Genetic Epidemiology 3

Genetic association studies

Heather J Cordell, David G Clayton

We review the rationale behind and discuss methods of design and analysis of genetic association studies. There are similarities between genetic association studies and classic epidemiological studies of environmental risk factors but there are also issues that are specific to studies of genetic risk factors such as the use of particular family-based designs, the need to account for different underlying genetic mechanisms, and the effect of population history. Association differs from linkage (covered elsewhere in this series) in that the alleles of interest will be the same across the whole population. As with other types of genetic epidemiological study, issues of design, statistical analysis, and interpretation are very important.

Lancet 2005; 366: 1121–31

This is the third in a Series of seven papers on genetic epidemiology.

University of Cambridge, Department of Medical Genetics, Juvenile Diabetes Research Foundation/ Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, Wellcome Trust/Medical Research Council Building, Addenbrookes Hospital, Hills Road, Cambridge, CB2 2XY, UK (H J Cordell DPhil, Prof D G Clayton MA)

Correspondence to: Dr Heather J Cordell heather.cordell@cimr.cam.ac.uk

Genetic association studies aim to detect association between one or more genetic polymorphisms and a trait, which might be some quantitative characteristic or a discrete attribute or disease. Association differs from linkage in that the same allele (or alleles) is associated with the trait in a similar manner across the whole population, while linkage allows different alleles to be associated with the trait in different families. However, genetic associations arise only because human populations share common ancestry and it has been argued that association studies are really just a special form of linkage study in which the extended family is the wider population. In linkage analysis, data from distantly related individuals are more powerful for detecting small effects than data from closely related individuals, but this advantage is offset by the fact that, owing to increased possibility for linkage to be destroyed by recombination, linkage extends over shorter distances in distantly related individuals, necessitating a greater density of markers. Association in apparently unrelated people represents the extreme of this effect: association analysis has greater power than linkage studies to detect small effects, but requires many more markers to be examined. The fact that association operates only over short distances in the genome has for long guaranteed association studies an important place in fine mapping genetic loci initially detected by linkage. More recently, it has been realised that genetic susceptibility to common complex disorders probably involves many genes, most of which have small effects. This fact, together with the identification of large numbers of single nucleotide polymorphisms (SNPs) throughout the genome and rapidly falling genotyping costs, has led to the importance of association studies in genetic epidemiology. Indeed, it is possible to envisage the search for disease susceptibility genes being done by screening large numbers of SNPs across the whole genome.[1,2]

Although family-based studies still have a place in the study of population association (in addition to linkage), such research has much more in common with classic epidemiological studies of environmental and behavioural risk factors than do linkage studies. Consequently, issues of study design and analysis have more in common with the rest of epidemiology. Parallels with classic epidemiology are also clear if we consider why association between a genetic polymorphism and a trait might exist in a given population: (1) the polymorphism has a causal role; (2) the polymorphism has no causal role but is associated with a nearby causal variant; or (3) the association is due to some underlying stratification or admixture of the population. In a mixed population in which strata have different environmental exposures or the founder populations entail different genetic risks, any locus whose allele frequencies differ between strata or founder populations will be associated with disease to some extent, whether or not it is near to a causal locus.

Direct association

The first of these forms of association is termed direct association, and studies of direct association target polymorphisms which are themselves putative causal variants. This type of study is the easiest to analyse and the most powerful, but the difficulty is the identification of candidate polymorphisms. A mutation in a codon which leads to an aminoacid change is a candidate causal variant. However, it is likely that many causal variants responsible for heritability of common complex disorders will be non-coding. For example, such variants may cause variation in gene regulation and expression, or differential splicing. We do not know enough to predict which variants may have such effects. Thus, direct association studies only have the potential to discover some of the genetic causes of disease and disease-related traits. However, some 10 000–15 000 aminoacid changing SNPs with minor allele frequency exceeding 1% in Europeans have been identified, and screening of these in whole genome studies is feasible.

Indirect association

In the second type of association, the polymorphism is a surrogate for the causal locus and this type of association allows us to search for causal genes in indirect association studies. However, indirect associations are even weaker than the direct associations they reflect, and it will usually be necessary to type several surrounding markers to have a high chance of picking up the indirect association. Indirect association studies are more difficult to analyse, and there is still debate as to the best methods. They are also less powerful than direct studies. Finally, by contrast with direct studies, until we can be sure that we have adequately charted the polymorphisms in a region, there cannot be a definitive negative result since we cannot exclude the possibility that a causal variant exists but is not picked up by the markers chosen. The next phase of the Human Genome Project—the International HapMap Project[3]—aims to improve our knowledge in this respect. This project will be discussed in more detail in a later paper in this series.[4] The imminent completion of the second phase of this study, plus rapid recent advances in high throughput genotyping technology, mean that screening of perhaps 80% of the genome for disease associations is becoming feasible, if costly. In the meantime, most indirect association studies concentrate on candidate genes identified either on the basis of their known function or from animal models. Even as whole genome studies are increasingly used, such candidate gene studies will continue to play an important part. Such studies will allow typing of markers more densely, not only to improve detection of true causal associations but also to increase confidence that negative findings represent true negatives.

Confounded association

The final type of association is that due to confounding by stratification and admixture (substructure) within the population. Confounding, as in the rest of epidemiology, raises the possibility both of generating false findings (positive confounding) or obscuring true causal associations (negative confounding). However, although the problem of unobserved confounding is intractable in classic epidemiology, dictating limits on the size of causal effect that can be safely inferred from observational studies,[5] genetic epidemiology offers possibilities for circumventing the difficulty.

The most obvious way of avoiding this difficulty is to measure association in well-mixed, outbred populations. Failing this, any stratification and admixture effects could be reduced by matching (in the design or the analysis, or both) by geographical region and by any markers of ethnic origin. In this manner, comparisons can be made, as far as possible, within homogeneous subpopulations. It has been argued that such devices will avoid the small confounding effects expected to arise from stratification and admixture.[6] However, this view has been questioned. Meta-analyses[7] have indicated that causal variants for complex disease might, when looked at one at a time, have rather small effects and large studies will be necessary to detect them.[8] Against this background, even modest confounding by stratification and admixture could have important repercussions. It is not yet known how serious this problem will be for association studies in populations of European origin, but it poses a grave difficulty in admixed populations such as African Americans or Afro-Caribbeans.[9] Admixture does present opportunities for gene mapping by exploiting a back-crossing experiment of nature, but such studies are beyond the scope of this article.

The first method for dealing with confounding by population structure is matching by family; if comparisons are made between siblings with the same parents, confounding by population structure is excluded. However, such studies are not always very powerful and they are difficult, or even impossible, to undertake on a sufficiently large scale to detect genetic associations reliably. The role of such studies will probably be to confirm findings generated by less expensive methods and to answer more complex secondary questions.

The second method for dealing with the problem is to seek genetic markers for population substructure, or ancestry informative markers—loci whose allele frequencies differ between the founder populations.[9–12] Inevitably there will be some loss of power due to the imperfect measurement of admixture proportions. This loss of power might be modest for populations in which founder populations are very different and there are good markers of substructure, such as African Americans,[9] but it remains to be seen whether this method can be applied efficiently to control for the smaller differences which might exist, for example, within European populations.

The third approach is genomic control.[13–15] Confounding is regarded as a random process, potentially affecting all loci, such that the effect of positive confounding is to increase the type 1 error (false positive) rate for association tests; although conventional tests for association are correct if regarded as tests for association within the population studied, they will have an inflated false positive rate when judged as tests of causal effects in the presence of stratification or admixture (or both). Another perspective (more intuitive to geneticists) is that, although people in a population-based association study can be regarded as having been independently sampled from the particular population studied, they are not independently sampled when regarded as a sample of all mankind; they are cryptically related because they have been drawn from the same population. As a result, when regarded as tests of the causal null hypothesis, conventional χ^2 tests for association have greater variance than they should and use of conventional significance levels will lead to a higher false-positive rate.

Genomic control is less ambitious than other methods that control for confounding by substructure in that it seeks only to control the false positive rate by increasing the threshold required for statistical significance. The factor by which the variance is inflated by confounding can be estimated by typing a large number of unselected markers across the genome and estimating the variance of association test statistics empirically. This method is simple to do. However, no attempt is made to deal with negative confounding, which increases the false negative rate and reduces power. Use of more stringent test criteria to control the false positive rate will accentuate loss of power. It also remains to be empirically tested whether the distribution of test statistics is inflated by the same multiple, irrespective of allele frequency and throughout the entire distribution.

It remains to be seen whether correcting for confounding by substructure by statistical modelling will be more powerful than accepting some degree of confounding and controlling the resultant type 1 error rate. Much will depend on how serious the problem turns out to be, and whether sufficiently informative markers will be identified for the former approach to work efficiently. However, the approaches could turn out to be complementary, with gross effects addressed by statistical models and surrogate measures of substructure and more subtle effects, such as those due to cryptic relatedness between cases and/or controls, left to genomic control.

Direct association: patterns of genotype–phenotype relationship

We shall consider a diallelic locus, directly related to either a quantitative trait or to a discrete trait such as presence (prevalence), or occurrence (incidence), of a disease. Multiallelic loci lead to more complicated scenarios and generate tests with many degrees of freedom. Even in the simplest diallelic case, different patterns for the genotype–phenotype relationship must be considered. Since there are three possible genotypes, which have a natural order (1/1, 1/2, and 2/2), the question of linearity of the relationship must be considered.

Linear dose-response modelling

In classic mendelian genetics of fully penetrant discrete traits, the description of an allele as dominant implies that the corresponding phenotype will occur irrespective of the number of copies of the allele carried. A recessive allele requires both copies to be present for the phenotype to be evident. In a diallelic system, if neither allele is dominant, 1/2 heterozygotes will display an intermediate phenotype. Fisher[16] used the term dominance in a different way to describe the related concept of linearity of the genotype–phenotype relationship for quantitative traits. He defined absence of dominance to imply the linear relationship:

$$\text{Mean trait value}=\alpha+\beta x$$

where x codes genotypes 1/1, 1/2, and 2/2 as 0, 1, and 2 respectively and β is the additive effect of each copy of the 2 allele. Since this model predicts that the trait mean for heterozygotes will lie precisely midway between the means for the two types of homozygote, it is easy to see why Fisher identified linearity with absence of dominance, but this idea is based on a stronger model than the earlier concept.

The importance of a simplifying model such as the linear dose-response model above is that the strength of genotype–phenotype relationship is expressed in a single parameter (β) and statistical tests for existence of such a relationship only have one degree of freedom. To extend the model to allow a quite general pattern of relationship we must introduce an additional parameter to measure deviation from linearity. For example, we might introduce a variable z coded as 0 for homozygotes and 1 for heterozygotes, to give the following model:

$$\text{Mean trait value}=\alpha+\beta x$$

γ is then said to represent a dominance effect. In this extended model, all patterns of relationship between phenotype mean and the three genotypes are possible, but two parameters now code the association and statistical tests have two degrees of freedom. Consideration of this broader class of models inevitably carries the penalty of reduced power if the pattern of relationship truly is linear. Some have argued that in most cases we would wish to constrain the two-parameter model so that the trait mean for heterozygotes cannot lie outside the range delimited by the means for homozygotes. This approach leads to tests that are intermediate between conventional tests with one and two degrees of freedom.[17] In any situation, the choice of the most powerful test depends on the pattern of association that actually exists and, unless we are simply doing confirmatory studies, this pattern is unknown a priori—a ubiquitous problem for statistical analysis. Perhaps for most complex disease genetics the model in which heterozygote risk is constrained to lie within the range defined by the two homozygote risks is the best compromise between generality and parsimony. However, such a model is little used, perhaps because of lack of software implementations.

To model gene effects on binary qualitative traits that are not fully penetrant, Wright[18] introduced the notion of an underlying, unobserved, and normally distributed quantitative trait (liability) governed by Fisher's linear model; the discrete trait is assumed to manifest when liability exceeds some threshold value. The predictions from Wright's model are very close to those from the logistic regression model, which is the mainstay of statistical analysis in the rest of epidemiology.[19,20] With this approach, absence of dominance means that the log odds of response for 1/2 heterozygotes is midway between that for 1/1 and 2/2 homozygotes, and so each allele contributes multiplicatively to the odds. For uncommon traits (as most diseases are), this model is nearly the same as the model of multiplicative effects of each allele on risk.

The multiplicative risk model could be argued to be the natural model for lack of dominance in this context. The multiplicative risk model has one particularly useful property. Hardy-Weinberg equilibrium is defined by genotype frequencies consistent with the two alleles being independently sampled from a population of alleles. Genotypes of controls, in a case-control study, should therefore be in Hardy-Weinberg equilibrium. But if disease risk is related to genotype multiplicatively, such that genotype risk can be decomposed into a product of effects of the two alleles, then genotypes of the cases of disease are also expected to be in Hardy-Weinberg equilibrium, with alleles being independently drawn from a population in which the frequency of high risk alleles is increased. This result justifies the common practice of counting alleles rather than genotypes in statistical analyses.[21]

Epistasis

The general issue of dominance relates to the extent to which the joint effect of two alleles at a single autosomal locus might be different from the sum (or product in a multiplicative model) of the effects anticipated for each allele independently. A related issue is the degree to which the combined effect of alleles at two or more loci can reasonably be modelled by the individual locus contributions. The fact that inheritance of some traits could only be explained by joint action of two unlinked loci was first demonstrated by Bateson,[22] who termed the effect epistasis. In these first examples, variation of phenotype with genotype at one locus was only apparent in those with certain genotypes at the second locus; others would show no effect. Thus epistasis was defined as one locus masking the effect of another. Fisher[16] used a similar term, epistacy, to refer to a statistical interaction meaning deviation from additive effects of the two loci upon the trait mean. The term epistacy soon evolved into epistasis,[23] and in modern genetics the two uses of the word coexist, often causing confusion.[24,25]

Epistasis in Fisher's sense is dependent on scale and, in general, does not have a clear interpretation in terms of mechanism. The interpretation of the causal implications of statistical interaction in epidemiology has been vigorously debated over at least three decades.[26,27] A similar debate continues in relation to interaction between genes and environmental risk factors. Some have argued that the interaction of genes and environment will become a major influence on the epidemiological study of disease causation and on public heath interventions,[28–30] whereas others have been more sceptical.[31]

If interpretation of statistical interaction between genes is problematic, an important reason to consider such interaction relates to our ability to discover the genes related to complex diseases in the first place. If such genes act together, epistatically, with several genes acting in the same pathway, the marginal effect of each gene on its own might be small, but might reflect much larger effects of collections of genes.[32,33] Some have even postulated scenarios in which marginal effects are absent altogether.[34] But this hypothesis requires one gene to reverse the direction of effect of another, which although possible, is perhaps unlikely to happen widely. Such arguments have led these same researchers to suggest that the analysis of association studies should move away from analysis of genes one at time, focusing instead on pairs or even larger constellations of genes. It is not yet clear whether the gains in effect size realised in practice by consideration of several genes at a time will be sufficient to compensate for the requirement for more stringent correction for the number of hypotheses to be tested.[35] A further debate concerns the relative merits of recursive partitioning methods, which derive from the automatic interaction detection methods of Sonquist and Morgan,[36] originating in the social sciences but now widely used in the computer science and bioinformatics communities, over more standard regression-based approaches.

Indirect association: patterns of linkage disequilibrium

The mapping of susceptibility genes for common complex disorders and genes for other common traits by the indirect method depends on the existence of association, at the population level, between the causal variants and nearby markers. Such association, because of the proximity of loci on the genome, is termed linkage disequilibrium. (Some use this term to describe any population-wide association between loci, whether due to proximity or to another reason such as population stratification and admixture. We prefer the term allelic association for this more general circumstance. The term gametic phase disequilibrium is also used to describe allelic association due to proximity). Success of this strategy depends upon some understanding of patterns of linkage disequilibrium and the forces that determine them—mutation, recombination, and population history.

The figure shows the genealogy of the same small segment of eight versions of the same chromosome. It is assumed that they will be descended from a common ancestor, and that the segment is so small that no recombination will have arisen within the segment. This latter assumption is necessary because the recombination in the sample history is an important complication: an adjacent segment separated by recombination will have an entirely different genealogy above this point. We assume that a mutant allele cannot revert back to wild type (lightning does not strike twice), and every copy of the mutant allele in the population is descended from the same ancestral mutation. The scale for the height of the genealogy is meioses (ie, generations). In this example there are four haplotypes. Labelling the initial allele at each locus as *1* and the new allele created by mutation as *2* these are *111* (individuals 6, 7, and 8), *122* (individuals 4 and 5), *211* (individuals 1 and 2), and *121* (individual 3). Alleles that are common in the sample are older mutations, and the number of different haplotypes increases in direct proportion to the number of polymorphisms, unless some polymorphisms correspond to mutations on the same branch of the genealogy. Less obvious is that fact that, even under these simplifying assumptions, the pattern and strength of association between polymorphisms is very variable.

The situation in the figure represents complete linkage disequilibrium between the three loci. This fact is apparent when looking at loci two at a time; each pair of loci define only three haplotypes. Table 1 shows the two-locus haplotype frequencies as 2×2 contingency tables. Complete linkage disequilibrium between pairs of loci is evident because at least one cell of the corresponding table is zero, since this is the maximum degree of association possible given the row and column totals. Linkage disequilibrium decays for three reasons: (1) recombination(s) in the genealogy occurring at some point between the two loci; (2) recurrence of the same mutation; and (3) gene conversion (transfer of information between alleles or loci).[37] The first reason is the most important for this decay. Since the probability of recombination increases with the distance between the loci, the strength of linkage disequilibrium is expected to decline with distance.

Various different measures of pairwise linkage disequilibrium have been proposed,[38] including Lewontin's D′,[39] which has also been termed the "association probability".[40] Lewontin's D′ is an important measure for identification of regions in which there has been little recombination and, therefore, in which there is the potential to map causal loci by indirect association studies. However, this measure does not directly determine the power of indirect association studies. Formally, the power of tests for indirect association depends largely on the index r^2, the square of the conventional correlation coefficient between the allele at the typed locus, scored 1 or 2, and the allele at the causal locus, scored similarly. The dependence of power on r^2 rather than on any other measure of association is complete for quantitative traits in the absence of a dominance variance component due to the causal locus.[41] The nature of the relationship between r^2 and the power to detect association is such that, if B is causal, we would require a sample size 2·8 times as large (0·56/0·2) to detect the indirect association with A than to detect the association with C. The panel shows that, even when loci are in complete disequilibrium (D′=1), the pairwise r^2 values can vary widely, because they are related to the allele frequencies and to the position of the corresponding mutations in the genealogy.

Linkage disequilibrium is also relevant to the more recent discussion of "haplotype blocks".[42] Genetic loci across large areas of the genome were suggested to divide into blocks characterised by little disequilibrium between blocks and limited haplotype diversity within blocks. These two aspects of blocks, physical extent and haplotype diversity are, in a sense, reflected by the measures D′ and r^2, respectively. However, they are not necessarily linked since they are determined by the different random processes of recombination and mutation; both the extent and haplotype diversity of blocks are extremely variable. Further, haplotypic diversity almost inevitably increases as more polymorphisms are discovered.

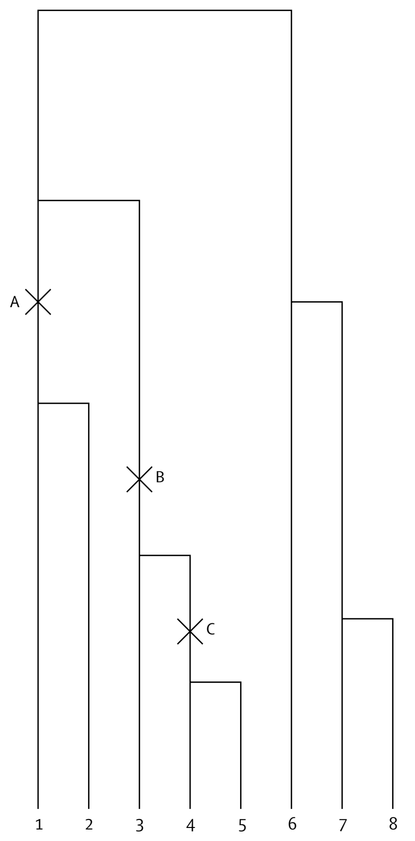

Figure: **Genealogy of a sample of eight chromosomes, showing ancestry of three SNPs**
Crosses=mutations, each of which will generate new (diallelic) polymorphism.

Locus A	Locus B			Locus C	Locus B			Locus C	Locus A		
	1	2	Total		1	2	Total		1	2	Total
1	3	3	6	1	5	1	6	1	4	2	6
2	2	0	2	2	0	2	2	2	2	0	2
Total	5	3		Total	5	3		Total	6	2	
r^2=0·2				r^2=0·56				r^2=0·11			

Lewontin's D′=1·0 in all cases.

Table 1: **Pairwise linkage disequilibrium for the eight-chromosome genealogy**

There has been some discussion as to whether blocks have clear boundaries, coincident with so-called recombination hot spots, or whether they arise as a result of purely random forces.[43,44] Sperm-typing experiments show the existence of hot spots[45] but, random forces undoubtedly also have an important role. This debate is important in relation to the stability of block structures across populations and to the sharpness of block boundaries. If the extent of haplotypes is determined by random recombination, then all haplotypes encompassing a given point in the genome will not be the same length and we should not be surprised to see a few high values of D′ extending well outside the main block of linkage disequilibrium.

The idea of haplotype blocks tends to be linked with the idea of haplotype tagging SNPs,[46] largely because the ideas were published simultaneously. However, the idea of haplotype tagging SNPs arose from studies of candidate genes after it was noted that, after discovering large numbers of SNPs by a combination of searching databases and exon resequencing, there is usually substantial redundancy—a few haplotype tagging SNPs capture, in some sense, most of the polymorphism of the gene. Many different methods have been proposed for the choice of such SNPs.[46–50] Consideration of the power to detect indirect association via haplotype tagging SNPs suggest that the important criterion is the coefficient of determination, a generalisation of r^2 to multiple regression models and usually denoted by R^2. This quantity measures the ability of a set of tag SNPs to predict another dimorphism.[48,51] The R^2 values with which tag SNPs predict the remaining known polymorphisms provide an estimate of the likely ability to predict a causal variant, but, with our limited knowledge of human polymorphism, its accuracy cannot be guaranteed.

Study designs

Familiar epidemiological designs such as population-based case-control or cohort designs[19,52] are often used for genetic association studies and the data are analysed much the same way too, risk factors such as smoking and obesity etc, being replaced by the presence or absence of a particular genetic polymorphism. Risk can be considered in terms of either a predisposing allele or genotype, or in terms of multiple categories of disease risk such as the risks associated with different alleles at a multiallelic genetic locus, or the risks associated with the three possible genotypes 1/1, 1/2, 2/2 at a single diallelic locus.

Other designs have been specifically proposed for genetic studies. Family-based designs such as the case-parent triad design,[53,54] case-parent-grandparent design,[55] or analysis of general pedigrees have been proposed to counteract confounding due to population stratification that can occur in case-control or other population-based designs.[56] In family designs, alleles or genotypes

	Details	Advantages	Disadvantages	Statistical analysis method
Cross-sectional	Genotype and phenotype (ie, note disease status or quantitative trait value) a random sample from population	Inexpensive. Provides estimate of disease prevalence	Few affected individuals if disease rare	Logistic regression, χ^2 tests of association or linear regression
Cohort	Genotype subsection of population and follow disease incidence for specified time period	Provides estimate of disease incidence	Expensive to follow-up. Issues with drop-out	Survival analysis methods
Case-control	Genotype specified number of affected (case) and unaffected (control) individuals. Cases usually obtained from family practitioners or disease registries, controls obtained from random population sample or convenience sample	No need for follow-up. Provides estimates of exposure effects	Requires careful selection of controls. Potential for confounding (eg, population stratification)	Logistic regression, χ^2 tests of association
Extreme values	Genotype individuals with extreme (high or low) values of a quantitative trait, as established from initial cross-sectional or cohort sample	Genotype only most informative individuals hence save on genotyping costs	No estimate of true genetic effect sizes	Linear regression, non-parametric, or permutation approaches
Case-parent triads	Genotype affected individuals plus their parents (affected individuals determined from initial cross-sectional, cohort, or disease-outcome based sample)	Robust to population stratification. Can estimate maternal and imprinting effects	Less powerful than case-control design	Transmission/disequilibrium test, conditional logistic regression or log-linear models
Case-parent-grandparent septets	Genotype affected individuals plus their parents and grandparents	Robust to population stratification. Can estimate maternal and imprinting effects	Grandparents rarely available	Log-linear models
General pedigrees	Genotype random sample or disease-outcome based sample of families from general population. Phenotype for disease trait or quantitative trait	Higher power with large families. Sample may already exist from linkage studies	Expensive to genotype. Many missing individuals	Pedigree disequilibrium test, family-based association test, quantitative transmission/disequilibrium test
Case-only	Genotype only affected individuals, obtained from initial cross-sectional, cohort, or disease-outcome based sample	Most powerful design for detection of interaction effects	Can only estimate interaction effects. Very sensitive to population stratification	Logistic regression, χ^2 tests of association
DNA-pooling	Applies to variety of above designs, but genotyping is of pools of anywhere between two and 100 individuals, rather than on an individual basis	Potentially inexpensive compared with individual genotyping (but technology still under development)	Hard to estimate different experimental sources of variance	Estimation of components of variance

Table 2: **Study designs for genetic association studies**

transmitted to affected individuals are compared with untransmitted alleles or genotypes, providing a control sample that is inherently matched to the case sample with regard to population structure. An alternative approach is to use population–based studies and correct for population stratification.[11–13] However, these methods involve typing of either a large number of unselected markers or a panel of markers chosen to be highly informative for the type of admixture in the study population. Such corrections will only be possible in large studies. The case-parent triad design typically requires the same number of triads (consisting of a case and two parents) to be typed as the number of cases required in a case-control design (assuming an equal number of controls), to give the same power. Thus a sample of 500 case-parent triads will have roughly the same power as 500 cases and 500 controls, but the case-parent triad design requires 1·5 times the amount of genotyping, and could also be more difficult to obtain (except when family samples had already been collected for previous linkage study). For this reason, many prefer the case-control approach; however, family-based approaches provide a useful complementary strategy because of their robustness to population stratification and because they allow estimates of effects due to direct maternal genotype or maternal-fetal interaction and parent-of-origin (imprinting) effects.[57–60] Case-parent triad designs allow such effects to be estimated at the expense of rather weak assumptions concerning population distributions of parental genotypes. These assumptions may be avoided by use of the case-parent-grandparent design (but such families may be difficult to obtain in practice).

Some designs try to reduce the genotyping effort, for instance by only typing individuals at the extremes of the phenotype distribution. DNA pooling studies[61] reduce the amount of genotyping by typing DNA pooled from a group of individuals as opposed to genotyping each person separately. With the haplotype tagging approach,[46] genotypes from an initial sample (say 32 people) are used to select loci to genotype in the larger sample. This strategy essentially exploits the indirect association approach to gene mapping. Further cost savings and efficiency can be obtained by a staged strategy.[62,63]

Various different designs are commonly used in genetic association studies (table 2). Methods and programs have been developed for power and sample size calculations (table 3).[2,52,64–74]

	Reference	URL
Analytical calculation	52, 2, 64, 65	
QUANTO	66, 67	http://hydra.usc.edu/gxe
Genetic power calculator	68	http://statgen.iop.kcl.ac.uk/gpc/index.html
Stata power and sample size programs	69, 70	http://cruk.leeds.ac.uk/katie
TDTPOWER	71	http://www.uni-bonn.de/~umt70e/soft.htm
TDTASP	72, 73	http://odin.mdacc.tmc.edu/anonftp/
TDT-PC	74	http://www.biostat.jhsph.edu/~wmchen/pc.html

Table 3: **Resources for power and sample size calculations**

Statistical analysis

The analysis of data depends crucially on the study design. In the simplest case, familiar methods such as logistic regression, χ^2 tests of association, and odds ratios may be suitable. At a single marker, the issue arises as to whether to analyse on the basis of allele counts or genotype counts. Suppose we have case and control data for a single diallelic genetic locus (table 4). A simple χ^2 test for independence has 2 degrees of freedom. Two odds ratios can be calculated: af/be (for genotype 2/2 *vs* 1/1) and cf/de (for 1/2 *vs* 1/1). Alternatively, if there is a reason to expect dominance or recessiveness in the effect of allele 2, we could group the top two rows or bottom two rows together to provide a χ^2 test with 1 degree of freedom and an odds ratio of (a+c)f/(b+d)e or a(d+f)/b(c+e), respectively. Another approach might be a test of trend, with a dose-response effect in regard to the number of copies of the 2 allele. A similar test could be done by uncoupling the alleles within a genotype and constructing a test in terms of case and control chromosomes (table 5). A χ^2 test of association with 1 degree of freedom on the data in table 4 assumes that chromosomes or alleles are independent units,[21] which essentially means Hardy-Weinberg equilibrium (and, for estimation of effects under the alternative hypothesis, multiplicative effects of alleles). All of these tests can be done with statistical software.

Table 6 lists some sources for statistical methods commonly used in genetic association studies.[20,52,54,57–61,71,74–91] Although many can be done in standard packages, some (particularly for family data) require special software. Some are designed to be simple, others require some specialist knowledge.

The simplest and most powerful statistical analyses arise in the direct association studies in which causal hypotheses, and hence analyses, are specific to single, typed polymorphisms. However, in indirect studies, which exploit linkage disequilibrium between typed markers and causal variants, analysis of marker loci one at a time might not be ideal—the r^2 between single

	Cases	Controls
2/2	a	b
1/2	c	d
1/1	e	f

Table 4: **Counts of genotypes in case-control study**

	Chromosomes	
Allele	Cases	Controls
2	2a+c	2b+d
1	2e+c	2f+d

Table 5: **Counts of chromosomes in case-control study**

	Approach	Reference	Software	URL
Logistic regression	Model log odds of disease as linear function of underlying genotype variables	20, 74, 20	Standard statistical package (eg, Stata, SAS, S-Plus, R)	http://www.stata.com/ http://www.sas.com/ http://www.insightful.com/products/splus/ http://www.r-project.org/
χ^2 test of association	Test for independence of disease status and genetic risk factor	20	Standard statistical package	See above
Linear regression	Model quantitative trait as linear function of underlying genotype variables	75	Standard statistical package	See above
Survival analysis	Model survivor function or hazard as function of underlying genotype variables	20, 52	Standard statistical package	See above
Transmission/ disequilibrium test	Test departure of transmission of alleles from heterozygous parents to affected offspring from null hypothesis of half	71, 76–78	Various (eg, Genehunter, RC-TDT, Genassoc, Transmit, Unphased	http://fhcrc.org/labs/kruglyak/Downloads/index.html http://www.uni-bonn.de/~umt70e/soft.htm http://www-gene.cimr.cam.ac.uk/clayton/software/ http://www.mrc-bsu.cam.ac.uk/personal/frank/
Conditional logistic regression	Calculate conditional probability of affected offspring genotypes, given parental genotypes	54, 60, 79, 80	Genassoc Unphased	http://www-gene.cimr.cam.ac.uk/clayton/software/ http://www.mrc-bsu.cam.ac.uk/personal/frank/
Log linear models	Model counts of genotype combinations for mother, father, and affected offspring	57, 58, 59	Standard statistical package	See above
Pedigree disequilibrium test	Test departure of transmission of alleles to affected pedigree members from null expectation	81, 82	Pedigree disequilibrium test Unphased	http://www.chg.duke.edu/software/pdt.html http://www.mrc-bsu.cam.ac.uk/personal/frank/
Family-base association test	Tests for association or linkage between disease phenotypes and haplotypes by utilising family-based controls	83–86	Family-based association test	http://www.biostat.harvard.edu/~fbat/fbat.htm
Quantitative transmission/ disequilibrium test	Linkage disequilibrium analysis of quantitative and qualitative traits based on variance components	87, 88	Quantitative transmission/ disequilibrium test	http://www.sph.umich.edu/csg/abecasis/QTDT/
DNA pooling	Test for differences in allele frequencies in different pooled samples while estimating components of variance due to experimental error	61, 89–91	Standard statistical package	See above

Table 6: **Statistical methods for genetic association studies**

markers and the causal locus could be much lower than the R^2 for prediction of the causal locus from a group of markers. For such studies, therefore, it will be preferable to use multilocus approaches to analysis. However, these methods are still at an early stage of development.

Multilocus approaches are generally assumed to involve consideration of haplotypes. Analysis at the haplotype level can reveal an effect marked by an ancestral haplotype but has two main drawbacks: since the number of haplotypes could be large, the potential gain might be offset by an excessive increase in the degrees of freedom in the test; and haplotype phase will often be uncertain.

There are several ways in which the first problem might be approached. The simplest, and most common is to pool rare haplotypes, which will certainly sacrifice some information. Instead some have proposed grouping strategies based on cladistic considerations.[93] However, for markers in regions in which linkage disequilibrium is strong, it is questionable whether the use of any haplotype information is worth the increase in degrees of freedom since, in such circumstances, a simple multiple regression equation with one parameter per marker can achieve prediction of untyped loci with R^2 only slightly worse than haplotype-based predictions.[48,51] This finding suggests testing for indirect associations either by regression of trait on marker loci, without inclusion of the interaction terms, or by an appropriate variant of Hotelling's T^2 statistic.[48,51,94,95] These analyses have the additional attraction of not requiring resolution of haplotype phase and can often be done with conventional statistical packages.

When linkage disequilibrium is less strong, haplotype analyses remain important, especially for fine mapping (eg, in a stepwise logistic regression strategy).[80] Long haplotypes, spanning several blocks of linkage disequilibrium, are particularly important for identifying rare variants,[96] although these would have to have large effects to be detectable. The resolution of phase can then present a serious practical problem. Various computational algorithms address the phase-estimation problem, both in unrelated individuals[97–99] and in families.[100] In a two-stage procedure (whereby haplotype scoring based on a first haplotype analysis stage are used in a second stage test of association), these algorithms can be satisfactory in population-based studies since significance can be assessed by permutation arguments. For estimation of relative risks, however, association and haplotype phase must be considered simultaneously.[101] In family studies, too, simple two-stage approaches break down since information about association is given by transmission patterns, which are relevant to haplotype phase resolution. Even for transmission/disequilibrium studies in which both parents of cases are genotyped successfully at all loci, care is necessary since simply restricting analysis to families in which phase can be assigned can cause bias.[78] This problem can be avoided by judicious selection of the information used in the analysis,[80] albeit with some loss of information.

Alternatively, as in population–based studies, phase uncertainty can be accounted for explicitly in the association analysis.[77] The problem of uncertain phase can be avoided altogether by use of molecular methods for haplotyping,[102,103] but such methods usually have lower throughput and are more expensive than those that yield only the diplotype.

In addition to associations between phenotypes and single genes, interaction effects between genes or between genes and environment can also be studied. After taking account of the vast increase in the number of potential tests, the expected power to detect interactions is low. However, power can be increased if we can safely assume independence between genes, or between a gene and an environmental exposure, within the population as a whole and, therefore, within controls. Evidence for statistical interaction can then be obtained from examination of cases only.[104,105] However, as we have emphasised, the relationship between statistical and biological interactions (eg, functional interaction between proteins) is complex. Such analyses are more relevant to prediction of disease risk than to elucidation of the underlying trait pathogenesis.[25–27]

Significance and importance

The standards of statistical proof that have become acceptable in the general biomedical literature are not appropriate for genetic association studies. Something akin to a multiple testing problem pervades the discipline, although there has been no clear consensus about how it should be dealt with. Approaches such as the Bonferroni correction are not appropriate because it is not the number of tests in any one investigation that is important. Rather it is that the vast majority of loci tested will not be associated, so that even a small false positive probability will mean that most positive results will turn out to be false. Thus, it is the a-priori probability of association that needs to be accounted for, rather than the number of tests. Thus, it has been suggested that Bayesian methods are more appropriate;[106,107] when prior probability of association is known, they allow calculation of the posterior probability that an association is genuine. However, the mathematics require knowledge not only of the prior probability of association but also of the distribution of the size of effects that will be encountered.

In gene expression array studies, so many tests are done simultaneously that these unknowns can be estimated within the experiment, and empirical Bayes methods can be used.[108,109] When genome-wide association studies with many thousands of SNPs become feasible, such methods will become appropriate for association studies,[110] but in the meantime, studies of candidate regions will dominate, and here the prior probabilities that determine appropriate standards of evidence remain largely subjective. However, such considerations show that, given the small a-priori probability that any genetic locus is associated with disease and the small effect sizes that seem to be typical and the inadequate study sizes that have also been typical, it should not be at all surprising that most findings judged positive with conventional levels of statistical significance have not been replicated.[6]

Some would respond by pointing to the very low population attributable fractions that correspond to these small genetic effects and asking whether there is any utility in their discovery. However, no one would claim that the interventions that will follow from advances in genetic epidemiology will simply correct the less beneficial genetic variation. Instead, the important role of such studies will be the elucidation of mechanisms. In epidemiology, the role of genetic variation can be important in establishing the causal nature of environmental associations in which intervention could have major effects.[111]

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

Funding was provided by the Wellcome Trust and the Juvenile Diabetes Research Foundation.

References

1 Livak K, Marmaro J, Todd J. Towards fully automated genome-wide polymorphism screening. *Nat Genet* 1995; **9**: 341–42.

2 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–17.

3 The International HapMap Consortium. The International HapMap project. *Nature* 2003; **426**: 789–96.

4 Hattersley AT, McCarthy MI. What makes a good genetic association study? *Lancet* (in press).

5 Taubes G. Epidemiology faces its limits. *Science* 1996; **269**: 164–69.

6 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* 2002; **11**: 513–20.

7 Ioannidis J, Trikalinos T, Ntzani E, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. *Lancet* 2003; **361**: 567–71.

8 Dahlman I, Eaves IA, Kosoy R, et al. Parameters for reliable results in genetic association studies in common disease. *Nat Genet* 2002; **30**: 149–50.

9 Hoggart C, Parra E, Shriver M, et al. Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* 2003; **72**: 1492–504.

10 Pritchard J, Rosenberg N. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–28.

11 Pritchard J, Stephens M, Rosenberg N, et al. Association mapping in structured populations. *Am J Hum Genet* 2000; **67**: 170–81.

12 Satten G, Flanders W, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001; **68**: 466–77.

13 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.

14 Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* 2000; **66**: 1933–1944.

15 Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. *Genet Epidemiol* 2002; **22**: 78–93.

16 Fisher R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edin* 1918; **52**: 399–433.

17 Chiano M, Clayton D. Genotype relative risks under ordered restriction. *Genet Epidemiol* 1998; **15**: 135–46.

18 Wright S. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proc Natl Acad Sci USA* 1920; **6**: 320–22.

19 Breslow N, Day N. Statistical Methods in Cancer Research. Volume I—The Analysis of Case-Control Studies. IARC Scientific Publications. Lyon: IARC, 1980.

20 Clayton D, Hills M. Statistical Models in Epidemiology. Oxford: Oxford University Press, 1993.

21 Sasieni P. From genotypes to genes: doubling the sample size. *Biometrics* 1997; **53**: 1253–61.

22 Bateson W. Mendel's principles of heredity. Cambridge: Cambridge University Press, 1909.

23 Phillips P. The language of gene interaction. *Genetics* 1998; **149**: 1167–71.

24 Cordell H, Todd J, Hill N, et al. Statistical modeling of interlocus interactions in a complex disease: Rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* 2001; **158**: 357–67.

25 Cordell H. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002; **11**: 2463–68.

26 Siemiatycki J, Thomas D. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol* 1981; **10**: 383–87.

27 Thompson W. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991; **44**: 221–32.

28 Khoury M, Wagener D. Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors. *Am J Hum Genet* 1995; **56**: 835–44.

29 Shpilberg O, Dorman J, Ferrel M, et al. The next stage: molecular epidemiology. *J Clin Epidemiol* 1997; **50**: 635–38.

30 Khoury M. Genetic and epidemiological approaches to the search for gene-environment interaction: the case of osteoporosis. *Am J Epidemiol* 1998; **147**: 1–2.

31 Clayton D, McKeigue P. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358**: 1357–60.

32 Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 2003; **4**: 701–09.

33 Moore J. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003; **56**: 73–82.

34 Culverhouse R, Suarez B, Lin J, et al. A perspective on epistasis: limits of models displaying no main effects. *Am J Hum Genet* 2001; **70**: 461–71.

35 Marchini J, Donnelly P, Cardon L. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005; **37**: 413–17.

36 Sonquist J, Morgan J. The detection of interaction effects. volume Monograph 35 of *Survey Research Center, Institute of Social Research*. Ann Arbor: University of Michigan, 1964.

37 Jeffreys AJ, May C. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* 2004; **36**: 151–56.

38 Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–22.

39 Lewontin R. The interaction of selection and linkage I. General considerations. *Genetics* 1964; **49**: 49–67.

40 Morton N, Zhang W, Taillon-Miller P, et al. The optimal measure of allelic association. *Proc Natl Acad Sci US* 2001; **98**: 5217–21.

41 Sham PC, Cherny SS, Purcell S, et al. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 2000; **66**: 1616–30.

42 Daly M, Rioux J, Schaffer S, et al. High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–32.

43 Ardlie K, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 2002; **3**: 299–309.

44 Wall J, Pritchard J. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003; **4**: 587–97.

45 Jeffreys AJ, Kauppi L, Neumann H. Intensely punctate meiotic recombination in the class II region of the major histocompatibilty complex. *Nat Genet* 2001; **29**: 217–22.

46 Johnson G, Esposito L, Barratt B, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–37.

47 Stram D, Haiman C, Altshuler D, 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* 2003; **55**: 27–36.

48 Chapman JM, Cooper JD, Todd JA, et al. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 2003; **56**: 18–31.

49 Byng MC, Whittaker JC, Cuthbert AP, et al. SNP subset selection for genetic association studies. *Ann Hum Genet* 2003; **67**: 543–56.

50 Carlson CS, Eberle MA, Rieder MJ, et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–20.

51 Clayton D, Chapman J, Cooper J. The use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 2004; **27**: 415–28.

52 Breslow N, Day N. Statistical Methods in Cancer Research. Volume II: The Design and Analysis of Cohort Studies. IARC Scientific Publications. Lyon: IARC, 1987.

53 Falk C, Rubinstein P. Haplotype relative risks: an easy and reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 1987; **51**: 227–33.

54 Schaid D, Sommer S. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* 1993; **53**: 1114–26.

55 Weinberg C. Studying parents and grandparents to assess genetic contributions to early-onset disease. *Am J Hum Genet* 2003; **72**: 438–47.

56 Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361**: 598–604.

57 Weinberg C, Wilcox A, Lie R. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 1998; **62**: 969–78.

58 Weinberg C. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am Hum Genet* 1999; **65**: 229–35.

59 Sinsheimer J, Palmer C, Woodward J. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. *Genet Epidemiol* 2003; **24**: 1–13.

60 Cordell H, Barratt B, Clayton D. Case/pseudo-control analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions and parent-of-origin effects. *Genet Epidemiol* 2004; **26**: 167–85.

61 Sham P, Bader J, Craig I, et al. DNA pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002; **3**: 862–71.

62 Satagopan J, Elston R. Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 2003; **25**: 149–57.

63 Lowe C, Cooper J, Chapman J, et al. Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun* 2004; **5**: 301–05.

64 Camp NJ. Genomewide transmission/disequilibrium testing: consideration of the genotypic relative risks at disease loci. *Am J Hum Genet* 1997; **61**: 1424–30.

65 Camp N. Genomewide transmission/disequilibrium testing: a correction. *Am J Hum Genet* 1999; **64**: 1485–87.

66 Gauderman W. Sample size calculations for matched case-control studies of gene-environment interaction. *Stat Med* 2002; **21**: 35–50.

67 Gauderman W. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 2002; **155**: 478–84.

68 Purcell S, Cherny S, Sham P. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; **19**: 149–50.

69 Saunders C, Bishop D, Barrett J. Power and sample size calculations for studies of gene-gene and gene-environment interactions. *Genet Epidemiol* 2002; **23**: 302–03.

70 Saunders C, Bishop D, Barrett J. Sample size calculations for main effects and interactions in case-control studies using Stata's nchi2 and npnchi2 functions. *Stata J* 2003; **3**: 47–56.

71 Knapp M. A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet* 1999; **64**: 861–70.

72 McGinnis R. General equations for Pt, Ps, and the power of the TDT and the affected-sib-pair test. *Am J Hum Genet* 2000; **67**: 1340–47.

73 McGinnis R, Shifman S, Darvasi A. Power and efficiency of the TDT and case-control design for association scans. *Behav Genet* 2002; **32**: 135–44.

74 Chen W, Deng H. A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genet Epidemiol* 2001; **21**: 53–67.

75 McCullagh P, Nelder J. Generalized Linear Models. London: Chapman & Hall, 1989.

76 Spielman R, McGinnis R, Ewens W. Transmission test for linkage disequilibrium: the insulin gene region and insulin–dependent diabetes mellitus. *Am J Hum Genet* 1993; **52**: 506–16.

77 Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* 1999; **65**: 1170–77.

78 Dudbridge F, Koeleman B, Todd J, et al. Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet* 2000; **66**: 2009–12.

79 Schaid D. General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996; **13**: 423–49.

80 Cordell H, Clayton D. A unified stepwise regression procedure 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* 2002; **70**: 124–41.

81 Martin E, Monks S, Warren L, et al. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hu Genet* 2000; **67**: 146–54.

82 Dudbridge F. Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol* 2003; **25**: 115–21.

83 Laird N, Horvath S, Xu X. Implementing a unified approach to family based tests of association. *Genet Epidemiol* 2000; **19** (suppl 1)**:** S36–S42.

84 Lake S, Blacker D, Laird N. Family-based tests of association in the presence of linkage. *Am J Hum Genet* 2000; **67**: 1515–25.

85 Lunetta K, Faraone S, Biederman J, et al. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum Genet* 2000; **66**: 605–14.

86 Horvath S, Xu X, Laird N. The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet* 2001; **9**: 301–06.

87 Abecasis G, Cardon L, Cookson W. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 2000; **66**: 279–92.

88 Abecasis G, Cookson W, Cardon L. Pedigree tests of transmission disequilibrium. *Eur J Hum Genet* 2000; **8**: 545–51.

89 Darvasi A, Soller M. Selective DNA pooling for determination of linkage between a molecular markers and a quantitative trait locus. *Genetics* 1994; **138:** 1365–73.

90 Barcellos L, Klitz W, Field L, et al. Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet* 1997; **61**: 734–47.

91 Bader J, Bansal A, Sham P. Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. *GeneScreen* 2001; **1**: 143–50.

92 Barratt B, Payne F, Rance H, et al. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet* 2002; **66**: 393–405.

93 Seltman H, Roeder K, Devlin B. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 2001; **68**: 1250–63.

94 Hotelling H. The generalization of Student's ratio. *Ann Math Stat* 1931; **2**: 360–78.

95 Fan R, Knapp M. Genome association studies of complex diseases by case-control designs. *Am J Hum Genet* 2003; **72**: 850–68.

96 Lin S, Chakravarti A, Cutler DJ. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* 2004; **35**: 1181–88.

97 Stephens M, Smith N, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–89.

98 Fallin D, Schork N. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 2000; **67**: 947–59.

99 Clayton D. SNPHAP, a program for estimating frequencies of haplotypes of large numbers of diallelic markers from unphased genotype data from unrelated subjects. http://www-gene.cimr.cam.ac.uk/clayton/software/ (accessed Aug 4, 2005).

100 O'Connell J. Zero-recombinant haplotyping: application of fine mapping usings SNPs. *Genet Epidemiol* 2000; **19**: S64–S70.

101 Epstein M, Satten G. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003; **73**: 1316–29.

102 Michalatos-Beloin S, Tishkoff S, Bentley K, et al. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res* 1996; **24**: 4841–43.

103 Eitan Y, Kashi Y. Direct micro-haplotyping by multiple double PCR amplifications of specific alleles (MD-PASA). *Nucleic Acids Res* 2002; **30**: e62.

104 Piegorsch W, Weinberg C, Taylor J. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994; **13**: 153–62.

105 Weinberg C, Umbach D. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* 2000; **152**: 197–203.

106 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* 2004; **96**: 434–42.

107 Thomas D, Clayton D. Betting odds and genetic associations. *J Natl Cancer Inst* 2004; **96**: 421–23.

108 Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 2002; **23**: 70–86.

109 Storey J, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci US A* 2003; **100**: 9440–45.

110 Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* 2003; **164**: 829–33.

111 Davey Smith G, Ebrahim S. Mendelian randomisation. *Int J Epidemiol* 2003; **32**: 1–22.