads — but if sequencing costs continue to drop[73], this could, in theory, be accomplished in the near future.

However, there are reasons to question the rationale that underlies this approach. By definition, causal alleles for monogenic disorders are highly penetrant and often lead to severe phenotypes. Accordingly, these alleles often cause severe changes in protein function, and the spectrum of disease alleles usually includes not only missense mutations but also nonsense mutations, severe splicing mutations and insertion or deletion mutations, which can induce frameshifts (although even for Mendelian disorders, an appreciable fraction of mutations are outside the coding region). Clearly, these mutations are often subject to negative selection. By contrast, the alleles that underlie complex traits have more subtle effects on disease risk and might be more likely to include non-coding regulatory variants with a modest impact on expression. In addition, given the modest effects of these alleles on disease risk and the late-onset of many common diseases, the causal alleles are far less likely to be subject to strong negative selection and might therefore comprise different types of variants to those that underlie Mendelian disorders.

In support of the missense variant approach, Botstein and Risch[71] have pointed out that the small list of common variants that have been reliably associated with common disease include a large proportion of missense variants[53]. However, this argument is undermined by a significant ASCERTAINMENT BIAS; until recently, missense variants have been preferentially discovered and preferentially tested for association with disease, so the true proportion of causal variants that involve missense changes cannot be estimated. Furthermore, SNPs in coding regions are implicitly accepted as 'the answer' when an association with a missense variant is detected,

TAG SNPs
Single nucleotide polymorphisms that are correlated with, and therefore can serve as a proxy for, much of the known remaining common variation in a region.

ASCERTAINMENT BIAS
A consequence of collecting a nonrandom subsample with a systematic bias, so that results based on the subsample are not representative of the entire sample.

often without the functional scrutiny that is required for a SNP in a non-coding region, and often despite the presence of many nearby variants that might be equally or more strongly associated with disease. Indeed, one of the missense variants that has been shown to be associated with complex disease — the Thr17Ala polymorphism in the gene encoding cytotoxic T-lymphocyte-associated protein 4 (*CTLA4*) — is reliably associated with autoimmune disease only because it is in strong LD with a regulatory polymorphism in a non-coding region, which is more strongly associated with disease and is therefore more likely to be causal[44].

Nevertheless, some missense variants have been reliably associated with complex disease and, as a group, missense variants are more likely to have functional consequences. Therefore, the genome-wide testing of large collections of missense variants is likely to remain a productive approach. However, given our current lack of knowledge about common disease risk alleles, it remains unclear what fraction of these would be discovered even by a comprehensive survey of missense polymorphisms.

New methods are emerging that might help recognize variants that affect gene function without affecting the encoded amino-acid sequence. By comparing the human and mouse genomes[74], it was shown that a significant amount of non-coding DNA is highly conserved[75]. This indicates that conserved non-coding regions are often functionally important — a hypothesis that has been supported experimentally[75–80]. Polymorphisms in these non-coding regions could also have an important role in the genetics of biomedical traits. Indeed, a modification of the missense approach to include SNPs in these conserved regions has also been proposed[71]. However, the large number of additional SNPs required would sacrifice the efficiency of the missense approach and would result in studies that are similar in scale to the indirect LD approach.

A convenience-based approach. A third approach to choosing markers for genome-wide association studies is to select variants on the basis of logistical considerations, such as the ease and cost of genotyping. Such a set of variants will be less efficient per variant for surveying the genome for disease alleles than a set that is based on LD or functional considerations, but will nevertheless achieve some degree of coverage of the genome. However, for some sets of variants, the coverage is so poor that calling them 'genome-wide' is misleading. The least comprehensive of such so-called genome-wide association studies are linkage studies that are converted into association studies by looking for associations between disease and the 400–1,000 microsatellites that are typed in linkage studies. Even under the optimistic assumption that testing a single microsatellite for association completely surveys variation in a surrounding 50-kb block of LD (blocks are on average ~20 kb (REF. 60), so this is also optimistic), such a study would cover 20–50 Mb — 1–3% of the genome or less — and cannot truly be considered a genome-wide association study.

A proposed alternative approach is to type a few SNPs in or near the coding region of each gene[81,82]. This method, like all association approaches, only surveys those variants that have been chosen for genotyping and those variants that are in LD with the chosen variants. Unless the LD patterns of each gene are empirically determined, even missense SNPs might well be missed using this approach, because choosing SNPs on the basis of physical proximity does not guarantee that nearby SNPs will be captured[52]. Furthermore, regulatory variants further away from a gene will almost certainly not be surveyed.

More recently, large collections of many thousands[83] (for example, the Affymetrix Centurion and ParAllele and MegAllele mapping sets) or over a million SNPs (K. Frazer and D. Cox, personal communication; see also the Perlegen Whole Genome Scanning collection) have been developed, and these can be genotyped at a significantly lower cost per SNP. The degree of coverage has not yet been published for these SNP sets, but they are likely to cover a significant fraction of the genome, even if they are less efficient per marker than an LD-based set. If the savings are large enough, the cost might be lower than with an LD-based set of markers for the same degree of genome coverage.

Before using such a set of variants, it will be important to genotype them in a well-defined set of samples (such as those used in the HapMap project), to determine how well the genome is covered and how best to supplement the set of markers, if necessary. Without such an assessment, even a large set of SNPs might seem to be genome-wide but might actually fail to survey a large amount of genomic variation. The ideal set of markers would be chosen with regard to LD and would be amenable to genotyping using the most cost-effective method. At present, published data and data that are emerging from the HapMap project indicate that although 100,000 markers (1 every 30 kb of the genome) would provide a prodigious amount of data, it is far from a complete scan of the genome and might only provide an adequate proxy for fewer than 50% of common variants (I. Pe'er and M.J.D., personal communication). A million randomly selected SNPs (or a few hundred thousand that have been optimally selected on the basis of LD) seem to provide much more

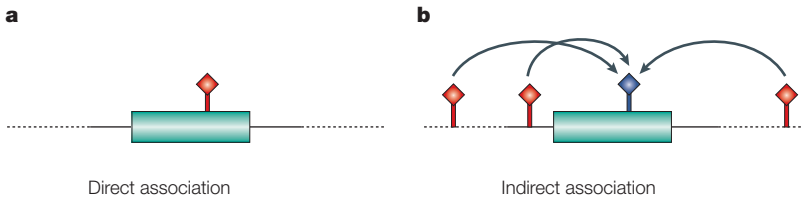


Figure 1 | **Testing SNPs for association by direct and indirect methods. a** | A case in which a candidate SNP (red) is directly tested for association with a disease phenotype. For example, this is the strategy used when SNPs are chosen for analysis on the basis of prior knowledge about their possible function, such as missense SNPs that are likely to affect the function of a candidate gene (green rectangle). **b** | The SNPs to be genotyped (red) are chosen on the basis of linkage disequilibrium (LD) patterns to provide information about as many other SNPs as possible. In this case, the SNP shown in blue is tested for association indirectly, as it is in LD with the other three SNPs. A combination of both strategies is also possible.

MULTIPLE-HYPOTHESIS TESTING
Testing more than one hypothesis within an experiment. As a result, the probability of an unusual result from within the entire experiment occurring by chance is higher than the individual *p*-value associated with that result.

BONFERRONI CORRECTION
The simplest correction of individual *p*-values for multiple-hypothesis testing: $p_{corrected} = 1 - (1 - p_{uncorrected})^n$, where n is the number of hypotheses tested. This formula assumes that the hypotheses are all independent, and simplifies to $p_{corrected} = np_{uncorrected}$ when $np_{uncorrected} << 1$.

ODDS RATIO
A measure of relative risk that is usually estimated from case-control studies.

Box 1 | Can rare alleles be detected by association methods?

Both the frequency and penetrance of causal alleles affect the statistical power to detect these alleles; power increases with increasing frequency and increasing penetrance. This indicates that a more constructive and intuitive measure than either of these factors is a single parameter, such as the amount of variation in a phenotype that can be explained by the genetic variant in question. If such as measure is used, rare, highly penetrant alleles and common, low-penetrance alleles are on an equal footing. The power to detect an allele therefore depends on what is ultimately the most relevant measure of a genetic variant's contribution: the proportion of the phenotypic variance in the population that is explained by a particular variant. This means that rare variants with modest effects will be difficult to detect by any method because they explain only a trivial fraction of the variance in a trait.

However, rare alleles might also be more difficult to detect by association for other reasons. Even if rare alleles have strong effects, they might be difficult to detect by association methods because they are less well represented in SNP databases and because tag SNP approaches are currently designed to tag common SNPs (usually with frequencies >5%). However, population-genetic considerations indicate that most rare alleles with frequencies <5% are likely to have arisen relatively recently (because old alleles tend to either disappear or become common), so there will have been less time for recombination and mutation to disrupt the haplotype on which they arose. Therefore, rare variants are expected to be on single, long haplotypes, as has been observed[131]. Recently, Cutler and colleagues have proposed an exhaustive search of all haplotypes that could greatly increase power to detect rare variants with strong effects[132]. In addition, Rioux and colleagues recently showed that the alleles of *CARD15* with frequencies of <5% that contribute to inflammatory bowel disease could have been detected indirectly with haplotypes composed of common variants[133]. Of course, rare variants with strong effects, such as those in *CARD15*, should also be detectable by well-powered linkage analyses, which should be attempted for common diseases in advance of, or in conjunction with, a genome-wide association study.

complete coverage. However, it is important to note that none of these marker sets will be optimal for detecting the effects of rare variants with frequencies of 1% or less. Nevertheless, there is hope that haplotypes of common variants will be sufficient to capture modestly rare disease alleles in the 1–5% frequency range (BOX 1).

Strategies to increase efficiency

In this section, we discuss strategies that have been proposed for implementing a whole-genome association study that has adequate power in the context of MULTIPLE-HYPOTHESIS TESTING, while minimizing the amount of genotyping required. These strategies include multi-stage designs to minimize sample sizes, the use of special populations and the pooling of samples.

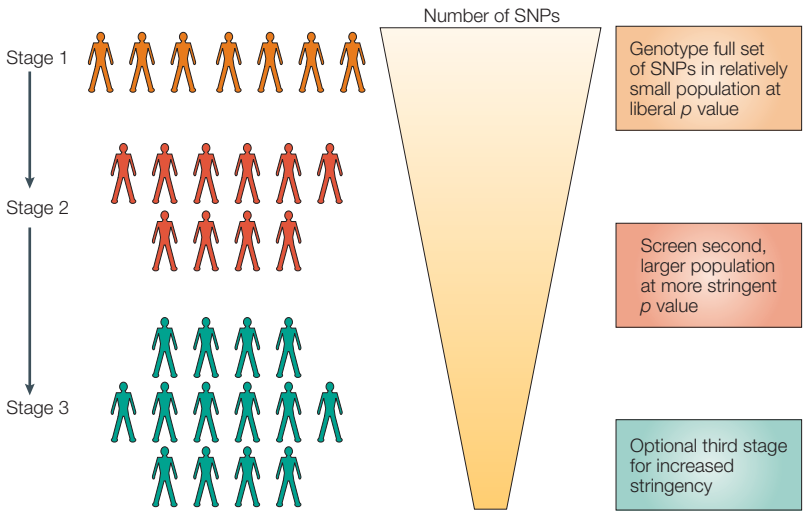
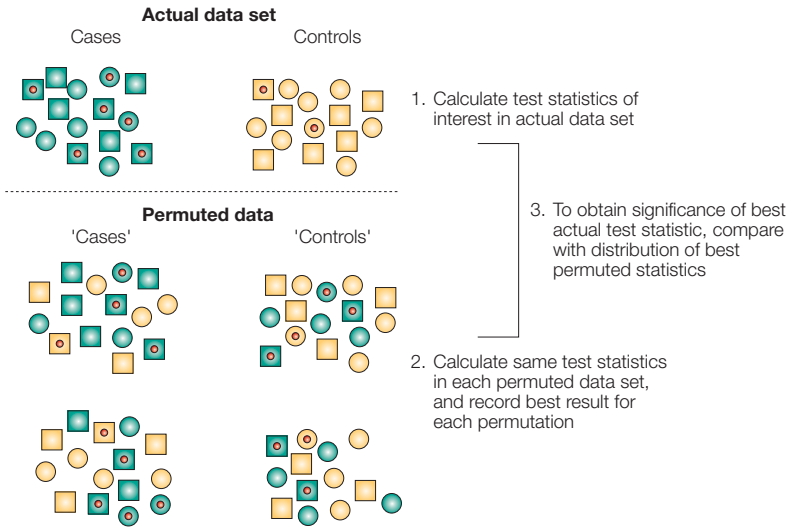


Figure 2 | **Using a multistage approach to minimize sample sizes.** In association studies, a multistage approach can reduce the amount of genotyping required, without sacrificing power. In stage 1, the full set of SNPs is genotyped in a fraction of samples, and a liberal *p*-value threshold is used to identify a subset of SNPs with putative associations. In the second and possibly third stages, the SNPs identified from the first stage are re-tested in populations that are larger or of a similar size. The results of this can then be used to distinguish the few true-positive associations identified in stage 1 from the many false-positive results that occur by chance.

Selecting an appropriate sample size. The most obvious limitation of genome-wide studies is the high cost and significant effort required to genotype hundreds of thousands of SNPs per individual. Because of this high cost, there is pressure to limit the sample size, with a consequent reduction in power. However, because variants that contribute to complex traits are likely to have modest effects, large sample sizes are crucial. The sample sizes required are further increased by the large number of hypotheses that are tested in a genome-wide association study, because *p*-values must be corrected for multiple-hypothesis testing. Risch and Merikangas[28] propose that a *p*-value of 5×10^{-8} (equivalent to a *p*-value of 0.05 after a BONFERRONI CORRECTION for 1 million independent tests) is a conservative threshold for declaring a significant association in a genome-wide study.

To understand the consequences of this threshold, consider an allele with a frequency of 15% and an ODDS RATIO of 1.25 (similar to that of the *PPARG* Pro12Ala variant associated with type 2 diabetes). For such a variant, even assuming that the causal SNP (or another SNP that serves as a perfect proxy) has been typed, nearly 6,000 cases and 6,000 controls are required to provide 80% statistical power to detect associations with a *p*-value of 5×10^{-8}. For 500,000 independent SNPs, this sample size would require 6 billion genotypes, which would be prohibitively costly. Alternatively, a more liberal *p*-value threshold could be used, but to achieve 80% power for even a nominally significant *p*-value of 0.05 for a variant such as *PPARG* Pro12Ala, 1,200 cases and 1,200 controls — or 1.2 billion genotypes — would be required. Sample sizes smaller than this risk missing

Box 2 | **Permutation testing**



When evaluating potentially interesting results with very low *p*-values, it is necessary to determine how often they would arise by chance if the study were repeated and if there were no true-positive findings. This can be achieved empirically by permutation testing, as illustrated in the figure. Green squares and circles represent men and women with the disease, respectively; yellow squares and circles represent controls; and red circles indicate the presence of a potential disease-susceptibility allele. All of the statistics of interest are calculated both in the actual data and in many different permuted data sets. No biologically meaningful association should be observed in the permuted data because the case and control labels — the phenotype values — have been shuffled. Therefore, the collection of the best test statistics observed in each permuted data set — that is, the results that show the greatest apparent association by chance — represents a null distribution with which the best actual test statistic can be compared. This can be used to obtain a true estimate of statistical significance that is corrected for multiple-hypothesis testing.

variants that are similar, in terms of frequency and risk, to variants that have already reliably been shown to affect disease susceptibility. Indeed, because the early stages of gene discovery (which we are still in with respect to common disease) are biased towards the detection of stronger effects, even larger sample sizes might be required to detect alleles with more modest effects. However, despite the benefits of permitting a smaller sample size, using a relaxed *p*-value threshold of 0.05 also guarantees that 5% of all genotyped SNPs will be 'associated' by chance. For a 500,000-SNP study, this translates into a rather useless list of 25,000 false-positive associations, within which are buried a few genuine causal alleles. Therefore, a liberal *p*-value threshold requires follow-up studies to distinguish false-positives from real associations.

A multi-stage approach. A simple procedure for overcoming these problems is the use of a two- or three-stage screening process, in which a more modest threshold for 'passing' markers as positive is used during the evaluation of the initial scan of the genome for association[41,84,85] (FIG. 2). Formal statistical considerations for such a strategy can be found in REFS 84,85; here we describe the general principle. A threshold is set that maintains the power to detect loci that explain only a small fraction of the phenotypic variance (thereby losing little power compared with a study that screens the entire genome in all samples), while bearing in mind that a large but reasonably controlled number of false-positive results will also pass this threshold. This generous threshold would also be set so that markers that do not meet it are unlikely to achieve significance in the whole sample and can therefore be safely discarded. All markers that pass the threshold are then tested in a second, independent population sample, which is similar in size or larger than the initial population. This second stage is carried out for only a small fraction of the markers that were contained in the initial screen, so there is a large efficiency gain. It might be better to genotype these markers using a different method, as different technologies are better suited for SNP sets of different sizes. This would also further minimize the chance that false associations that arise as a result of technical genotyping artefacts will be repeated. The joint thresholds from the first two stages can be set so that fewer than 5% of studies would produce a double-positive, taking all of the tests into account. A third stage could also be used for even greater stringency.

How should the thresholds used in these analyses be defined? One option is to use the standard Bonferroni correction. However, with a high density of markers, significant LD between many markers, and redundancy between single markers and the multimarker haplotypes that might also be tested for association, the assumption of independence among tests is strongly violated. Therefore, the Bonferroni correction will generally be punitively conservative, requiring inappropriately low *p*-values (and therefore inappropriately large sample sizes).

Many alternative strategies for defining significance thresholds have been proposed; we feel that permutation testing (BOX 2) offers a good solution to empirically assessing the probability of having observed a particular

Box 3 | **Founder populations**

Founder populations are those that have been recently derived — 100 or fewer generations ago — from a limited pool of individuals, and such populations have been proposed to be advantageous for studying multigenic diseases. For rare alleles with population frequencies that are less than the reciprocal of the effective number of founding chromosomes (such as the alleles that cause most single-gene disorders), there is a strong advantage to using founder populations. This is because populations with a sufficiently small and recent origin have repeatedly been shown to carry such rare mutations on a single chromosome, which can be readily dated to the founding of that population. The populations that have been used most successfully in such studies include non-geographically isolated groups such as Finnish people, French-Canadians and Ashkenazi Jews.

However, theory and early data indicate that such mapping advantages might be quite modest in the search for alleles that are associated with complex disease, which might have higher population frequencies. In these populations, it has been repeatedly found that rare mutations are present on a long shared haplotype, which arises because one (or a small number) of original founding chromosomes carried the mutation in question, and as a result, most modern-day members of that population who bear the mutation are likely to carry that particular chromosome for a long distance. As the allele is completely or largely located on a single chromosome at the founding of the population, the length of the shared haplotype depends on the time since the founding event over which recombination has acted to disrupt that chromosome; that is, it depends on the age of the population. For more common alleles, the situation will generally be more complicated, as they will enter even a small founder population so many times that the length of shared haplotypes around these alleles will be indistinguishable from that of the much larger ancestral population[134]. Indeed, studies of common variation in several of these more commonly used founder populations[135] show little or no difference in the LD patterns of common genetic variants. Therefore, these populations do not seem to provide a significant reduction in the labour required to perform a genome-wide association study.

Of course, more isolated populations will be more homogeneous and therefore might have the advantage of a more consistent environment. It at least seems to be no more difficult to detect associations in isolated populations than in more diverse populations, so isolated populations (or populations that are homogeneous for other reasons) might very well be recommended on other grounds. Several particularly isolated European and Pacific Island populations have been reported to have modestly extended general LD even around common alleles when compared with neighbouring populations[136]. So, further exploration of this idea might be warranted for populations with particularly small numbers of founders that contributed substantially to the current gene pool.

result by chance. There are also other less computationally intensive methods, including FREQUENTIST approaches for estimating experiment-wide significance[86–89] and BAYESIAN approaches for assessing the likelihood that an association is genuine[90]. Importantly, genuine associations can achieve thresholds that survive correction for multiple testing if enough samples are genotyped (either as a single study or as a combined analysis of several studies). For example, the association of the variable number of tandem repeats (VNTR) mutation in the insulin gene (*INS*) with type 1 diabetes[91], the association between *PPARG* Pro12Ala and type 2 diabetes[48] and a number of other associations, have overall *p*-values that survive even conservative genome-wide corrections for multiple testing[28].

Founder populations and pooled samples. Other methods have been proposed that increase the efficiency of association studies, most notably the use of FOUNDER POPULATIONS, which reduce the number of markers that need to be genotyped, and of pooled samples, which reduce the number of samples genotyped. Founder populations offer powerful advantages for efficiently localizing genes that underlie Mendelian disorders, but might provide less of an advantage for common diseases. The use of such populations in association studies is described in BOX 3.

Pooled samples, in which equal amounts of DNA from multiple individuals are mixed into a single well before genotyping, have the potential to markedly reduce the amount of genotyping required in whole-genome association studies (reviewed in REF. 92), perhaps up to 30-fold[93]. However, this requires the extremely accurate determination of small differences in allele frequency to detect alleles with modest effects, while still using high-throughput, low-cost genotyping platforms. To test haplotypes for association, accurate estimates of haplotype frequencies can, in theory, be reconstructed from allele-frequency data for multiple markers[43,92,93]. However, in practice, this might be difficult without either extremely accurate estimates for each marker or genotyping of multiple redundant markers, which would reduce the cost savings. Furthermore, variants with pure recessive effects might be more difficult to identify using this strategy: under a recessive model, the difference in genotype frequencies between cases and controls will be more pronounced than the difference in allele frequencies, but pooled genotyping only measures allele frequencies.

One other important limitation of pooling strategies arises from the fact that if a genome-wide association study is undertaken, at great effort and expense, it will often be desirable to genotype individuals with many different measured phenotypes, so that the information obtained from the genotype data can be maximized. Indeed, family-based and haplotype-based tests of association for quantitative traits have greatly increased the number of phenotypes that can be examined[94–101]. However, to study multiple phenotypes using a pooling strategy, a different pool must be made for each phenotype, thereby reducing the cost savings. In addition — and this might be of crucial importance — for some diseases, more complex analyses of gene–gene or gene–environment interactions will not be possible without individual genotype data. Nevertheless, because of the potentially dramatic cost savings, large-scale empirical trials of several different pooled genotyping technologies might be valuable, including direct comparisons with individual genotyping data on the same samples to assess sensitivity and specificity.

Avoiding false-positive associations

In a study involving hundreds of thousands of markers, minimizing false positives is essential. Sources of false-positive associations can be divided into three main categories: statistical fluctuations that arise by chance and result in in low *p*-values (which are likely to occur when testing multiple hypotheses); underlying systematic biases due to study design; and technical artefacts. The issue of false-positives that result from multiple-hypothesis testing is best addressed using robust criteria for declaring significant associations, such as those mentioned above. Here, we discuss systematic biases and technical causes of false-positive associations.

FREQUENTIST
A statistical approach for assessing the likelihood that a hypothesis is correct (such as an association being valid), by assessing the strength of the data that supports the hypothesis and the number of hypotheses that are tested.

BAYESIAN
A statistical approach that assesses the probability of a hypothesis being correct (for example, whether an association is valid) by incorporating the prior probability of the hypothesis and the experimental data supporting the hypothesis.

FOUNDER POPULATIONS
Populations that that have been derived from a limited pool of individuals within the last 100 or fewer generations.

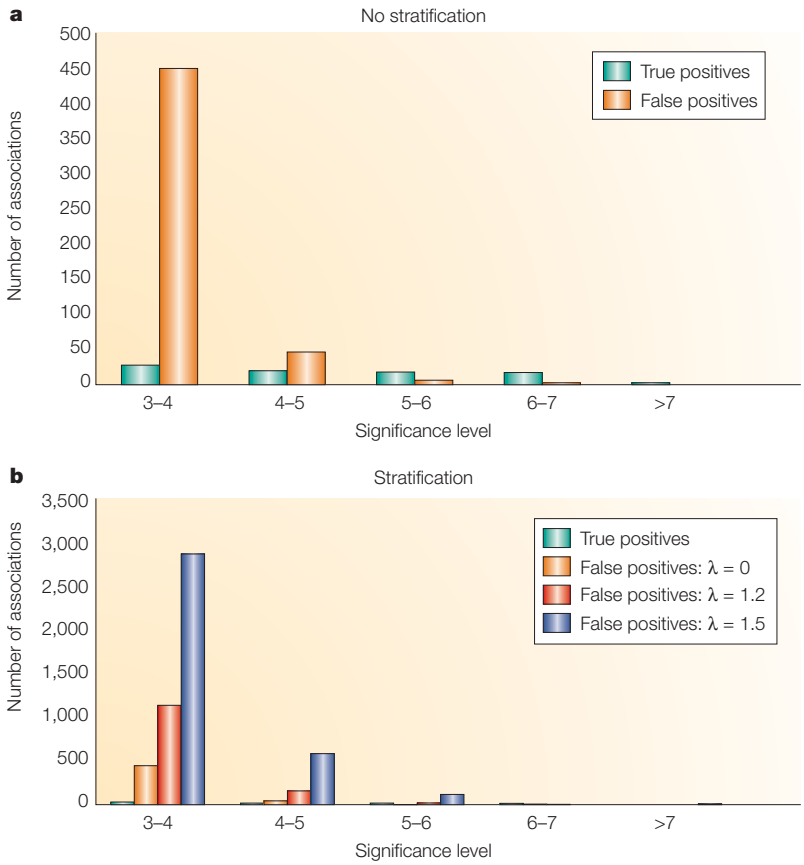


Figure 3 | **Effects of population stratification in whole-genome association studies.** A simulation of the effect of population stratification on the ratio of true-positive and false-positive results in a genome-wide association study. **a** | The expected number of false-positive and true-positive associations detected in a simulated genome-wide association study at different significance levels (shown as the negative log of *p*-values). Results are shown for a simulated study of 1,000 cases and 1,000 controls, for which 500,000 independent tests are carried out but only 100 represent true associations, with allele frequencies of 15% and odds ratios ranging from 1.2 to 1.5. **b** | The same scenario is simulated under two different levels of mild stratification, $\lambda = 1.2$ and $\lambda = 1.5$ (that is, modelling degrees of stratification such that the typical chi-square statistic for association is inflated by either 1.2 or 1.5; see REF. 103). These levels of λ for studies of 1,000 cases and controls are plausible on the basis of recent empirical studies[106]. We modelled the effect of stratification on false-positives by simulating association results under the scenario of no stratification and no true effects and then multiplying the simulated chi-square statistics by a factor of λ (see REF. 104).

ADMIXTURE
Combining two or more populations into a single group. This has implications for studies of genotype–disease associations if the component populations have different genotypic distributions.

DISCORDANT SIB STUDY
A family-based association approach that uses only sibs who are phenotypically discordant (that is, different). Like the transmission disequilibrium test, this approach is immune to population stratification.

Bias due to population stratification. The most widely discussed source of systematic bias is population stratification due to ethnic ADMIXTURE. Population stratification is the presence of multiple subgroups within a population that differ in disease prevalence (or average trait value, for quantitative traits). This can lead to the over-representation of one or more subgroups among the individuals chosen as disease cases in association studies. If a genetic marker has different frequencies in the different subgroups, false-positive associations can ensue.

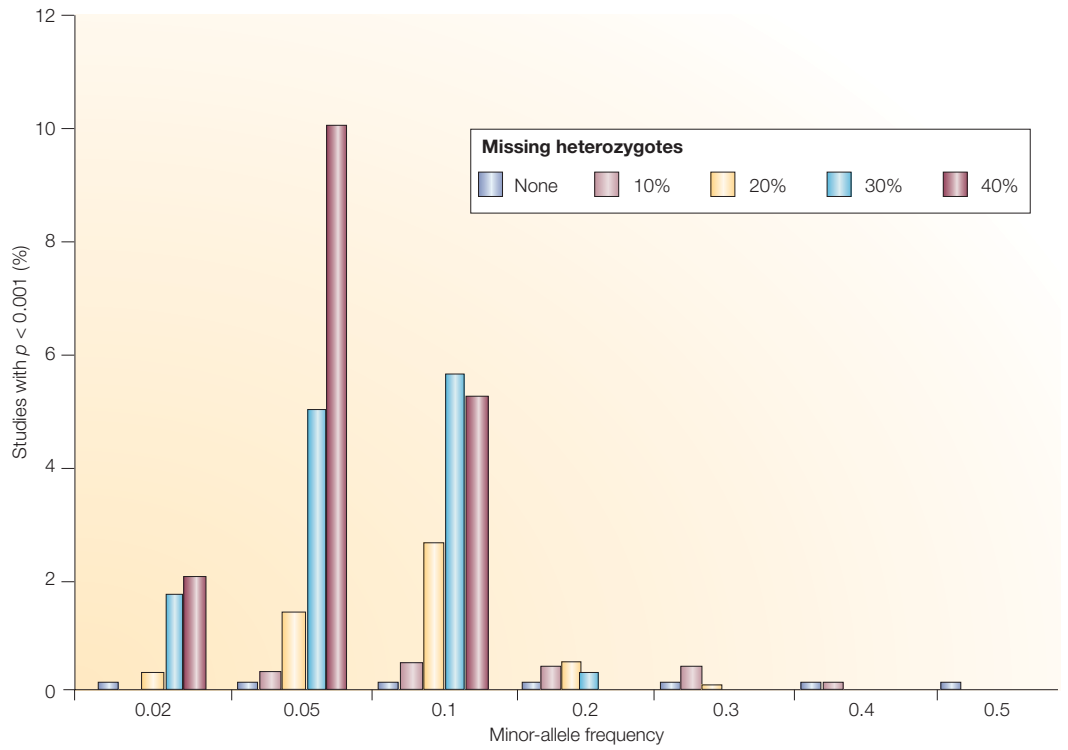
Techniques have been developed to detect[102–104] and correct[103,105,106] for population stratification by typing dozens of unlinked markers. Whether 'well-matched' association studies (that is, matched by self-described ethnicity) are subject to stratification is controversial[107–110], but until recently there has been little empirical data on this subject. Two research groups have recently reported the results of typing unlinked markers in well-matched case-control studies[106,111]. The consensus from these studies is that large-scale stratification is unlikely in well-matched populations[111]. However, mild stratification is difficult to rule out[106], and can be shown to exist in well-matched studies of historically admixed populations if enough markers are typed[106] or if the right phenotype is chosen[112]. Mild stratification might also exist even in less admixed populations at a level that can be a significant problem when looking for alleles with modest effects on disease[106,112] (FIG. 3). Fortunately, the large number of markers typed in a genome-wide association study will permit the extremely precise assessment of stratification, allowing for either the appropriate upward correction of *p*-values (with a concurrent loss of power) or the re-matching of cases and controls on the basis of genotypes at a large number of random loci[105]. Rematching samples will also sacrifice power, as not all samples will be analyzed in the final set, and the loss of power will depend on the severity of the stratification. Ideally, re-matching would take place in advance by typing a smaller number of markers in a large pool of potential controls; markers that provide information regarding ancestry might be particularly useful for this[113].

Another way of avoiding stratification is to use family-based samples. This approach has several theoretical advantages: as well as being immune to stratification[114], these samples can be used to determine whether an allele has different effects on disease when it is inherited maternally or paternally[115], and DISCORDANT SIB designs[116–118] can control for the effects of shared environment. Furthermore, more complex family-based designs are possible[119] that might allow combined association and linkage analysis[120], and family-based association tests have also been developed for quantitative traits[94–98]. However, pure sibship-based association studies are underpowered relative to case-control studies[107,116,117], and the requirement for living parents might introduce an age-of-onset bias towards younger patients for diseases that usually arise late in life. Furthermore, family-based samples are often much more difficult to collect, particularly if larger pedigrees are sought. Finally, the most commonly used family-based design, the TRANSMISSION DISEQUILIBIRIUM TEST (TDT; see REF. 114) is susceptible to technical artefacts (see below).

Bias due to technical artefacts. Because a large number of markers is typed in a genome-wide association study, even unusual technical artefacts are likely to arise occasionally. These are less likely to cause false-positives in case-control studies, because genotyping errors or missing data should affect cases and controls equally. The exception to this rule occurs when cases and controls are not genotyped in an identical manner, which can result from obvious methodological flaws such as genotyping cases and controls on separate days or on separate plates, or from more subtle issues concerning laboratory methods.

By contrast, the TDT is more susceptible to false-positive results from laboratory difficulties. The potential contribution of genotyping errors to false-positive

Box 4 | **Missing genotypes and false-positive transmission disequalibrium tests**



To estimate how missing heterozygote genotypes influence the number of false-positive results in transmission disequalibrium tests (TDTs), pedigrees with a total of 1000 parent–offspring trios were simulated, and 1,000 different sets of genotypes were generated for each set of trios under a model of no association. Heterozygote genotypes were randomly removed at different rates, and TDT statistics[114] were calculated using only complete trios. The percentage of *p*-values <0.001 — that is, the percentage of false-positives that would be generated in a real study — was recorded for each level of missing heterozygotes. This simulation was carried out for alleles with different minor allele frequencies. The results of the simulations are plotted in the graph. For each minor allele frequency, the percentage of false-positive results that would be obtained ($p<0.001$) is shown under scenarios in which 0%, 10%, 20%, 30% and 40% of heterozygote genotypes are missing. (With no missing data, the percentage of times the *p*-value fell below 0.001 was 0.1%, as expected under the null distribution.) Studies where the parental genotype counts violated Hardy–Weinberg equilibrium at $p<0.01$ or where the total number of failed genotypes exceeded 10% were not counted as false-positives, as these would be detected and eliminated by standard quality-control measures.

In real studies, false-positive TDT associations can ensue, even in the presence of standard quality-control measures, for example if heterozygote genotypes are harder to identify correctly than other genotypes. As illustrated in the figure, this effect is greatest with relatively rare alleles for which quality-control measures (such as checking for violation of Hardy–Weinberg equilibrium) are less effective, but the alleles are common enough that there are sufficient numbers of informative trios to produce small *p*-values.

TRANSMISSION DISEQUILIBRIUM TEST
A family-based test for association that is immune to population stratification. The transmission of alleles from heterozygous parents to affected offspring is compared to the expected 1:1 ratio.

HARDY–WEINBERG EQUILIBRIUM
The binomial distribution of genotypes in a population, such that frequencies of genotypes AA, Aa and aa will be p^2, 2pq, and q^2, respectively, where p is the frequency of allele A, and q is the frequency of allele a. Hardy–Weinberg equilibrium applies in a population when there are no factors such as migration or admixture that cause deviations from p^2, 2pq and q^2.

TDT associations has been described previously[121], as have possible corrective measures, including tests that take errors into account, with some accompanying loss of power[122]. The error rates considered in these studies are generally greater than those observed with most current genotyping technologies, and would usually be detected by the occurrence of apparent inheritance errors in families. However, these studies nevertheless reaffirm the importance of corroborating genotype data that show putative associations. Less widely appreciated is the fact that missing data can also result in false-positives if samples with a particular genotype are more likely not to be classified during genotyping; many genotyping methods have lower success rates for heterozygotes, so this scenario is not unrealistic.

We have examined the impact of missing data on the TDT, and find that this can be an important cause of false-positive associations, even when reasonable quality-control measures are used. As discussed in BOX 4, the impact of missing heterozygotes is most important with alleles that are rare enough to avoid violation of HARDY–WEINBERG EQUILIBRIUM, but are still common enough to achieve low *p*-values. False-positives are also observed under other scenarios, especially missing homozygotes of the rare allele (J.N.H., unpublished results). Therefore, data completeness is important for TDTs, especially for rarer alleles.

A complete analysis of the impact of missing data is beyond the scope of this review, but the implication is that for family-based designs, genotyping technologies

that are not only accurate but also have low failure rates will be crucial for avoiding false-positive associations. In general, before declaring an association to be significant, particularly with TDT analyses, genotyping should be repeated to obtain accurate and complete data; compared with the effort of a genome-wide association screen, the effort of improving data quality for a few potential associations is minimal. Despite these potential artefacts and other associated difficulties, family-based approaches might still be of significant value because they avoid stratification and might allow the use of more powerful statistical methods.

Analysing gene–gene interactions

Gene–gene interactions — EPISTASIS — are thought to have an important role in complex traits, but the analysis of how these interactions contribute to complex disease is likely to be challenging for some time to come. In an excellent review, Cordell identifies several reasons why establishing the biological importance of interactions that have been identified statistically might be nearly impossible[123]. Beyond this, with respect to genome-wide association studies, the sample sizes that we will be able to study in the next few years will not support the massive number of hypotheses that are involved in even a two-dimensional screen (testing the association of all pairs of markers). Therefore, fully powered, unconstrained scans for epistasis that account for multiple-hypothesis testing might not be possible in the near future.

Fortunately however, efficient procedures for detecting interactions might not require boundless searches of the data. In standard models of pure biological epistasis (for example, if there is only a phenotypic effect when specific alleles in two genes are present together), there will nonetheless be detectable associations for at least one of the variants; that is, one can still observe a main effect that does not require the consideration of interactions with the other factor. Therefore an effective scan for epistasis could involve simply searching for modest individual effects and then either querying for interactions among the set of positive markers or rescanning, taking into account potential interactions with the markers that have main effects. The latter conditional analysis is likely to have an important role, regardless of epistasis; scans that are conditional on known positive results are also more powerful for detecting other independent effects, as the variance explained by the major loci has been controlled for, thereby enhancing the signal from the minor contributors.

Several studies, involving both simulated and real data, have revealed that there are only limited advantages to screens using more complex models that incorporate interactions between loci[124–126]. However, we acknowledge that more unusual models of interaction can be postulated that result in no main effects for either variant; in such cases, several statistical approaches can be used, for example, MULTIFACTOR-DIMENSIONALITY REDUCTION[127] (see REF. 128 for a more detailed discussion of these methods).

There are theoretical and empirical reasons to think that epistasis is important in complex disease. In theory, if individual alleles explain only small fractions of variance, fewer are required to explain the heritability of common phenotypes if interactions between them are important. Empirical evidence of epistasis is provided by repeated observations of the genetic background modifying the effect of transgenes, knockouts and spontaneous mutations in mice. Nevertheless, recent studies[45,125,129] indicate that, in some cases, many primarily additive, independent factors might define the heritability of common phenotypes. This indicates that initial screens for main effects in genome-wide association studies are likely to be successful in many cases, without considering epistasis in the initial analysis.

What has been done so far?

No truly genome-wide association study has yet been carried out, although relatively comprehensive studies using pooled samples will undoubtedly be reported in the near future, and comprehensive studies based on individual genotypes are unlikely to be far behind. The largest published study using individual genotyping was accomplished by Nakamura and colleagues[130], although this was carried out using small samples. By genotyping over 50,000 SNPs in a sample of just under 100 cases and controls, and following up initial results in a replication panel, they identified a strong potential association between myocardial infarction and variation in the lymphotoxin-α gene. Although this was a large project involving many millions of genotypes, the low power of the initial screening sample means that the rates of both false-negatives[38] and false-positives[90] in this study were probably high. Furthermore, it is not clear what fraction of the genome was surveyed by this set of SNPs. Therefore, the performance of genome-wide association methods has not yet been assessed.

Conclusions and future directions

Several objectives need to be met before genome-wide association studies become truly practical. First, a set of SNPs must be chosen that comprehensively captures the common variation across the genome. Accumulating the data necessary to choose such SNPs is one of the main goals of the human HapMap project, which is due to be completed in the next 2 years. Methods for selecting such SNPs, and for using them efficiently for tests of association, are being developed and refined. Other proposed large sets of markers should be similarly assessed to determine how completely they survey variation across the genome. It is crucially important that the cost of genotyping continues to decrease. Finally, standardized criteria for establishing significance (perhaps based on permutation testing) are needed.

Before numerous expensive genome-wide association studies are attempted, we suggest that pilot experiments should be used to test the merits of this approach. These could include, for example, the application of one or more of the approaches described above to survey variation over a small fraction of the genome. This would ideally be carried out in a population where multiple phenotypes have been measured, and/or in a genomic region that has convincing linkage to the phenotypes of

EPISTASIS
In statistical genetics, this term refers to an interaction of multiple genetic variants (usually at different loci) such that the net phenotypic effect of carrying more than one variant is different than would be predicted by simply combining the effects of each individual variant (mathematically, this means that the gene–gene interaction is significant).

MULTIFACTOR-DIMENSIONALITY REDUCTION
An approach that attempts to reduce the number of tests required to search for interactions between multiple variables.

interest, to maximize the chances of finding true associations. Pilot experiments would also provide insights into how carefully investigators must design their studies to avoid false-positives.

Looking further into the future, we note that the most comprehensive approach towards understanding complex disease would be complete genome resequencing in a large population of cases and controls. This approach would not be limited by the choice of candidate genes, it would cover the complete spectrum of coding and non-coding variants and, unlike genome-wide association studies, it would be able to test both rare and common variants for roles in disease. Unfortunately, this approach is not close to becoming feasible. However, if a major breakthrough in sequencing technology made complete genome sequencing rapid and affordable, this approach would be the most thorough, if perhaps the most challenging to interpret. Unfortunately, over the past few years, sequencing technologies have remained fundamentally unchanged, and it is the automation and refinement of existing methods that has led to a reduction in cost[73]. Current costs are still four orders of magnitude higher than would be required for affordable whole-genome sequencing. Therefore, although new technologies are being explored[73], waiting for whole-genome sequencing to become a reality would ignore the increasingly feasible genome-wide association approach.

Association studies that are genuinely genome-wide offer great promise; in the near future, we will be able to efficiently and comprehensively test common genetic variation across the genome for a role in common disease and complex traits. On the basis of initial successes in candidate-gene association studies that represent only a tiny fraction of the genome, more comprehensive genome-wide association studies should greatly advance our understanding of the genetic basis of common diseases and complex traits.

1. International human genome sequencing consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
4. Gibbs, R. A. *et al.* The international HapMap project. *Nature* **426**, 789–796 (2003).
A description of the HapMap project, which will empirically determine LD patterns across the human genome, allowing the efficient selection of SNPs for genome-wide association studies.
5. Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157 (2000).
6. Blangero, J. Localization and identification of human quantitative trait loci: king harvest has surely come. *Curr. Opin. Genet. Dev.* **14**, 233–240 (2004).
7. McKeigue, P. M. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am. J. Hum. Genet.* **63**, 241–251 (1998).
8. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
9. Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G. & McKeigue, P. M. Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* **74**, 965–978 (2004).
10. Zhu, X., Cooper, R. S. & Elston, R. C. Linkage analysis of a complex disease through use of admixed populations. *Am. J. Hum. Genet.* **74**, 1136–1153 (2004).
11. Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853–855 (2001).
12. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
13. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
14. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
15. Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
16. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
17. Stoll, M. *et al.* Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nature Genet.* **36**, 476–480 (2004).
18. Stefansson, H. *et al.* Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.* **71**, 877–892 (2002).
19. Nistico, L. *et al.* The *CTLA-4* gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Hum. Mol. Genet.* **5**, 1075–1080 (1996).
20. Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* **69**, 936–950 (2001).
21. Daly, M. J. & Rioux, J. D. New approaches to gene hunting in IBD. *Inflamm. Bowel Dis.* **10**, 312–317 (2004).
22. Evans, D. M. & Cardon, L. R. Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am. J. Hum. Genet.* **75**, 687–692 (2004).
23. Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. *Hum. Mol. Genet.* **13**, 1943–1949 (2004).
24. John, S. *et al.* Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet.* **75**, 54–64 (2004).
25. Middleton, F. A. *et al.* Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am. J. Hum. Genet.* **74**, 886–897 (2004).
26. Levy, D. *et al.* Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. *Hypertension* **36**, 477–483 (2000).
27. Cox, N. J. *et al.* Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families. *Am. J. Hum. Genet.* **69**, 820–830 (2001).
28. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
A discussion of the power of association studies versus linkage studies for common alleles of modest effect, also anticipating the requirement to take multiple-hypothesis testing into account in genome-wide association studies.
29. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
30. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
31. Tabor, H. K., Risch, N. J. & Myers, R. M. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Rev. Genet.* **3**, 391–397 (2002).
32. Wang, W. Y. S., Barratt, B. J., Clayton, D. G & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
33. Harris, H. *The Principle of Human Biochemical Genetics* 211–242 (American Elsevier Publishing Company, New York, 1970).
34. Chakravarti, A. Population genetics — making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
35. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
36. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
37. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nature Genet.* **29**, 306–309 (2001).
38. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182 (2003).
A meta-analysis of association studies between common variants and common diseases, which indicates that a fraction (but much fewer than half) of reported associations are correct. Modest effects are the rule, indicating the need for large sample sizes.
39. Gloyn, A. L. *et al.* Large-scale association studies of variants in genes encoding the pancreatic α-cell K_{ATP} channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes* **52**, 568–572 (2003).
40. Florez, J. C. *et al.* Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes* **53**, 1360–1368 (2004).
41. Altshuler, D. *et al.* The common *PPARG* Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
This study uses large sample sizes to demonstrate a modest but consistent association between a missense polymorphism in a candidate gene and type 2 diabetes.
42. Stefansson, H. *et al.* Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. *Am. J. Hum. Genet.* **72**, 83–87 (2003).
43. Yang, J. Z. *et al.* Association study of neuregulin 1 gene with schizophrenia. *Mol. Psychiatry* **8**, 706–709 (2003).
44. Ueda, H. *et al.* Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
By testing many variants in large samples, and using logistic regression, this study shows that a 3′ UTR variant is more strongly associated with autoimmune diseases than the previously studied missense variant in the same gene.
45. Negoro, K. *et al.* Analysis of the *IBD5* locus and potential gene–gene interactions in Crohn's disease. *Gut* **52**, 541–546 (2003).
46. Giallourakis, C. *et al.* IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of a novel association with ulcerative colitis. *Am. J. Hum. Genet.* **73**, 205–211 (2003).
47. Lindgren, C. & Hirschhorn, J. Genetics of type 2 diabetes. *Endocrinologist* **11**, 178–187 (2001).
48. Florez, J. C., Hirschhorn, J. & Altshuler, D. The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu. Rev. Genomics Hum. Genet.* **4**, 257–291 (2003).

49. Vaisse, C. *et al.* Melanocortin-4 receptor mutations are a frequent and heterogeneous cause of morbid obesity. *J. Clin. Invest.* **106**, 253–262 (2000).
50. Hirschhorn, J. N. & Altshuler, D. Once and again — issues surrounding replication in genetic association studies. *J. Clin. Endocrinol. Metab.* **87**, 4438–4441 (2002).
51. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
52. Carlson, C. S., Eberle, M. A., Kruglyak, L. & Nickerson, D. A. Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452 (2004).
A useful and clear recent review of genome-wide association studies.
53. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
54. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
55. Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet.* **2**, 930–942 (2001).
56. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
57. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
58. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
The first description of long segments of strong LD with low haplotype diversity ('haplotype blocks').
59. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
A survey of chromosome 21 that reveals long segments of LD with low haplotype diversity.
60. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
A survey of over 50 genomic regions that reveals long segments of LD with low haplotype diversity, including relatively large samples from multiple populations.
61. Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
62. Dawson, E. *et al.* A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).
63. Crawford, D. C. *et al.* Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74**, 610–622 (2004).
64. Goldstein, D. B., Ahmadi, K. R., Weale, M. E. & Wood, N. W. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* **19**, 615–622 (2003).
65. Zhang, K., Deng, M., Chen, T., Waterman, M. S. & Sun, F. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA* **99**, 7335–7339 (2002).
66. Stram, D. O. *et al.* Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55**, 27–36 (2003).
67. Ke, X. & Cardon, L. R. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19**, 287–288 (2003).
68. Weale, M. E. *et al.* Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene *SCN1A*: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73**, 551–565 (2003).
69. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
70. Halldorsson, B. V. *et al.* Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res.* **14**, 1633–1640 (2004).
71. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33** Suppl. 228–237 (2003).
A proposal to focus on missense SNPs in the search for the variants that underlie common disease.
72. Cambien, F. *et al.* Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.* **65**, 183–191 (1999).
73. Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nature Rev. Genet.* **5**, 335–344 (2004).
74. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
75. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
The identification of functional regulatory sequences using evolutionary conservation.
76. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100–109 (2001).
77. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
78. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
79. Frazer, K. A. *et al.* Noncoding sequences conserved in a limited number of mammals in the *SIM2* interval are frequently functional. *Genome Res.* **14**, 367–372 (2004).
80. Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. *Nature Rev. Genet.* **5**, 456–465 (2004).
81. Buetow, K. H. *et al.* High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc. Natl Acad. Sci. USA* **98**, 581–584 (2001).
82. De La Vega, F. M. *et al.* New generation pharmacogenomic tools: a SNP linkage disequilibrium map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. *Biotechniques* (Suppl.), 48–50, 52, 54 (2002).
83. Matsuzaki, H. *et al.* Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14**, 414–425 (2004).
84. van den Oord, E. J. & Sullivan, P. F. False discoveries and models for gene discovery. *Trends Genet.* **19**, 537–542 (2003).
85. Lowe, C. E. *et al.* Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun.* **5**, 301–305 (2004).
86. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125**, 279–284 (2001).
87. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
88. Dudbridge, F. & Koeleman, B. P. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am. J. Hum. Genet.* **75**, 424–435 (2004).
89. Nyholt, D. R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.* **74**, 765–769 (2004).
90. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.* **96**, 434–442 (2004).
A Bayesian perspective on the interpretation of association studies, which emphasizes the negative impact of low prior probabilities and inadequate power on the likelihood that an association is valid.
91. Barratt, B. J. *et al.* Remapping the insulin gene/*IDDM2* locus in type 1 diabetes. *Diabetes* **53**, 1884–1889 (2004).
92. Sham, P., Bader, J. S., Craig, I., O'Donovan, M. & Owen, M. DNA Pooling: a tool for large-scale association studies. *Nature Rev. Genet.* **3**, 862–871 (2002).
93. Barratt, B. J. *et al.* Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.* **66**, 393–405 (2002).
94. Allison, D. B. Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**, 676–690 (1997).
95. Rabinowitz, D. A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* **47**, 342–350 (1997).
96. Fulker, D. W., Cherny, S. S., Sham, P. C. & Hewitt, J. K. Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259–267 (1999).
97. Abecasis, G. R., Cookson, W. O. & Cardon, L. R. Pedigree tests of transmission disequilibrium. *Eur. J. Hum. Genet.* **8**, 545–551 (2000).
98. 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).
99. Zaykin, D. V. *et al.* Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.* **53**, 79–91 (2002).
100. Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70**, 425–434 (2002).
101. Stram, D. O. *et al.* Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum. Hered.* **55**, 179–190 (2003).
102. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. **65**, 220–228 (1999).
103. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
104. Reich, D. E. & Goldstein, D. B. Detecting association in a case-control study while correcting for population stratification. *Am. J. Hum. Genet.* **20**, 4–16 (2001).
105. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
Description of software for detecting and correcting for the presence of multiple population subgroups in an association study.
106. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nature Genet.* **36**, 388–393 (2004).
107. Morton, N. E. & Collins, A. Tests and estimates of allelic association in complex inheritance. *Proc. Natl Acad. Sci. USA* **95**, 11389–11393 (1998).
108. Wacholder, S., Rothman, N. & Caporaso, N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol. Biomarkers Prev.* **11**, 513–520 (2002).
109. Thomas, D. C. & Witte, J. S. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol. Biomarkers Prev.* **11**, 505–512 (2002).
110. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
111. Ardlie, K. G., Lunetta, K. L. & Seielstad, M. Testing for population subdivision and association in four case-control studies. *Am. J. Hum. Genet.* **71**, 304–311 (2002).
112. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nature Genet.* **36**, 512–517 (2004).
113. Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422 (2003).
114. 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).
115. Frayling, T. M. *et al.* Parent-offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes* **48**, 2475–2479 (1999).
116. 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).
117. Horvath, S. & Laird, N. M. A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am. J. Hum. Genet.* **63**, 1886–1897 (1998).
118. Boehnke, M. & Langefeld, C. D. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am. J. Hum. Genet.* **62**, 950–961 (1998).
119. 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).
120. Lazzeroni, L. C. Allele sharing and allelic association I: sib pair tests with increased power. *Genet. Epidemiol.* **22**, 328–344 (2002).
121. Mitchell, A. A., Cutler, D. J. & Chakravarti, A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am. J. Hum. Genet.* **72**, 598–610 (2003).
122. Gordon, D. *et al.* A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur. J. Hum. Genet.* **12**, 752–761 (2004).
123. Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).
A discussion of epistasis, including the usefulness of searching first for main effects.

124. Leal, S. M. & Ott, J. Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. *Am. J. Hum. Genet.* **66**, 567–575 (2000).
125. Cordell, H. J., Wedig, G. C., Jacobs, K. B. & Elston, R. C. Multilocus linkage tests based on affected relative pairs. *Am. J. Hum. Genet.* **66**, 1273–1286 (2000).
126. Cordell, H. J. *et al.* Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* **158**, 357–367 (2001).
127. Ritchie, M. D. *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69**, 138–147 (2001).
128. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Rev. Genet.* **4**, 701–709 (2003).
129. Singer, J. B. *et al.* Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* **304**, 445–448 (2004).
130. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genet.* **32**, 650–654 (2002).
131. Kamatani, N. *et al.* Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. *Am. J. Hum. Genet.* **75**, 190–203 (2004).
132. Lin, S., Chakravarti, A. & Cutler, D. J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genet.* **36**, 1181–1188 (2004).
133. Vermeire, S. *et al. CARD15* genetic variation in a Quebec population: prevalence, genotype-phenotype relationship, and haplotype structure. *Am. J. Hum. Genet.* **71**, 74–83 (2002).
134. Kruglyak, L. Genetic isolates: separate but equal? *Proc. Natl Acad. Sci. USA* **96**, 1170–1172 (1999).
135. Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. & Darvasi, A. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* **12**, 771–776 (2003).
136. Kaessmann, H. *et al.* Extensive linkage disequilibrium in small human populations in Eurasia. *Am. J. Hum. Genet.* **70**, 673–685 (2002).

Acknowledgements
We thank David Altshuler, Paul DeBakker, Chris Newton-Cheh and Nick Patterson for useful discussions. J.N.H. is the recipient of a Burroughs Wellcome Career Award in Biomedical Science and a Smith Family Foundation New Investigator Award.

Competing interests statement
The authors declare no competing financial interests.

Online links

FURTHER INFORMATION
International HapMap Project: http://www.hapmap.org
dbSNP database: http://www.ncbi.nlm.nih.gov/projects/SNP
The ENCODE project: http://www.genome.gov/10005107
Par Allele Meg Allele genotyping products: http://www.parallelebio.com/products-services/genotyping-products.html
Perlegen Whole Genome Scanning: http://www.perlegen.com/science/scanning.html
Affymetrix gene chip arrays: http://www.affymetrix.com/products/arrays/specific/100k.affx
Access to this interactive links box is free online.