# A tutorial on statistical methods for population association studies

*David J. Balding*

Abstract | Although genetic association studies have been with us for many years, even for the simplest analyses there is little consensus on the most appropriate statistical procedures. Here I give an overview of statistical approaches to population association studies, including preliminary analyses (Hardy–Weinberg equilibrium testing, inference of phase and missing data, and SNP tagging), and single-SNP and multipoint tests for association. My goal is to outline the key methods with a brief discussion of problems (population structure and multiple testing), avenues for solutions and some ongoing developments.

Haplotype
A combination of alleles at different loci on the same chromosome.

Population stratification
Refers to a situation in which the population of interest includes subgroups of individuals that are on average more related to each other than to other members of the wider population.

Multiple-testing problem
Refers to the problem that arises when many null hypotheses are tested; some significant results are likely even if all the hypotheses are false.

Hardy–Weinberg equilibrium
Holds at a locus in a population when the two alleles within an individual are not statistically associated.

*Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK.*
*e-mail: d.balding@imperial.ac.uk*
doi:10.1038/nrg1916

The goal of population association studies is to identify patterns of polymorphisms that vary systematically between individuals with different disease states and could therefore represent the effects of risk-enhancing or protective alleles (BOXES 1,2). That sounds easy enough: what could be difficult about spotting allele patterns that are overrepresented in cases relative to controls?

One fundamental problem is that the genome is so large that patterns that are suggestive of a causal polymorphism could well arise by chance. To help distinguish causal from spurious signals, tight standards for statistical significance need to be established; another tactic is to consider only patterns of polymorphisms that could plausibly have been generated by causal genetic variants, given our current understanding of human genetic history[1] and evolutionary processes such as mutation and recombination. Checking for systematic errors and dealing with missing values present further challenges. Upstream of the study itself, at the study design phase, several questions need to be considered, such as: How many individuals should be genotyped? At how many markers? And how should markers and individuals be chosen?

In this article I survey current approaches to such challenges. My goal is to give a broad-brush view of different statistical problems and how they relate to each other, and to suggest some solutions and sources of further information. I look first at statistical analyses that precede association testing and then move on to the tests of association, based on single SNPs, multiple SNPs and haplotypes. I also briefly introduce adjustments to allow for possible population stratification (or population structure) and approaches to the problem of multiple testing. My hope is that those handling genetic-association data will obtain a clearer picture of the statistical issues and gain some ideas for new or modified approaches.

In this review I cover only population association studies in which unrelated individuals of different disease states are typed at a number of SNP markers. I do not address family-based association studies, admixture mapping or linkage studies (BOX 3), which also have an important role in efforts to understand the effects of genes on disease[2].

## Preliminary analyses

Data quality is of paramount importance, and data should be checked thoroughly, for example, for batch or study-centre effects, or for unusual patterns of missing data. Testing for Hardy–Weinberg equilibrium (HWE) can also be helpful, as can analyses to select a good subset of the available SNPs ('tag' SNPs) or to infer haplotypes from genotypes.

***Hardy–Weinberg equilibrium.*** Deviations from HWE can be due to inbreeding, population stratification or selection. They can also be a symptom of disease association[3], the implications of which are often under-exploited[4]. Apparent deviations from HWE can arise in the presence of a common deletion polymorphism, because of a mutant PCR-primer site or because of a tendency to miscall heterozygotes as homozygotes. So far, researchers have tested for HWE primarily as a data quality check and have discarded loci that, for example, deviate from HWE among controls at significance level $\alpha = 10^{-3}$ or $10^{-4}$. However, the possibility that a deviation from HWE is due to a deletion polymorphism[5] or a segmental duplication[6] that could be important in disease causation should now be considered before discarding loci.

Testing for deviations from HWE can be carried out using a Pearson goodness-of-fit test, often known simply as 'the $\chi^2$ test' because the test statistic has approximately a $\chi^2$ null distribution. Be aware, however, that there are many different $\chi^2$ tests. The Pearson test is easy to compute, but the $\chi^2$ approximation can be poor when there are low genotype counts, and it is better to use a Fisher exact test, which does not rely on the $\chi^2$ approximation[7–9]. The open-source data-analysis software R (see online links box) has an R genetics package that implements both Pearson and Fisher tests of HWE, and PEDSTATS also implements exact tests[9]. (All statistical genetics software cited in the article can be found at the Genetic Analysis Software website, which can be found in the online links box).

A useful tool for interpreting the results of HWE and other tests on many SNPs is the log quantile–quantile (QQ) *P*-value plot (FIG. 1): the negative logarithm of the *i*th smallest *P* value is plotted against $-\log(i / (L + 1))$, where *L* is the number of SNPs. Deviations from the $y = x$ line correspond to loci that deviate from the null hypothesis[10].

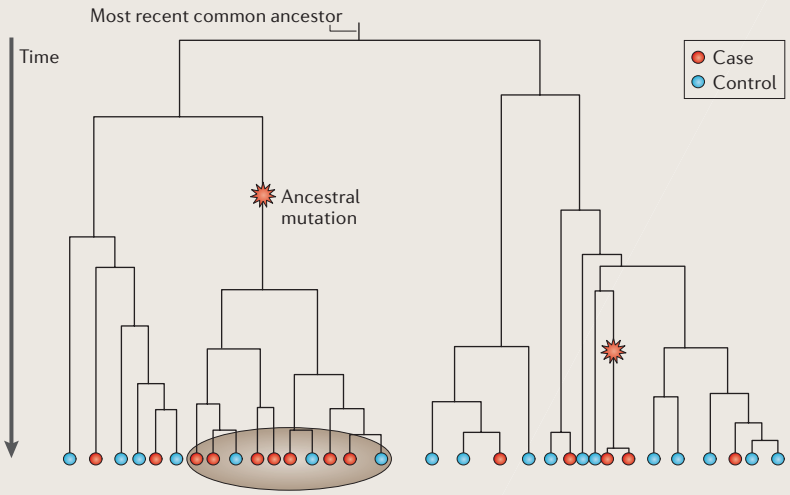
***Missing genotype data.*** For single-SNP analyses, if a few genotypes are missing there is not much problem. For multipoint SNP analyses, missing data can be more problematic because many individuals might have one or more missing genotypes. One convenient solution is data imputation: replacing missing genotypes with predicted values that are based on the observed genotypes at neighbouring SNPs. This sounds like cheating, but for tightly linked markers data imputation can be reliable, can simplify analyses and allows better use of the observed data. Imputation methods either seek a 'best' prediction of a missing genotype, such as a maximum-likelihood estimate (single imputation), or randomly select it from a probability distribution (multiple imputations). The advantage of the latter approach is that repetitions of the random selection can allow averaging of results or investigation of the effects of the imputation on resulting analyses[11].

Most software for phase assignment (see below) also imputes missing alleles. There are also more general imputation methods: for example, 'hot-deck' approaches[11], in which the missing genotype is copied from another individual whose genotype matches at neighbouring loci, and regression models that are based on the genotypes of all individuals at several neighbouring loci[12].

These analyses typically rely on missingness being independent of both the true genotype and the phenotype. This assumption is widely made, even though its validity is often doubtful. For example, as noted above, heterozygotes might be missing more often than homozygotes. What is worse, case samples are often collected differently from controls, which can lead to differential rates of missingness even if genotyping is carried out blind to case–control status. The combination of these two effects can lead to serious biases[13]. One simple way to investigate differential missingness between cases and controls is to code all observed genotypes as 1 and missing genotypes as 0, and test for association of this variable with case–control status.

***Haplotype and genotype data.*** Underlying an individual's genotypes at multiple tightly linked SNPs are the two haplotypes, each containing alleles from one parent. I discuss below the merits of analyses that are based on phased haplotype data rather than unphased genotypes, and consider here only ways to obtain haplotype data.

**Box 1 | Rationale for association studies**




Population association studies compare unrelated individuals, but 'unrelated' actually means that relationships are unknown and presumed to be distant. Therefore, we cannot trace transmissions of phenotype over generations and must rely on correlations of current phenotype with current marker alleles. Such a correlation might be generated by one or more groups of cases that share a relatively recent common ancestor at a causal locus. Recombinations that have occurred since the most recent common ancestor of the group at the locus can break down associations of phenotype with all but the most tightly linked marker alleles, permitting fine mapping if marker density is sufficiently high (say, ≥1 marker per 10 kb, but this depends on local levels of linkage disequilibrium).

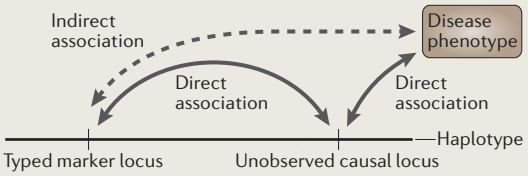
This principle is illustrated in the figure, in which for simplicity I assume haploidy, such as for X-linked loci in males. The coloured circles indicate observed alleles (or haplotypes), and the colours denote case or control status; marker information is not shown. The alleles within the shaded oval all descend from a risk-enhancing mutant allele that perhaps arose some hundreds of generations in the past (red star), and so there is an excess of cases within this group. Consequently, there is an excess of the mutant allele among cases relative to controls, as well as of alleles that are tightly linked with it. The figure also shows a second, minor mutant allele at the same locus that might not be detectable because it contributes to few cases.

Although the SNP markers that are used in association studies can have up to four nucleotide alleles, because of their low mutation rate most are diallelic, and many studies only include diallelic SNPs. With increasing interest in deletion polymorphisms[5], triallelic analyses of SNP genotypes might become more common (treating deletion as a third allele), but in this article I assume all SNPs to be diallelic.

Broadly speaking, association studies are sufficiently powerful only for common causal variants. The threshold for 'common' depends on sample and effect sizes as well as marker frequencies[90], but as a rough guide the minor-allele frequency might need to be above 5%. Arguments for the common-disease common-variant (CDCV) hypothesis essentially rest on the fact that human effective population sizes are small[1]. A related argument is that many alleles that are now disease-predisposing might have been advantageous in the past (for example, those that favour fat storage). In addition, selection pressure is expected to be weak on late-onset diseases and on variants that contribute only a small risk. Although some common variants that underlie complex diseases have been identified[91], we still do not have a clear idea of the extent to which the CDCV hypothesis holds.

Box 2 | **Types of population association study**

Population association studies can be classified into several types; for example, as follows:




**Candidate polymorphism**
These studies focus on an individual polymorphism that is suspected of being implicated in disease causation.

**Candidate gene**
These studies might involve typing 5–50 SNPs within a gene (defined to include coding sequence and flanking regions, and perhaps including splice or regulatory sites). The gene can be either a positional candidate that results from a prior linkage study, or a functional candidate that is based, for example, on homology with a gene of known function in a model species.

**Fine mapping**
Often refers to studies that are conducted in a candidate region of perhaps 1–10 Mb and might involve several hundred SNPs. The candidate region might have been identified by a linkage study and contain perhaps 5–50 genes.

**Genome-wide**
These seek to identify common causal variants throughout the genome, and require ≥300,000 well-chosen SNPs (more are typically needed in African populations because of greater genetic diversity). The typing of this many markers has recently become possible because of the International HapMap Project[32] and advances in high-throughput genotyping technology (see also BOX 5).

These classifications are not precise: some candidate-gene studies involve many hundreds of genes and are similar to genome-wide scans. Typically, a causal variant will not be typed in the study, possibly because it is not a SNP (it might be an insertion or deletion, inversion, or copy-number polymorphism). Nevertheless, a well-designed study will have a good chance of including one or more SNPs that are in strong linkage disequilibrium with a common causal variant, as illustrated in the figure: the two direct associations that are indicated cannot be observed, but if $r^2$ (see main text) between the two loci is high then we might be able to detect the indirect association between marker locus and disease phenotype.

Statistical methods that are used in pharmacogenetics are similar to those for disease studies, but the phenotype of interest is drug response (efficacy and/or adverse side effects). In addition, pharmacogenetic studies might be prospective whereas disease studies are typically retrospective. Prospective studies are generally preferred by epidemiologists, and despite their high cost and long duration some large, prospective cohort studies are currently underway for rare diseases[92,93]. Often a case–control analysis of genotype data is embedded within these studies[2], so many of the statistical analyses that are discussed in this review can apply both to retrospective and prospective studies. However, specialized statistical methods for time-to-event data might be required to analyse prospective studies[94].

Significance level
Usually denoted $\alpha$, and chosen by the researcher to be the greatest probability of type-1 error that is tolerated for a statistical test. It is conventional to choose $\alpha = 5\%$ for the overall analysis, which might consist of many tests each with a much lower significance level.

Test statistic
A numerical summary of the data that is used to measure support for the null hypothesis. Either the test statistic has a known probability distribution (such as $\chi^2$) under the null hypothesis, or its null distribution is approximated computationally.

Common-disease common-variant hypothesis
The hypothesis that many genetic variants that underlie complex diseases are common, and therefore susceptible to detection using current population association study designs. An alternative possibility is that genetic contributions to complex diseases arise from many variants, all of which are rare.

Effective population size
The size of a theoretical population that best approximates a given natural population under an assumed model. Human effective population size is often taken to mean the size of a constant-size, panmictic population of breeding adults that generates the same level of polymorphism under neutrality as observed in an actual human population.

Maximum-likelihood estimate
The value of an unknown parameter that maximizes the probability of the observed data under the assumed statistical model.

Phase
The information that is needed to determine the two haplotypes that underlie a multi-locus genotype within a chromosomal segment.

Direct, laboratory-based haplotyping or typing further family members to infer the unknown phase are expensive ways to obtain haplotypes. Fortunately, there are statistical methods for inferring haplotypes and population haplotype frequencies from the genotypes of unrelated individuals. These methods, and the software that implements them, rely on the fact that in regions of low recombination relatively few of the possible haplotypes will actually be observed in any population. These programs generally perform well[14], given high SNP density and not too much missing data. SNPHAP is simple and fast, whereas PHASE[15] tends to be more accurate but comes at greater computational cost. Recently FASTPHASE has emerged[16], which is nearly as accurate as PHASE and much faster.

True haplotypes are more informative than genotypes, but inferred haplotypes are typically less informative because of uncertain phasing. However, the information loss that arises from phasing is small when linkage disequilibrium (LD) is strong.

Note that phasing cases and controls together allows better estimates of haplotype frequencies under the null hypothesis of no association, but can lead to a bias towards this hypothesis and therefore a loss of power. Conversely, phasing cases and controls separately can inflate type-1 error rates. A similar issue arises in imputing missing genotypes.

***Measures of LD and estimates of recombination rates.*** LD will remain crucial to the design of association studies until whole-genome resequencing becomes routinely available. Currently, few of the more than 10 million common human polymorphisms are typed in any given study. If a causal polymorphism is not genotyped, we can still hope to detect its effects through LD with polymorphisms that are typed. To assess the power of a study design to achieve this, we need to measure LD. However, LD is a non-quantitative phenomenon: there is no natural scale for measuring it. Among the measures that have been proposed for two-locus haplotype data[17], the two most important are $D'$ and $r^2$.

$D'$ is sensitive to even a few recombinations between the loci since the most recent mutation at one of them. Textbooks emphasize the exponential decay over time of $D'$ between linked loci under simple population-genetic models, but stochastic effects mean that this theoretical relationship is of limited usefulness. A disadvantage of $D'$ is that it can be large (indicating high LD) even when one allele is very rare, which is usually of little practical interest.

$r^2$ reflects statistical power to detect LD: $nr^2$ is the Pearson test statistic for independence in a $2 \times 2$ table of haplotype counts. Therefore, a low $r^2$ corresponds to a large sample size, $n$, that is required to detect the LD between the markers. If disease risk is multiplicative

Box 3 | **Linkage and other approaches**

In all approaches to gene mapping, the key idea is that a disease-predisposing allele will pass from generation to generation together with variants at tightly linked loci. Linkage studies directly examine the transmission across generations of both disease phenotype and marker alleles within a known pedigree, seeking correlations that suggest that the marker is linked with a causal locus. In parametric linkage analysis[62,95], disease and marker transmission are evaluated under a specified disease model using likelihood-based statistical analyses of extended pedigrees. In nonparametric linkage analysis[96], excess allele sharing is sought in affected relatives, which avoids the need to posit a disease model.

An important advantage of linkage methods is that information is combined across families such that evidence for a causal role of a locus can accumulate even if different variants segregate at that locus in different families. Therefore, linkage analysis is appropriate when many rare variants at a locus each contribute to disease risk. However, linkage approaches can require many and/or large families to achieve satisfactory power and resolution.

There are various strategies for combining linkage with association analyses for family-based data sets. The best-known of the family-based association methods is the transmission disequilibrium test (TDT)[97], which implements a matched-pair study design by comparing alleles that are transmitted to an affected child with the untransmitted parental alleles. More general and more powerful family-based association tests are available[98,99].

Admixture mapping[100,101] has some similarities with nonparametric linkage. It can use case-only samples from a population formed by recent admixture of two or more populations with very different disease prevalences. An excess sharing among cases of an allele that is more common in the high-risk ancestral population could be a signal that the allele contributes to disease risk.

across alleles, and HWE holds, $r^2$ between a marker and a causal SNP gives the sample size that would have been required to detect the disease association by directly typing the causal SNP, relative to the sample size required to achieve the same power when typing the marker.

Both $D'$ and $r^2$ are two-locus measures; however, with dense markers it is of interest to summarize LD over a region. One approach is to compute local averages of pairwise values of $D'$ and $r^2$. Alternatively, values over a region can be illustrated diagrammatically with colours encoding different values[18,19]. LD maps[20,21] provide another solution: these fit an exponential decay function to $D'$ values, and the decay parameter provides a measure of local LD. The resulting LD unit is usually strongly correlated with underlying recombination rate, but also reflects the history of the mutations that generated the SNPs.

Fine-scale estimates of recombination rate might provide the most satisfactory solution to the problem of summarizing LD in a region because recombination is the most important biological phenomenon underlying LD. PHASE provides estimates[22], and other available software includes LDHAT[23] and HOTSPOTTER[24]. Analyses that are based on such software, and empirical studies[25,26], have shown that recombination rates are highly variable on fine scales. This is consistent with the observation that much of the human genome is 'block like'[27,28], with little or no recombination within blocks but block boundaries that are often hotspots of intense recombination.

***SNP tagging.*** 'Tagging' refers to methods to select a minimal number of SNPs that retain as much as possible of the genetic variation of the full SNP set[29–31]. Simple pairwise methods discard one (preferably that with most missing values) of every pair of SNPs with, say, $r^2 > 0.9$. More sophisticated methods can be more efficient[32], but the most efficient tagging strategy will depend on the statistical analysis to be used. In practice, tagging is only effective in capturing common variants.

There are two principal uses for tagging. The first is to select a 'good' subset of SNPs to be typed in all the study individuals from an extensive SNP set that has been typed in just a few individuals. Until recently, this was frequently a laborious step in study design, but the International HapMap Project[33] and related projects now allow selection of tag SNPs on the basis of publicly available data. The population that underlies a particular study will typically differ from the populations for which public data are available, and a set of tag SNPs that have been selected in one population might perform poorly in another. However, recent studies indicate that tag SNPs often transfer well across populations[34,35].

A secondary use for tagging is to select for analysis a subset of SNPs that have already been typed in all the study individuals. Although it is undesirable to discard available information, the amount of information lost might be small, and reducing the SNP set in this way can simplify analyses and lead to more statistical power by reducing the degrees of freedom (df) of a test[29].

## Tests of association: single SNP

I now come to testing for association, first dealing with single-SNP analyses. I will discuss case–control, quantitative (continuous) and categorical disease outcomes, starting with the simplest tests and moving on to more advanced regression-based tests[36], and also the score procedure.

***Case–control phenotype.*** Perhaps the most natural analysis of SNP genotypes and case–control status at a single SNP is to test the null hypothesis of no association between rows and columns of the 2 × 3 matrix that contains the counts of the three genotypes (the two homozygotes and the heterozygote) among cases and controls. Users have a choice between, among others, a Pearson test (2 df) or a Fisher exact test. Again, the latter is preferred: it is computationally more demanding but is implemented in R and other software.

For complex traits, it is widely thought that contributions to disease risk from individual SNPs will often be roughly additive — that is, the heterozygote risk will be intermediate between the two homozygote risks. The general tests (Pearson 2 df and Fisher) have reasonable power regardless of the underlying risks, but if the genotype risks are additive they will not be as powerful as tests that are tailored to this scenario. One way to improve power to detect additive risks is to count alleles rather than genotypes so that each individual contributes twice to a 2 × 2 table and a Pearson 1-df test can be applied. However, this procedure is not recommended[37] because it requires an assumption of HWE in cases and controls combined and does not lead to interpretable risk estimates.

Regression models
A class of statistical models that relate an outcome variable to one or more explanatory variables. The goal might be to predict further values of the outcome variable given the explanatory variables, or to identify a minimal set of explanatory variables with good predictive power.

Prospective study design
Studies in which individuals are followed forward in time and disease events are recorded as they arise. DNA and biomarker samples, and data on environmental exposures and lifestyle factors, are usually obtained at the start of the study.

Retrospective study design
Studies in which individuals are identified for inclusion in the study on the basis of their disease state. Data on previous environmental exposures and lifestyle factors are then recorded, and samples for DNA and biomarker studies might be obtained.

Time to event
Refers to data in which the time to an event of interest is recorded, such as the time from the start of the study to disease onset, if any. This is potentially more informative than simply recording case or control status at the end of the study.

Linkage disequilibrium
The statistical association, within gametes in a population, of the alleles at two loci. Although linkage disequilibrium can be due to linkage, it can also arise at unlinked loci; for example, because of selection or non-random mating.

Type-1 error
The rejection of a true null hypothesis; for example, concluding that HWE does not hold when in fact it does. By contrast, the power of a test is the probability of correctly rejecting a false null hypothesis.

Degrees of freedom
This term is used in different senses both within statistics and in other fields. It can often be interpreted as the number of values that can be defined arbitrarily in the specification of a system; for example, the number of coefficients in a regression model. It is often sufficient to regard degrees of freedom as a parameter that is used to define particular probability distributions.

Bayesian
A statistical school of thought that, in contrast to the frequentist school, holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Bayes theorem is a celebrated result in probability theory that allows one to compute the posterior distribution for an unknown from the observed data and its assumed prior distribution.

Likelihood-ratio test
A statistical test that is based on the ratio of likelihoods under alternative and null hypotheses. If the null hypothesis is a special case of the alternative hypothesis, then the likelihood-ratio statistic typically has a $\chi^2$ distribution with degrees of freedom equal to the number of additional parameters under the alternative hypothesis.

The Cochran–Armitage test[38] (also known as just the Armitage test and called within R the proportion trend test) is similar to the allele-count test. It is more conservative and does not rely on an assumption of HWE. The idea is to test the hypothesis of zero slope for a line that fits the three genotypic risk estimates best (FIG. 2).

There is no generally accepted answer to the question of which single-SNP test to use. We could design optimal analyses if we knew what proportion of undiscovered disease-predisposing variants function additively and what proportions are dominant, recessive or even over-dominant. Lacking this knowledge, researchers have to use their judgment to choose which 'horse' to back. Adopting the Armitage test implies sacrificing power if the genotypic risks are far from additive, in order to obtain better power for near-additive risks. Using the Fisher test spreads the research investment over the full range of risk models, but this inevitably means investing less in the detection of additive risks.

An intermediate choice is to take the maximum test statistic from those designed for additive, dominant or recessive effects[39]. This approach weights those three models equally but excludes possible overdominant effects. A possible modification is to give more weight to the additive-test statistics, reflecting the greater plausibility of the additive model, but to allow strong non-additive effects to be detected. A different approach is to adopt the Armitage test when the minor-allele frequency is low and the Fisher test when the counts for all three genotypes are high enough for it to have good power for non-additive models.

My emphasis on the role of the researcher's judgement hints at Bayesian approaches, in which researchers make explicit their *a priori* predictions about the nature of disease risks. Bayesian approaches do not yet have a big role in genetic association analyses, possibly because of the additional computation that they can require[40]. I expect this approach to have a more prominent role in future developments. (See Supplementary information S1 (box) for suggestions of single-SNP tests that are based on Bayes factors.)

***Continuous outcomes: linear regression.*** The natural statistical tools for continuous (or quantitative) traits are linear regression and analysis of variance (ANOVA). ANOVA is analogous to the Pearson 2-df test in that it compares the null hypothesis of no association with a general alternative, whereas linear regression achieves a reduction in degrees of freedom from 2 to 1 by assuming a linear relationship between mean value of the trait and genotype (FIG. 3). In either case, tests require the trait to be approximately normally distributed for each genotype, with a common variance. If normality does not hold, a transformation (for example, log) of the original trait values might lead to approximate normality.

Standard statistical procedures offer a hierarchy of $\chi^2_1$ tests in which the ANOVA model is compared with the linear regression model, which in turn is compared with the null model of no association. The convention is to accept the simplest model that is not significantly inferior to a more general model.

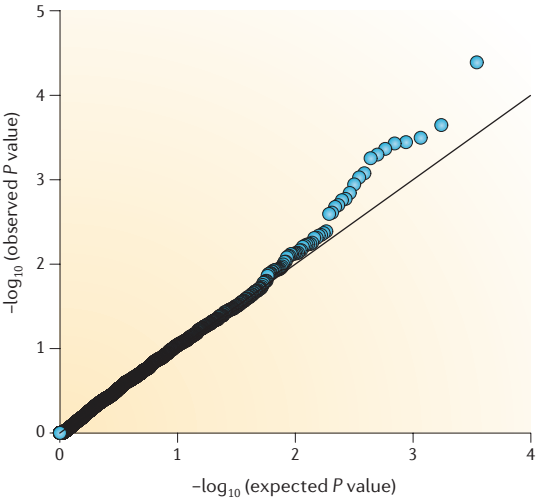



Figure 1 | **Log quantile–quantile (QQ) *P*-value plot for 3,478 single-SNP tests of association.** The close adherence of *P* values to the black line (which corresponds to the null hypothesis) over most of the range is encouraging as it implies that there are few systematic sources of spurious association. The use of the log scale helps to emphasize the smallest *P* values (in the top right corner of the plot): the plot is suggestive of multiple weak associations, but the deviation of observed small *P* values from the null line is unlikely to be sufficient to reach a reasonable criterion of significance.

***Logistic regression.*** Returning now to case–control outcomes, I consider a more advanced approach. The linear models that are outlined above for continuous traits cannot be applied directly to case–control studies, because case–control status is not normally distributed and there is nothing to stop predicted probabilities lying outside the range 0–1.

These problems are overcome in logistic regression, in which the transformation logit $(\pi) = \log(\pi / (1 - \pi))$ is applied to $\pi_i$, the disease risk of the $i$th individual. The value of logit $(\pi_i)$ is equated to either $\beta_0$, $\beta_1$ or $\beta_2$, according to the genotype of individual $i$ ($\beta_1$ for heterozygotes). The likelihood-ratio test of this general model, against the null hypothesis $\beta_0 = \beta_1 = \beta_2$, has 2 df, and for large sample sizes is equivalent to the Pearson 2-df test. Users can improve the power to detect specific disease risks, at the cost of lower power against some other risk models, by restricting the values of $\beta_0$, $\beta_1$ and $\beta_2$. For example, by requiring that the coefficients are linear, so that $\beta_1$ is half-way between $\beta_1$ and $\beta_2$, a 1-df test is obtained that is effectively equivalent to the Armitage test. Tests for recessive or dominant effects can be obtained by requiring that $\beta_0 = \beta_1$ or $\beta_1 = \beta_2$.

So far, logistic regression has not brought much that is new for single-SNP analyses. There is often a score procedure (see below) that is effectively equivalent to a logistic regression counterpart and is usually simpler and computationally faster. However, logistic regression offers a flexible tool that can readily accommodate multiple SNPs (see later section), possibly with complex epistatic and environmental interactions or covariates such as sex or age of onset.

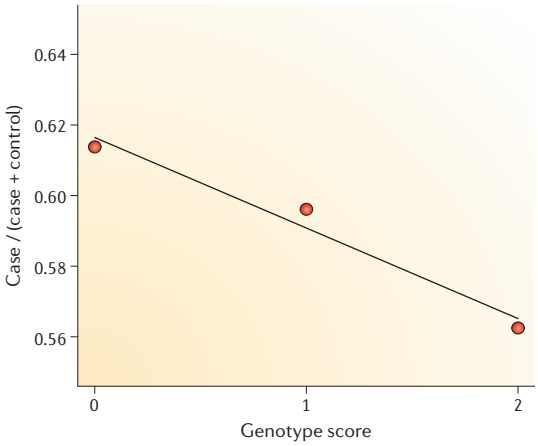



Figure 2 | **Armitage test of single-SNP association with case–control outcome.** The dots indicate the proportion of cases, among cases and controls combined, at each of three SNP genotypes (coded as 0, 1 and2), together with their least-squares line. The Armitage test corresponds to testing the hypothesis that the line has zero slope. Here, the line fits the data reasonably well as the heterozygote risk estimate is intermediate between the two homozygote risk estimates; this corresponds to additive genotype risks. The test has good power in this case but power is reduced by deviations from additivity. In an extreme scenario, if the two homozygotes have the same risk but the heterozygote risk is different (overdominance), then the Armitage test will have no power for any sample size even though there is a true association.

One potential problem with regression-based analyses is that they assume prospective observation of phenotype given the genotype, whereas many studies are retrospective: individuals are ascertained on the basis of phenotype, and genotype is the outcome variable. There is theory to show that the distinction often does not matter[41,42], but the theory does not hold in all settings, notably when missing genotypes or phase have been imputed.

***Score tests.*** There is a general procedure for generating tests that are asymptotically equivalent to likelihood-based tests: the score procedure[43]. These tests are based on the derivative of the likelihood with respect to the parameter of interest, with unknown parameters set to their null values. Both the Armitage and Pearson tests are score tests that correspond to the logistic regression models described above. The score procedure is flexible and can be adapted to incorporate covariates (such as sex or age), and to scenarios in which individuals are selected for genotyping on the basis of their phenotypes[44].

***Ordered categorical outcomes.*** In addition to binary and continuous variables, disease outcomes can also be categorical[45] — either ordered (for example, mild, moderate or severe) or unordered (for example, distinct disease subtypes). Unordered outcomes can be analysed using multinomial regression. For ordered outcomes, researchers might prefer an analysis that gives more weight to the most severely affected cases, perhaps because diagnosis is more certain or because genes that contribute to progression to the most severe state are the most important causal variants. One option is to adopt the 'proportional odds' assumption that the odds of an individual having a disease state in or above a given category is the same for all categories. Unfortunately, the score statistic under this model is complex and the equivalence of retrospective and prospective likelihoods does not apply. An alternative that does generate this equivalence is the 'adjacent categories' regression model, for which the risk of category $k$ relative to $k$–1 is the same for all $k$; the corresponding score test is a simple statistic that is a natural generalization of the Armitage test statistic.

## Dealing with population stratification

Population structure can generate spurious genotype–phenotype associations, as outlined in BOX 4. Here I briefly discuss some solutions to this problem. These require a number (preferably >100) of widely spaced null SNPs that have been genotyped in cases and controls in addition to the candidate SNPs.

***Genomic control.*** In Genomic Control (GC)[46,47], the Armitage test statistic is computed at each of the null SNPs, and $\lambda$ is calculated as the empirical median divided by its expectation under the $\chi^2_1$ distribution. Then the Armitage test is applied at the candidate SNPs, and if $\lambda > 1$ the test statistics are divided by $\lambda$. There is an analogous procedure for a general (2 df) test[48]. The motivation for GC is that, as we expect few if any of the null SNPs to be associated with the phenotype, a value of $\lambda > 1$ is likely to be due to the effect of population stratification, and dividing by $\lambda$ cancels this effect for the candidate SNPs. GC performs well under many scenarios, but it is limited in applicability to the simplest, single-SNP analyses, and can be conservative in extreme settings (and anti-conservative if insufficient null SNPs are used)[49,50].

***Structured association methods.*** These approaches[51–53] are based on the idea of attributing the genomes of study individuals to hypothetical subpopulations, and testing for association that is conditional on this subpopulation allocation. These approaches are computationally demanding, and because the notion of subpopulation is a theoretical construct that only imperfectly reflects reality, the question of the correct number of subpopulations can never be fully resolved.

***Other approaches.*** Null SNPs can mitigate the effects of population structure when included as covariates in regression analyses[50]. Like GC, this approach does not explicitly model the population structure and is computationally fast, but it is much more flexible than GC because epistatic and covariate effects can be included in the regression model. Empirically, the logistic regression approaches show greater power than GC, but their type-1 error rate must be assessed through simulation[50].

When many null markers are available, principal-components analysis provides a fast and effective way to diagnose population structure[54,55]. Alternatively, a mixed-model approach that involves estimated

Multinomial
Describes a variable with a finite number, say $k$, of possible outcomes; in the cases $k = 2$ and $k = 3$, the terms binomial and trinomial are also used.

Principal-components analysis
A statistical technique for summarizing many variables with minimal loss of information: the first principal component is the linear combination of the observed variables with the greatest variance; subsequent components maximize the variance subject to being uncorrelated with the preceding components.

kinship, with or without an explicit subpopulation effect, has recently been found to outperform GC in many settings[56]. Given large numbers of null SNPs, it becomes possible to make precise statements about the (distant) relatedness of individuals in a study so that a complete solution to the problem of population stratification — which has in the past been the cause of much concern — is probably not far away.

## Tests of association: multiple SNPs

Given *L* SNPs genotyped in cases and controls at a candidate gene that is subject to little recombination, or perhaps an LD block within a gene, we might want to decide whether or not the gene is associated with the disease and/or, given that there is association, find the SNP(s) that are closest to the causal polymorphism(s).

Analysing SNPs one at a time can neglect information in their joint distribution. This is of little consequence in the two extreme cases: when SNPs are widely spaced so as to have little or no LD between them or when almost all SNPs are typed so that any causal variant is likely to be typed in the study. In practice, most studies have SNP densities between these two extremes, in which case multipoint association analyses have substantial advantages over single-SNP analyses[57]. I first outline regression analyses of unphased SNP genotypes and then move on to haplotype-based analyses.

***SNP-based logistic regression.*** Logistic regression analyses for *L* SNPs are a natural extension of the single-SNP analyses that are discussed above: there is now a coefficient ($\beta_0$, $\beta_1$ or $\beta_2$) for each SNP, leading to a general test with *2L* df. By constraining the coefficients, tests with *L* df can be obtained. For example, a test for additive effects at each SNP is obtained by requiring that each $\beta_1 = (\beta_0 + \beta_2) / 2$. The corresponding score test, also with *L* df, is a generalization of the Armitage test, and is related to the Hotelling $T^2$ statistic[56]. Another test, with *L+1* df, uses only 1 df to capture gene-wide dominance effects[29].

Covariates such as sex, age or environmental exposures are readily included. Similarly, interactions between SNPs can be included. This conveys little benefit, and can reduce power to detect an association, if there is a single underlying causal variant and little or no recombination between SNPs[58], but it is potentially useful for investigating epistatic effects.

If the number of SNPs is large, tagging to eliminate near-redundant SNPs often increases power despite some loss of information. Alternatively, the problem of too many highly correlated SNPs in the model can be addressed using a stepwise selection procedure[59] or Bayesian shrinkage methods[60]. However, problems can arise in assessing the significance of any chosen model.

Essentially the same issues arise for a continuous phenotype; the same sets of coefficients are appropriate but they are equated to the expected phenotype value rather than the logit of disease risk.

***Haplotype-based methods.*** The multi-SNP analyses discussed above can suffer from problems that are associated with many predictors, some of which are highly correlated. A popular strategy, suggested by the block-like structure of the human genome, is to use haplotypes to try to capture the correlation structure of SNPs in regions of little recombination. This approach can lead to analyses with fewer degrees of freedom, but this benefit is minimized when SNPs are ascertained through a tagging strategy. Perhaps more importantly, haplotypes can capture the combined effects of tightly linked *cis*-acting causal variants[61].

An immediate problem is that haplotypes are not observed; instead, they must be inferred and it can be hard to account for the uncertainty that arises in phase inference when assessing the overall significance of any finding. However, when LD between markers is high, the level of uncertainty is usually low.

Given haplotype assignments, the simplest analysis involves testing for independence of rows and columns in a $2 \times k$ contingency table, where *k* denotes the number of distinct haplotypes[62]. Alternative approaches can be based on the estimated haplotype proportions among cases and controls, without an explicit haplotype assignment for individuals[63]: the test compares the product of separate multinomial likelihoods for cases and controls with that obtained by combining cases and controls. One problem with both these approaches is reliance on assumptions of HWE and of near-additive disease risk. A different approach, which leads to a test with fewer degrees of freedom, is to look for an excess sharing of haplotypes among cases relative to controls[64]. More sophisticated haplotype-based analyses treat haplotypes as categorical variables in regression analyses[65] or

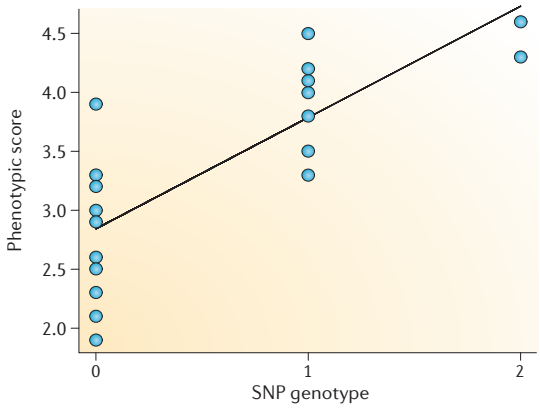
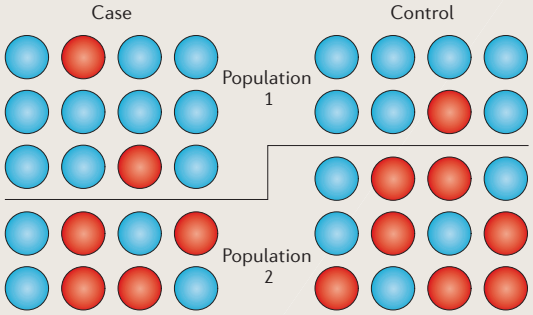



Figure 3 | **Linear regression test of single-SNP associations with continuous outcomes.** Values of a quantitative phenotype for three SNP genotypes, together with least-squares regression line. Note that here the line gives a predicted trait value for the rare homozygote (2) that exceeds the observed values, suggesting some deviation away from the assumption of linearity. Analysis of variance (ANOVA) does not require linearity of the trait means, at the cost of one more degree of freedom. Both tests also require the trait variance to be the same for each genotype: the graph is suggestive of decreasing variance with increasing genotype score, but there is not enough data to confirm this, and a mild deviation from this assumption is unlikely to have an important adverse effect on the validity of the test.

Stepwise selection procedure
Describes a class of statistical procedures that identify from a large set of variables (such as SNPs) a subset that provides a good fit to a chosen statistical model (for example, a regression model that predicts case–control status) by successively including or discarding terms from the model.

Shrinkage methods
In this approach a prior distribution for regression coefficients is concentrated at zero, so that in the absence of a strong signal of association, the corresponding regression coefficient is 'shrunk' to zero. This mitigates the effects of too many variables (degrees of freedom) in the statistical model.

Box 4 | **Spurious associations due to population structure**




The desired cause of a significant result from a single-SNP association test is tight linkage between the SNP and a locus that is involved in disease causation. The most important spurious cause of an association is population structure. This problem arises when cases disproportionately represent a genetic subgroup (population 1 in the figure), so that any SNP with allele proportions that differ between the subgroup and the general population will be associated with case or control status. In the figure, the blue allele is overrepresented among cases but only because it is more frequent in population 1.

Some overrepresented SNP alleles might actually be causal (the blue allele could be the reason that there are more cases in population 1), but these are likely to be 'swamped' among significant test results by the many SNPs that have no causal role. If the population strata are identified they can be adjusted for in the analysis[102].

Cryptic population structure that is not recognized by investigators is potentially more problematic, although the extent to which it is a genuine cause of false positives has been the topic of much debate[13,49,103,104]. There are at least three reasons for a subgroup to be overrepresented among cases:

- Higher proportion of a causal SNP allele in the subgroup;
- Higher penetrance of the causal genotype(s) in the subgroup because of a different environment (for example, diet);
- Ascertainment bias (for example, the subgroup is more closely monitored by health services than the general population, so that cases from the subgroup are more likely to be included in the study).

The first reason alone is unlikely to cause effects of worrying size[50], because of the genetic homogeneity of human populations and efforts by investigators to recruit homogeneous samples. Only the third reason is entirely non-genetic, so that there is unlikely to be a true causal variant among the strongest associations.

In fact, 'population structure' is a misnomer: the problem does not require a structured population. Indeed, populations are just a convenient way to summarize patterns of (distant) relatedness or kinship: the problem of spurious associations arises if cases are on average more closely related with each other than with controls. This insight might lead to more general and more powerful approaches to dealing with the problem.

corresponding score tests. Instead of inferring haplotypes in a separate step, ambiguous phase can be directly incorporated[66].

There are several problems with haplotype-based analyses. What should be done about rare haplotypes? Including them in analyses can lead to loss of power because there are too many degrees of freedom. One common but unsatisfactory solution is to combine all haplotypes that are rare among controls into a 'dustbin' category. How should similar but distinct haplotypes that might share recent ancestry be accounted for? Both might carry the same disease-predisposing variant but simple analyses will not consider their effects jointly and might miss the separate effects. Another problem with defining haplotypes is that block boundaries can vary according to the population sampled, the sample size, the SNP density and the block definition[67]. Often there will be some recombination within a block, and conversely there can be between-block LD that will not be exploited by a block-based analysis.

Many methods have emerged to try to overcome the problems of haplotype-based methods of analysis. These methods impose a structure on haplotype space to exploit possible evolutionary relationships among haplotypes, deal adequately with rare haplotypes and limit the number of tests that are required. One approach is to use clustering to identify sets of haplotypes that are assumed to share recent common ancestry and therefore convey a common disease risk[57,68–76]. Some of these approaches (often called cladistic) are based explicitly on evolutionary ideas or models and, for example, generate a tree that corresponds to the genealogical tree underlying the haplotypes. Others use more general haplotype-clustering strategies, but the underlying motivation is similar.

Although haplotype analysis seems to be a natural approach, it might ultimately confer little or no advantage over analyses of multipoint SNP genotypes. Even if recombination is entirely absent in a region, so that the block model applies perfectly, regression models can capture the variation without the need for interaction terms[58]. Furthermore, the widespread adoption of tagging strategies — facilitated by knowledge of LD that is obtained from the HapMap project and other sources — diminishes the potential utility of haplotype analyses. Nevertheless, haplotypes form a basic unit of inheritance and therefore have an interpretability advantage (as shown in BOX 1). Haplotype-based analyses[77] that are not restricted within block boundaries continue to hold promise for flexible and interpretable analyses that exploit evolutionary insights.

***Epistatic effects and gene–environment interactions.*** Most analyses of population association data focus on the marginal effect of individual variants. A variant with small marginal effect is not necessarily clinically insignificant: it might turn out to have a strong effect in certain genetic or environmental backgrounds, and in any case might give clues to mechanisms of disease causation.

Few researchers deny that genes interact with other genes and environmental factors in causing complex disease[78] but there is disagreement over whether tackling epistatic effects directly is a better strategy than the indirect approach of first seeking marginal effects[79,80]. The prospect of seeking multiple interacting variants simultaneously is daunting because of the many combinations of variants to consider, although this can be reduced by screening out variants that show no suggestion of a marginal effect. Both gene–gene (epistatic) and gene–environment interactions are readily incorporated into SNP-based or haplotype-based regression models and related tests[81,82]. A case-only study design[83] that looks for association between two genes or a gene and environmental exposure can give greater power.

The study of epistasis poses problems of interpretability[84]. Statistically, epistasis is usually defined in terms of deviation from a model of additive effects, but this might be on either a linear or logarithmic scale, which implies different definitions. Despite these problems, there is evidence that a direct search for epistatic effects can pay dividends[85] and I expect it to have an increasing role in future analyses.

Box 5 | **Genome-wide association studies**

The toolkit of statistical procedures for genome-wide association (GWA) studies is similar to that available for candidate genes, but there are important issues of computational and statistical efficiency as well as cost that lead to constraints on study design[87,91]. Genome-wide resequencing might not be far off, but at present typing all known variants genome-wide is unfeasible.

Fortunately, relatively few SNPs are required (approximately 300,000 in Caucasians) to capture most of the common genetic variation in a population. However, even this number implies a substantial cost and most researchers have adopted a two-stage strategy in which relatively few individuals are typed genome-wide, with the remaining individuals only typed at SNPs that seem promising from this first phase[105]. Although replication of results in different laboratories is always highly desirable, some researchers have in effect split their analysis into two phases in order to claim replication. This is undesirable because it does not achieve true replication and has an adverse effect on power compared with a joint analysis[106].

Because of the computational problems in analysing such large datasets, single-SNP tests remain the primary statistical tool used for GWA studies. Another strategy is to identify linkage disequilibrium blocks according to some criterion, and infer and analyse haplotypes within each block, while retaining for individual analysis those SNPs that do not lie within a block. Bayesian graphical models offer another computationally tractable option[107].

In the GWA setting, the computational demands of permutation procedures (see main text) can become excessive. One approach to reduce the computational burden is to perform relatively few permutations but to fit an extreme value distribution to the results and therefore extrapolate to the tail of the distribution[108]. To avoid a multiple-testing penalty for individual SNP tests, other approaches involve joint tests of groups of SNPs. For example, the sum or product of a statistic can be formed over sets of SNPs[87,109]. These approaches can detect a strong signal of association overall, even when each SNP only contributes modestly to disease effect, but strict adherence to the method does not permit identifying the most promising individual SNPs.

Using 300,000 common SNPs in a GWA study, the number of SNP pairs is about 100 billion, which leads to substantial issues of multiple testing and computational feasibility for exhaustive pairwise assessments of epistasis. However, even this number of tests is becoming feasible using logistic regression[85], and score procedures that are based on the case-only study design should be faster.

## Multiple testing

Multiple testing is a thorny issue, the bane of statistical genetics. The problem is not really the number of tests that are carried out: even if a researcher only tests one SNP for one phenotype, if many other researchers do the same and the nominally significant associations are reported, there will be a problem of false positives. The genome is large and includes many polymorphic variants and many possible disease models. Therefore, any given variant (or set of variants) is highly unlikely, *a priori*, to be causally associated with any given phenotype under the assumed model, so strong evidence is required to overcome the appropriate scepticism about an association. Although this Bayesian language provides a convenient description of the problem, the Bayesian remedy is complex because every possible disease model must be assigned a prior probability. The approach is appealing from the perspective of researchers who are engaged in the disinterested pursuit of knowledge, but less satisfactory in prescribing exacting standards for researchers who might be tempted to cut corners or exaggerate the prior plausibility of a model that is supported *a posteriori*.

The frequentist paradigm of controlling the overall type-1 error rate sets a significance level $\alpha$ (often 5%), and all the tests that the investigator plans to conduct should together generate no more than probability $\alpha$ of a false positive. In complex study designs, which involve, for example, multiple stages and interim analyses, this can be difficult to implement, in part because it was the analysis that was planned by the investigator that matters, not only the analyses that were actually conducted. However, in simple settings the frequentist approach gives a practical prescription: if $n$ SNPs are tested and the tests are approximately independent, the appropriate per-SNP significance level $\alpha'$ should satisfy $\alpha = 1 - (1 - \alpha')^n$, which leads to the Bonferroni correction $\alpha' \approx \alpha / n$. For example, to achieve $\alpha = 5\%$ over 1 million independent tests means that we must set $\alpha' = 5 \times 10^{-8}$. However, the effective number of independent tests in a genome-wide analysis depends on many factors, including sample size and the test that is carried out.

For tightly linked SNPs, the Bonferroni correction is conservative. A practical alternative is to approximate the type-1 error rate using a permutation procedure. Here, the genotype data are retained but the phenotype labels are randomized over individuals to generate a data set that has the observed LD structure but that satisfies the null hypothesis of no association with phenotype. By analysing many such data sets, the false-positive rate can be approximated. The method is conceptually simple but can be computationally demanding, particularly as it is specific to a particular data set and the whole procedure has to be repeated if the data set is altered.

Although the 5% global error rate is widely used in science, it is inappropriately conservative for large-scale SNP-association studies: most researchers would accept a higher risk of a false positive in return for greater power. The 5% value can of course be relaxed, but another approach is instead to monitor the false-discovery rate (FDR)[86,87], which is the proportion of false positive test results among all positives. Under the null hypothesis, *P* values should be uniformly distributed between 0 and 1; FDR methods typically consider the actual distribution as a mixture of outcomes under the null (uniform distribution of *P* values) and alternative

Frequentist
A name for the school of statistical thought in which support for a hypothesis or parameter value is assessed using the probability of the observed data (or more 'extreme' datasets) given the hypothesis or value. Usually contrasted with Bayesian.

(*P*-value distribution skewed towards zero) hypotheses. Assumptions about the alternative hypothesis might be required for the most powerful methods, but the simplest procedures avoid explicit assumptions. See BOX 5 for a discussion of issues that are relevant to genome-wide study designs.

The usual frequentist approach to multiple testing has a serious drawback in that researchers might be discouraged from carrying out additional analyses beyond single-SNP tests, even though these might reveal interesting associations, because all their analyses would then suffer a multiple-testing penalty. It is a matter of common sense that expensive and hard-won data should be investigated exhaustively for possible patterns of association. Although the frequentist paradigm is convenient in simple settings, strict adherence to it can be detrimental to science. Under the Bayesian approach, there is no penalty for analysing data exhaustively because the prior probability of an association should not be affected by what tests the investigator chooses to carry out.

## Conclusion

Of the vast public investment in genetic association studies in recent years, relatively little has been focused on efficient analyses of the data. There is, for example, a European Bioinformatics Institute but there is no equivalent institute for statistical genetics, the practitioners of which tend to work in relatively small groups that are scattered across institutions. Nevertheless, organized collaborations across institutions can achieve much, as shown by the achievements of the HapMap Analysis Group[33].

Progress is being made, but there is still much to be done. Ultimately, complex diseases will require complex analyses in which many variants are assessed simultaneously for their individual or joint contributions to disease risk. Fear of multiple-testing penalties should not deter researchers from making thorough analyses, but they need to deal honestly with the problem of chance associations. Bayesian regression and variable selection procedures are beginning to be developed for genome-wide microarray and genetic analyses[88,89], and they hold out promise for large-scale, simultaneous analyses of many SNPs in association studies.

To finish on an optimistic note, fear of the effects of population stratification should soon be banished: with genome-wide data a near-complete solution to the problem should be achievable, focusing directly on relatedness and not unreliable proxies such as geographical location or ethnic affiliation.


1. Jobling, M. A., Hurles, M. E. & Tyler-Smith, C. *Human Evolutionary Genetics: Origins Peoples & Disease* (Garland Science, New York, 2004).
2. Thomas, D. C. *Statistical Methods in Genetic Epidemiology* (Oxford Univ. Press, 2004).
**The best general reference for statistical methods in genetic epidemiology; for population association studies it discusses important general issues without specific details on tests and other analyses.**
3. Nielsen, D. M., Ehm, M. G. & Weir, B. S. Detecting marker–disease association by testing for Hardy–Weinberg disequilibrium at a marker locus. *Am. J. Hum. Genet.* **63**, 1531–1540 (1998).
4. Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 967–986 (2005).
5. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* **38**, 75–81 (2006).
6. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Rev. Genet.* **7**, 552–564 (2006).
7. Guo, S. W. & Thompson, E. A. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372 (1992).
8. Maiste, P. J. & Weir, B. S. A comparison of tests for independence in the FBI RFLP databases. *Genetica* **96**, 125–138 (1995).
9. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
10. Weir, B. S., Hill, W. G. & Cardon, L. R. Allelic association patterns for a dense SNP map. *Genet. Epidemiol.* **27**, 442–450 (2004).
11. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* (Wiley, New York, 2002).
12. Souverein, O. W., Zwinderman, A. H. & Tanck, M. W. T. Multiple imputation of missing genotype data for unrelated individuals. *Ann. Hum. Genet.* **70**, 372–381 (2006).
13. Clayton, D. G. *et al.* Population structure differential bias and genomic control in a large-scale case–control association study. *Nature Genet.* **37**, 1243–1246 (2005).
14 Marchini, J. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).
15. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
16. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
17. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
18. Abecasis, G. R. & Cookson, W. O. C. GOLD — graphical overview of linkage disequilibrium. *Bioinformatics* **16**, 182–183 (2000).
19. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
20. Maniatis, N. *et al.* The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA* **99**, 2228–2233 (2002).
21. Tapper, W. *et al.* A map of the human genome in linkage disequilibrium units. *Proc. Natl Acad. Sci. USA* **102**, 11835–11839 (2005).
22. Crawford, D. C. *et al.* Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genet.* **36**, 700–706 (2004).
23. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **23**, 581–584 (2004).
24. Li, N. & Stephens, M. Modelling LD and identifying recombination hotspots from SNP data. *Genetics* **165**, 2213–2233 (2003).
25. Jeffreys. A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatability complex. *Nature Genet.* **29**, 217–222 (2001).
26. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genet.* **36**, 151–156 (2004).
27. Ardlie, K. G., Krugylak, L. & Sielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.* **3**, 299–309 (2002).
28. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
29. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31 (2003).
30. Stram, D. O. Tag SNP selection for association studies. *Genet. Epidem.* **27**, 365–374 (2004).
31. 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).
32. Zeggini, E. *et al.* An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nature Genet.* **37**, 1320–1322 (2005).
33. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
34. Huang, W. *et al.* Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc. Natl Acad. Sci. USA* **103**, 1418–1421 (2006).
35. Gonzalez-Neira, A. *et al.* The portability of tagSNPs across populations: a worldwide survey. *Genome Res.* **16**, 323–330 (2006).
36. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* 2nd edn (Chapman and Hall, London, 1989).
**Still the best general reference on generalized linear models (includes linear, multinomial and logistic regression as special cases); it is relatively advanced and more gentle introductions are available.**
37. Sasieni, P. D. From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261 (1997).
**A useful reference for comparison of different single-SNP tests of association.**
38. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).
39. Freidlin, B., Zheng, G., Li, Z. H. & Gastwirth, J. L. Trend tests for case–control studies of genetic markers: power, sample size and robustness. *Hum. Hered.* **53**, 146–152 (2002).
40. Lunn, D. J., Whittaker, J. C. & Best, N. A Bayesian toolkit for genetic association studies. *Genet. Epidemiol.* **30**, 231–247 (2006).
41. Prentice, R. L. & Pyke, R. Logistic disease incidence models and case–control studies. *Biometrika* **66**, 403–411 (1979).
42. Seaman, S. R. & Richardson, S. Equivalence of prospective and retrospective models in the Bayesian analysis of case–control studies. *Biometrika* **91**, 15–25 (2004).
43. Cox, D. R. & Hinkley, D. V. *Theoretical statistics* (Chapman and Hall, London, 1974).

44. Wallace, C., Chapman J. M. & Clayton, D. G. Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *Am. J. Hum. Genet.* **78**, 498–504 (2006).
45. Agresti, A. *Categorical Data Analysis* 2nd edn (Wiley, New York, 2002).
46. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
47. Devlin, B. & Roeder, K. Genomic control a new approach to genetic-based association studies. *Theor. Pop. Biol.* **60**, 155–166 (2001).
48. Zheng, G., Freidlin, B. & Gastwirth. J. L. Robust genomic control. *Am. J. Hum. Genet.* **78**, 350–356 (2006).
49. 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).
50. Setakis, E., Stirnadel, H. & Balding D. J. Logistic regression protects against population structure in genetic association studies. *Genome Res.* **16**, 290–296 (2006).
51. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
52. Satten, G., Flanders, W. D. & Yang, Q. Accounting for unmeasured population structure in case–control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* **68**, 466–477 (2001).
53. Hoggart, C. J. *et al.* Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
54. Delrieu, O. & Bowman, C. Visualizing gene determinants of disease in drug discovery. *Pharmacogenomics* **7**, 311–329 (2006).
55. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
56. Yu, J. M. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet.* **38**, 203–208 (2006).
57. Waldron, E. R. B., Whittaker J. C. & Balding D. J. Fine mapping of disease genes via haplotype clustering. *Genet. Epidemiol.* **30**, 170–179 (2006).
58. Clayton, D., Chapman, J. & Cooper, J. The use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol.* **27**, 415–428 (2004).
59. Cordell, H. J. & Clayton, D. G. A unified stepwise regression approach for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am. J. Hum. Genet.* **70**, 124–141 (2002).
60. Wang, H. *et al.* Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**, 465–480 (2005).
61. Clark, A. G. The role of haplotypes in candidate-gene studies. *Genet. Epidemiol.* 27, 321–333 (2004).
62. Sham, P. *Statistics in Human Genetics* (Arnold, London, 1998).
**Still a useful reference for basic linkage and association analyses, but now a little out of date.**
63. Schaid, D. J. Evaluating associations of haplotypes with traits. *Genet. Epidemiol.* **27**, 348–364 (2004).
64. Tzeng, J. Y., Devlin, B., Wasserman, L. & Roeder, K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.* **72**, 891–902 (2003).
65. Lin, D. Y. & Zeng, D. Likelihood-based inference on haplotype effects in genetic association studies. *J. Am. Stat. Assoc.* **101**, 89–104 (2006).
66. 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).
67. Ke, X. Y. *et al.* The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13**, 577–588 (2004).
68. Templeton, A. R., Boerwinkle, E. & Sing C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics* **117**, 343–351 (1987).
**The first in a series of papers that initiated cladistic and more general clustering approaches to haplotype-based tests of association.**
69. Molitor, J., Marjoram, P. & Thomas, D. C. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am. J. Hum. Genet.* **73**, 1368–1384 (2003).
70. Seltman, H., Roeder, K. & Devlin, B. Evolutionary-based association analysis using haplotype data. *Genet. Epidemiol.* **25**, 48–58 (2003).
71. Durrant, C. *et al.* Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am. J. Hum. Genet.* **75**, 35–43 (2004).
72. Morris, A. P. Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genet. Epidemiol.* **29**, 91–107 (2005).
73. Beckmann, L., Thomas, D. C., Fischer, C. & Chang-Claude J. Haplotype sharing analysis using Mantel statistics. *Hum. Hered.* **59**, 67–78 (2005).
74. Templeton, A. R. *et al.* Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics* **169**, 441–453 (2005).
75. Tzeng, J. Y., Wang, C. H., Kao, J. T. & Hsiao, C. K. Regression-based association analysis with clustered haplotypes through use of genotypes. *Am. J. Hum. Genet.* **78**, 231–242 (2006).
76. Zollner, S. & Pritchard, J. K. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092 (2005).
77. Browning, S. R. Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* **78**, 903–913 (2006).
78. Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* **56**, 73–82 (2003).
79. Carlborg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nature Rev. Genet.* **5**, 618–625 (2004).
80. Todd, J. A. Statistical false positive or true disease pathway? *Nature Genet.* **38**, 731–733 (2006).
81. Lake, S. L. *et al.* Estimation and tests of haplotype–environment interaction when linkage phase is ambiguous. *Hum. Hered.* **55**, 56–65 (2003).
82. Millstein, J., Conti, D. V., Gilliland, F. D. & Gauderman, W. J. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet.* **78**, 15–27 (2006).
83. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Stat. Med.* **13**, 153–162 (1994).
84. 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).
85. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.* **37**, 413–417 (2005).
86. Storey, J. D. & Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
87. Dudbridge, F., Gusnanto, A. & Koeleman, P. C. Detecting multiple associations in genome-wide studies. *Hum. Genomics* **2**, 310–317 (2006).
88. Ishwaran, H. & Rao, J. S. Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Am. Stat. Assoc.* **98**, 438–455 (2003).
89. Yi, N. J. *et al.* Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170**, 1333–1344 (2005).
90. Zondervan, K. T. & Cardon, L. R. The complex interplay among factors that influence allelic association. *Nature Rev. Genet.* **5**, 238–238 (2004).
91. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
92. Bingham, S. & Riboli, E. Diet and cancer — the European prospective investigation into cancer and nutrition. *Nature Rev. Cancer* **4**, 206–215 (2004).
93. Ollier, W., Sprosen, T. & Peakman, T. UK Biobank: from concept to reality. *Pharmacogenomics* **6**, 639–646 (2005).
94. Leschzinger, G. *et al.* Clinical factors and ABCB1 polymorphisms in prediction of antiepileptic drug response: a prospective cohort study. *Lancet Neurol.* **5**, 668–676 (2006).
95. Thompson, E. in *Handbook of Statistical Genetics* 2nd edn (eds Balding D. J., Bishop, M. & Cannings, C.) 893–918 (Wiley, New York, 2003).
96. Holmans, P. in *Handbook of Statistical Genetics* 2nd edn (eds Balding D. J., Bishop, M. & Cannings, C.) 919–938 (Wiley, New York, 2003).
97. Ewens, W. J. & Spielman, R. S. in *Handbook of Statistical Genetics* 2nd edn (eds Balding D. J., Bishop, M. & Cannings, C.) 961–972 (Wiley, New York, 2003).
98. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. C. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
99. Van Steen, K. *et al.* Genomic screening and replication using the same data set in family-based association testing. *Nature Genet.* **37**, 683–691 (2005).
100. Smith, M. W. & O'Brien, S. J. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Rev. Genet.* **6**, 623–266 (2005).
101. Reich, D. *et al.* A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nature Genet.* **37**, 1113–1118 (2005).
102. Clayton, D. in *Handbook of Statistical Genetics* 2nd edn (eds Balding D. J., Bishop, M. & Cannings, C.) 939–960 (Wiley, New York, 2003).
103. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
104. Berger, M. *et al.* Hidden population substructures in an apparently homogeneous population bias association studies. *Eur. J. Hum. Genet.* **14**, 236–244 (2006).
105. Wang, H. S., Thomas, D. C., Pe'er I. & Stram, D. O. Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* **30**, 356–368 (2006).
106. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genet.* **38**, 209–213 (2006).
107. Verzilli, C. J., Stallard, N. & Whittaker, J. C. Bayesian graphical models for genomewide association studies. *Am. J. Hum. Genet.* **79**, 100–112 (2006).
108. Dudbridge, F. & Koeleman, P. C. 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).
109. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Rev. Genet.* **4**, 701–709 (2003).



### Acknowledgements
I thank W. Astle and E. Waldron for help with drawing figures, and W. Astle, L. Cardon, A. Lewin, D. Lunn, A. Morris, D. Schaid, J. Whittaker and D. Zabaneh for discussions and comments on drafts of the manuscript. The author is supported in part by the UK Medical Research Council.


### Competing interests statement
The author declares no competing financial interests.

### FURTHER INFORMATION
**European Bioinformatics Institute:** http://www.ebi.ac.uk
**Genetic Analysis Software (includes almost all freely available software for statistical genetic analyses, regularly updated):** http://linkage.rockefeller.edu/soft
**Genetic Power Calculator (a useful tool that calculates the power of many simple study designs):** http://pngu.mgh.harvard.edu/~purcell/gpc
**International HapMap Project:** http://www.hapmap.org
**Nature Reviews Genetics audio supplement:** http://www.nature.com/nrg/focus/stats/audio
**R genetics package:** http://rgenetics.org
**Wellcome Trust Case Control Consortium (a large, genome-wide association study for eight distinct diseases with common set of controls):** http://www.wtccc.org.uk

### SUPPLEMENTARY INFORMATION
**See online article:** S1 (box)
**Access to this links box is available online.**