

# Family-based designs in the age of large-scale gene-association studies

Nan M. Laird and Christoph Lange

**Abstract** | Both population-based and family-based designs are commonly used in genetic association studies to locate genes that underlie complex diseases. The simplest version of the family-based design — the transmission disequilibrium test — is well known, but the numerous extensions that broaden its scope and power are less widely appreciated. Family-based designs have unique advantages over population-based designs, as they are robust against population admixture and stratification, allow both linkage and association to be tested for and offer a solution to the problem of model building. Furthermore, the fact that family-based designs contain both within- and between-family information has substantial benefits in terms of multiple-hypothesis testing, especially in the context of whole-genome association studies.

## Linkage analysis

A method for localizing genes that is based on the co-inheritance of genetic markers and phenotypes in families over several generations.

## Association studies

A gene-discovery strategy that compares allele frequencies in cases and controls to assess the contribution of genetic variants to phenotypes in specific populations.

## Candidate gene

A gene for which there is evidence, usually functional, for a possible role in a disease or trait of interest.

## Power

The ability of a study to obtain a significant result if this result is true in the underlying population from which the study subjects were sampled.

Department of Biostatistics,  
Harvard School of Public  
Health, Boston,  
Massachusetts 02115, USA.  
Correspondence to: N.M.L.  
e-mail: laird@hsph.  
harvard.edu  
doi:10.1038/nrg1839

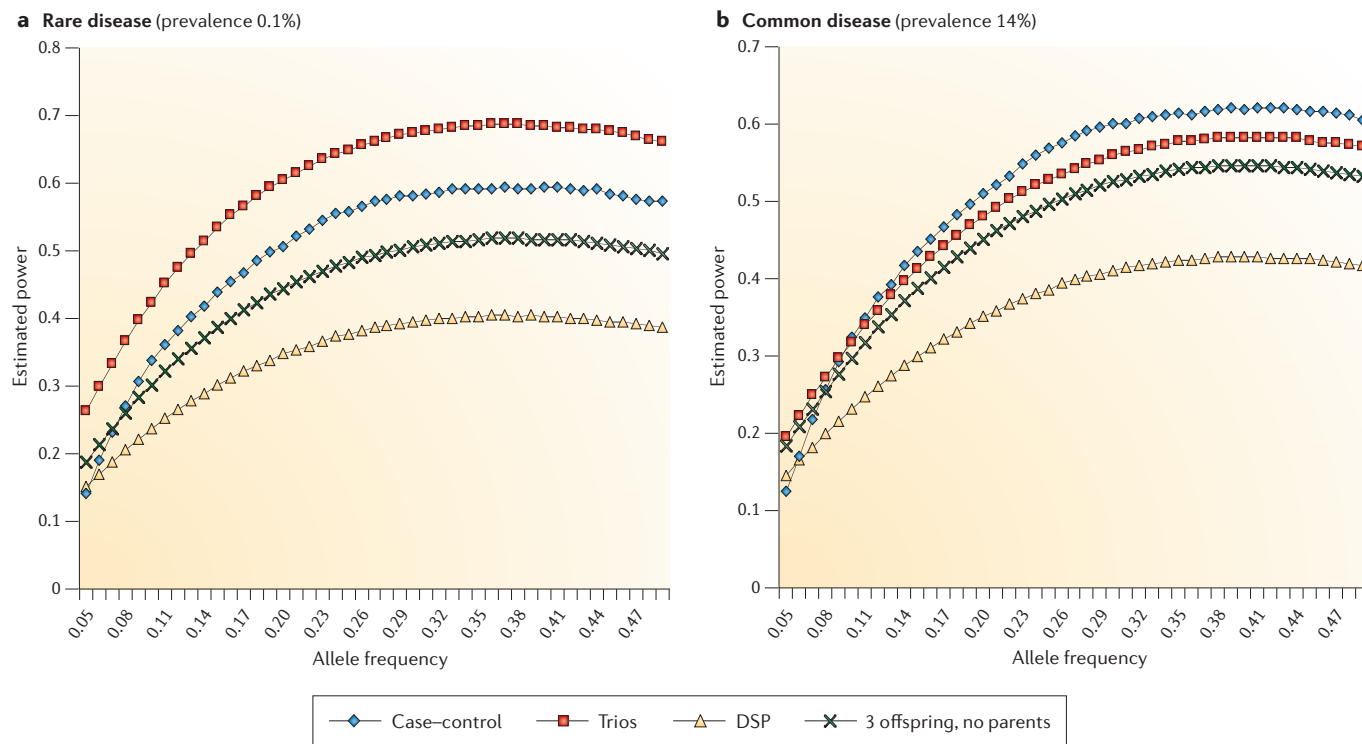
*'If you cannot get rid of the family skeleton, you may as well make it dance' (G.B. Shaw)*

For a long time, family studies were the *sine qua non* in genetics. In recent years, with the progress of the **Human Genome Project**, genetic markers that span the entire human genome have enabled widespread mapping efforts based on linkage analysis using families with multiple affected individuals, leading to the discovery of many genes for Mendelian diseases and traits. However, linkage studies have had only limited success in identifying genes for more complex diseases, such as heart disease, asthma, diabetes and psychiatric disorders. Along with improvements in genotyping technology, this has shifted the focus of gene mapping in humans to association studies, which use large numbers of SNPs or other markers that are genotyped in known linkage regions or candidate genes. Genetic association studies provide greater power and resolution of location than linkage studies<sup>1</sup>, offering renewed hope for mapping complex diseases and traits.

Although association studies have increasingly reported positive results, the number of replications of these findings is disappointingly low<sup>2</sup>. This can be attributed to a number of reasons: low statistical power, multiple-hypothesis testing, variability in study designs, phenotype definition and/or statistical modelling and population substructure<sup>3</sup>. Given that genome-wide association studies, which are now becoming possible, involve hundreds of thousands of markers, these issues become even more important.

Two fundamentally different designs are used in genetic association studies: those that use families and population designs that use unrelated individuals (case-control and case-cohort studies). We believe that the population and family designs, which have different strengths and weaknesses, should be viewed as complementary and not as competitive in the effort to overcome the challenges of association studies for complex diseases.

In terms of statistical power, the differences between the two approaches are generally small (when the use of trios in family designs is compared to case-control studies)<sup>4,5</sup> (FIG. 1). The recruitment of probands and their relatives in family-based association studies usually requires more resources in terms of time and money than that of unrelated subjects in population-based studies. Furthermore, more genotyping might be required for family-based studies, and together these factors have increased the popularity of population designs over family-based studies. An important exception is studies of childhood diseases/disorders, in which it might be easier to recruit parents than suitable controls. However, unlike population-based studies, family-based designs are robust against population substructure, and significant findings always imply both linkage and association. Furthermore, studies that use families offer a solution to the problems of model building and multiple-hypothesis testing, which are important issues in tests of association, and will become more pressing with the advent of genome-wide association studies.



**Figure 1 | Power comparison between case-control studies and family-based designs.** The estimated power levels for a case-control study with 200 cases and 200 controls are compared with those for various family-based designs: 200 trios (of an affected offspring plus parents); 200 discordant sibling (sib) pairs (DSPs; one affected and one unaffected) without parents; 200 '3 discordant offspring (at least 1 affected, at least 1 unaffected) and no parents'. Discordant-sib pair designs have 50% less power than case-control designs, as has been previously noted<sup>7,8</sup>. For the rare diseases (a), trio designs are more powerful than case-control designs. For common diseases (b), case-control designs are slightly more powerful than trio designs and designs with 3 discordant sibs. Although it is not shown here, for larger-effect sizes (for example, odd ratios greater than 2), unaffected probands contain more information, and the DSP design can achieve power levels that are similar to those of trios designs<sup>14</sup>. The power calculations for both the family designs and the case-control designs were done in PBAT (v3.3) using Monte-Carlo simulations.

**Multiple-hypothesis testing**  
Many different statistical tests are used on the same sample; for example, many genetic markers might be tested against many different phenotypes. Failure to account for multiple testing inflates the study-wide type-1 error rate.

**Population substructure**  
Characteristics of a population, such as admixture, population stratification and/or inbreeding, which might distort the distribution of the standard association statistics, leading to increased type-1 error and/or decreased power.

**Genome-wide association studies**  
Studies designed to look for association between disease and a dense set of markers covering the entire genome.

**Case-control study**  
An epidemiological study design in which cases with a defined condition and controls without this condition are sampled from the same population. Risk-factor information is compared between the two groups to investigate the potential role of these in the aetiology of the condition.

Here, we review the basics of family-based designs, starting with the transmission disequilibrium test (TDT), and then emphasize numerous recent advances which make these designs increasingly desirable. Various extensions have increased the power and generalizability of these designs to take into account factors such as missing parents because of late-onset disease, quantitative traits and the use of additional siblings (sibs). We focus on non-parametric extensions of the original TDT approach, the so-called 'family-based association tests' (FBATs), for several reasons. The FBAT approach readily incorporates additional features such as general pedigrees, missing founders and so on, without compromising robustness; it is easy to generalize to more complex phenotypes that characterize complex diseases and it has distinct advantages in handling the multiple-comparisons problem. We will also outline likelihood-based approaches to extending the TDT. Finally, we discuss issues that are specific to genome-wide association studies.

**The transmission disequilibrium test**

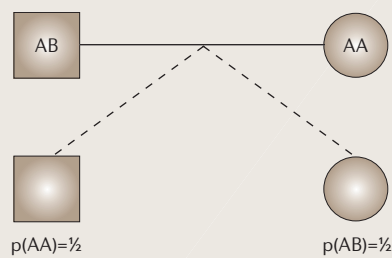
The simplest family-based design for testing association uses genotype data from trios, which consist of an

affected offspring and his or her two parents (BOX 1). The idea behind the TDT is intuitive: under the null hypothesis, Mendel's laws determine which marker alleles are transmitted to the affected offspring. The TDT compares the observed number of alleles that are transmitted with those expected in Mendelian transmissions. The assumption of Mendelian transmissions is all that is needed to ensure valid results of the TDT and the FBAT approach<sup>6</sup>. An excess of alleles of one type among the affected indicates that a disease-susceptibility locus (DSL) for a trait of interest is linked and associated with the marker locus.

Originally, the TDT was used to test for linkage in the presence of association. However, because both linkage and association between the trait and the marker have to be present for the TDT to reject the null-hypothesis<sup>7</sup> (BOX 1), the TDT is now typically used as a test for association<sup>8</sup>. This dual-alternative hypothesis also means that the TDT avoids false positives that arise when association is present but linkage is not, as might happen in the presence of admixture and/or population stratification. Moreover, the attractiveness of a rejection of the null hypothesis for a particular marker is that it

Box 1 | Trio designs — the TDT

Family trios are the basis of the transmission disequilibrium test (TDT)<sup>60</sup>. This test compares the observed number of alleles of type A that are transmitted to the affected offspring with those expected from Mendelian transmissions. An excess of A (or B) alleles among the affected offspring indicates that a disease-susceptibility locus (DSL) for the trait is in linkage and in linkage disequilibrium (LD) with the marker locus.



For the example in the figure, the mother can only transmit the A allele because she is homozygous for A. Such a parent is not informative about transmissions to affected offspring. However, the father transmits A and B with equal frequency, yielding offspring with AA or AB genotypes with equal frequency. With AB, BB parents, we expect to see genotypes AB and BB with equal frequency, and with AB, AB parents, we expect to see genotypes AA, AB, BB with frequencies of  $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ . The TDT discards all homozygous parents and just looks at transmissions from a heterozygous parent to an offspring. Assuming the null hypothesis is correct, each transmission of A occurs with a probability of  $\frac{1}{2}$ , and when there are  $n_{\text{het}}$  heterozygous parents, the distribution of the number of A alleles that are transmitted to affected offspring is binomial ( $n_{\text{het}}, \frac{1}{2}$ ).

There are three possible null-hypotheses for the TDT:

- $H_0$ : No linkage in the presence of association (used in the follow up of case-control association studies)
- $H_0$ : No linkage and no association (used for candidate-gene studies without having obtained a previous linkage signal, or for genome-wide association studies)
- $H_0$ : No association in the presence of linkage (used for the follow up of linkage signals)

There is only one testable hypothesis:

- $H_A$ : The marker is both linked and associated with a DSL affecting the trait

If there is linkage but no association, the marker and the DSL will tend to be transmitted together, but different marker alleles will be transmitted with the DSL in different families. This results in no overall association of a particular allele that is transmitted with the trait. If there is association between the marker and a DSL but no linkage there is no tendency for the marker and the DSL to be transmitted together to offspring. In this case, one would not expect to see an excess of a particular allele transmitted in affected offspring.

implies linkage disequilibrium (LD) between the marker and a DSL; that is, the association is due to lack of recombination between the two loci rather than population stratification. Because LD declines rapidly as the distance between two loci increases, the presence of LD indicates that the marker is tightly linked to a DSL. This makes association studies more valuable than linkage studies in pinpointing a narrow region where a DSL might lie.

The TDT is completely non-parametric — its validity does not require the proper specification of a disease model or assumptions about the distribution of the disease in the population. It is therefore also robust to potential misspecification of any features of the disease model or trait distribution. However, there are numerous cases in which the original TDT cannot be applied without extension: missing parents, general pedigrees, complex phenotypes and haplotypes with missing phase, for example. We argue here that extensions to the TDT should maintain these key features, namely robustness to population stratification and robustness to potential misspecification of the phenotype distribution. This is

the basis for the FBAT approach<sup>9,10</sup>, which is a widely used extension of the TDT. In the remainder of the review, we use the term FBAT to denote this particular approach.

**Generalizing the trio design: the FBAT approach**

The key to generalizing the TDT test to the FBAT approach is putting it into a general framework that both exposes the features that make it so robust and allows its easy extension to more general situations. First, a general formula is specified that shows how the FBAT statistic is computed from the available data. The distribution of the test statistic when the null hypothesis is true must then be described so that valid *p*-values can be computed. A natural basis for a test statistic of association is the covariance between genotype and phenotype. In the family-based setting, however, both the trait and genotype variables are centred in an unconventional way to provide flexibility for different sampling designs and to adjust for potential admixture and/or stratification.

**Defining the FBAT statistic.** Let *X* denote a variable that translates an offspring's genotype to a numeric value — the coded offspring genotype. For example, *X* might count the number of A alleles in an offspring genotype. Let *P* denote the genotype of the offspring's parents, and *T* denote the coded offspring trait. We define *T* as *Y* -  $\mu$ , where *Y* is the phenotypic variable and  $\mu$  is a fixed, pre-specified value that depends on the nature of the sample and phenotype. *Y* can be a measured variable, such as body-mass index, or a zero-one (all-or-nothing) indicator of disease, for example, obesity. The covariance statistic we use in the FBAT test is:

$$U = \Sigma T * (X - E(X|P)) \tag{1}$$

where *U* is the covariance, *E*(*X*|*P*) is the expected value of *X* computed under the null hypothesis, and summation is over all offspring in the sample. Mendel's laws underlie the calculation of *E*(*X*|*P*) for any null hypothesis given in BOX 1. Centring *X* by its expected value conditional on parental genotypes has the effect of removing contributions from homozygous parents and protecting against population stratification<sup>11</sup>.

**Specifying the distribution under the null hypothesis.**

To complete the specification of the test statistic, we derive its distribution under the null hypothesis. Following the TDT approach, we treat the offspring genotypes, *X*, as random, but the trait, *T*, and the parental genotypes are fixed<sup>9,10</sup>. Holding the trait fixed means we do not need to make distributional assumptions about the trait, and holding the parental genotypes fixed means we do not need to make assumptions about allele distributions in the population. Because *X* is centred around *E*(*X*|*P*), *U* has an expected value of zero under the conditional distribution. When there is more than one offspring, the computation of the variance of *U* depends on the null hypothesis (see below), but for any null hypothesis, the FBAT is defined by dividing *U*<sup>2</sup> by its variance, which is again computed

**Case-cohort study**

Similar to a case-control study, except both cases and controls are drawn from an existing cohort of subjects who are being followed to study a broad spectrum of diseases and risk factors.

**Proband**

In a family study, this is the individual who is first identified in the family as having the disease under study.

**Odds ratio**

The odds of exposure to the susceptible genetic variant in cases compared with controls. If the odds ratio is significantly greater than one, then the genetic variant is associated with the disease.

## Monte Carlo

A method for obtaining a  $p$ -value for a test statistic by drawing repeated samples from the null distribution of the data, computing the  $p$ -value for the same statistic for each sample, and comparing the observed  $p$ -value to the distribution of  $p$ -values obtained from the samples.

## Likelihood

A statistical model for analysing data that requires specifying a particular form for the distribution of the data.

## Admixture

This occurs when two or more subpopulations inbreed, so that two randomly chosen individuals in the population might have different degrees of genetic heritage from the original subpopulations.

under the appropriate null hypothesis by conditioning on  $T$  and  $P$  for each offspring. Given a sufficiently large sample, that is, at least 10 informative families, the FBAT statistic has a  $\chi^2$ -distribution with 1 degree of freedom.

The FBAT statistic is exactly the same as the TDT statistic under the following conditions: both parents are genotyped;  $T = 1$  for affected offspring and zero otherwise;  $X$  counts the number of a specific allele; and the null hypothesis specifies no linkage. By changing how the trait  $T$  is defined, either via  $Y$  or  $\mu$  or both, we can include unaffected offspring, fit alternative traits and multiple traits. Changing how  $X$  is defined allows us to test alternative genetic models (for example, recessive or dominant) and to incorporate multiple alleles at a marker.

**The general FBAT statistic.** This simple description for the FBAT statistic can be readily generalized using the conditioning approach to situations that involve arbitrary pedigrees, missing parents/founders, haplotypes, different null hypotheses, and so on. These situations are

outlined in BOX 2 along with extensions to handle multiallelic markers, arbitrary genetic models and more complex phenotypes, and these generalizations and extensions are discussed in detail in the following sections.

**Likelihood extensions.** There are numerous extensions to the basic TDT that are based on likelihood models; some of the most popular are outlined in BOX 3.

## Pedigrees, missing founders and haplotypes

There are several extensions to the basic TDT in which the simple binomial model that is described in BOX 1 does not apply. Many of these extensions can be characterized as situations in which the complete distribution of offspring genotypes under the null hypothesis of interest depends on unknown factors (for example, missing parents, missing phase or recombination fraction) that are not of direct interest. Such factors are often referred to as 'nuisance' parameters.

A standard statistical approach to handling nuisance parameters is to find sufficient statistics for them; the distribution of the full data, conditioning on the sufficient statistics, does not depend on the nuisance parameters<sup>12</sup>. This conditional distribution can then be used in testing the null hypothesis, without the need to estimate or specify the nuisance parameters. This forms the basis of the FBAT approach. This strategy is in contrast to likelihood approaches (BOX 3), which generally estimate the nuisance parameters from observed data; this can make them susceptible to confounding by population sub-structures<sup>13</sup>.

**General nuclear families and pedigrees.** Families come in many shapes and sizes and it might be desirable to include additional family members for various reasons. With a rare disease, it is most efficient to sample affected offspring and their parents, but with more common disorders, such as obesity, unaffected offspring can also contribute information<sup>14</sup>. Many family studies are based on pre-existing cohorts of families with more complex structures, which could have been ascertained for linkage studies, and some designs that include parents and grandparents have been suggested<sup>15</sup>. Genotyping additional family members is useful if founder genotypes are missing (see below).

Whether or not founders are known, the distribution of multiple offspring genotypes in pedigrees depends on which null hypothesis is tested. In testing the null hypothesis of no linkage, with or without association, there is no difficulty in incorporating multiple offspring; transmissions from all parents to all offspring are independent and families with multiple affected members can be treated as multiple trios. Pedigrees with founders of known genotype are a simple extension of nuclear families: one simply uses Mendel's first law (assuming a null hypothesis of no linkage and no association) to compute the joint genotype distribution of all offspring in the pedigree. For example, in the pedigree in BOX 4, the two related trios can be treated as two independent trios, but there is potentially a gain of 100% in the number of informative transmissions by treating the two offspring as arising in a pedigree.

## Box 2 | The general FBAT statistic

The general family-based association test (FBAT) statistic<sup>61</sup> is defined by:

$$U = \sum T_{ij}(X_{ij} - E(X_{ij}|S_i)) \quad (1)$$

where  $i$  indexes pedigree,  $j$  indexes non-founders in the pedigree, and summation is over all  $i$  and  $j$ ;  $T_{ij}$  is a coding function for the trait of interest, and  $X_{ij}$  is a coding function for the genotype. The coded genotype is chosen to reflect the selected mode of inheritance; for example, additive, dominant and recessive. Under the null hypothesis, the expected marker score,  $E(X_{ij}|S_i)$ , is computed conditional on the sufficient statistic<sup>10</sup>, which is denoted by  $S_i$ .

Typically, a phenotypic residual is used for the coded trait — that is,  $T_{ij} = (Y_{ij} - \mu)$ , where  $Y_{ij}$  is the original phenotype and  $\mu$  a user-defined offset parameter. For example, with quantitative traits,  $\mu$  should typically be the phenotypic sample mean. For complex phenotypes, for example, time-to-onset, longitudinal measurements or multivariate traits, more complex coding functions for  $T_{ij}$  can be derived<sup>43,45,46</sup>. The expected value of the coded genotype,  $X_{ij}$ , is computed conditional on the sufficient statistic  $S_i$  for any genetic information about the founders of the family, as described in REF. 10. For trios, the sufficient statistic is equivalent to the parental genotypes. The distribution of the FBAT statistic under the null hypothesis is obtained by treating the  $X_{ij}$  as random, but conditioning on the trait,  $T_{ij}$ , and the sufficient statistic. Because  $E(U) = 0$  by construction under  $H_0$ ,  $U$  can be normalized by its standard deviation ( $Z$ ), which can again be computed under the conditional distribution of offspring genotype, given offspring trait and  $S_i$  as follows:

$$Z = U / \sqrt{\text{var}(U)}, \text{ or equivalently, } \chi^2_{\text{FBAT}} = U^2 / \text{var}(U) \quad (2)$$

where

$$\text{Var}(U) = \sum_i \sum_{j,j'} T_{ij} T_{ij'} \cdot \text{cov}(X_{ij}, X_{ij'} | S_i, T_{ij}, T_{ij'}) \quad (3)$$

and  $\text{cov}(X_{ij}, X_{ij'} | S_i, T_{ij}, T_{ij'})$  is computed conditional on the traits and the sufficient statistics, assuming the null hypothesis is true. Note that this covariance only depends on  $S_i$  and not the traits when no linkage is part of the null hypothesis. For testing no association in the presence of linkage, an empirical variance can be used to estimate  $\text{var}(U)$ <sup>19</sup>.

For large samples,  $Z$  is approximately distributed as  $N(0,1)$ , and  $\chi^2_{\text{FBAT}}$  is distributed as approximately  $\chi^2$  on one degree of freedom. With multiallelic markers or haplotypes, a multiallelic version of the FBAT statistic is obtained by taking  $X$  to be a vector; each element of  $X$  codes for a specific allele or haplotype. Then  $U$  will be a vector,  $\text{var}(U|S)$  a variance/covariance matrix, and the test statistic is the quadratic form  $U^T \text{var}(U|S)^{-1} U$ , which is distributed as  $\chi^2$  with degrees of freedom equal to the rank of  $\text{var}(U|S)$ <sup>61</sup>. If the haplotype phase is unknown, the coding function and the computation of the expected value of  $X_{ij}$  and its variance are modified by weighting the possible phases.

## Box 3 | Likelihood methods

The likelihood method specifies a probability density for the observed data as a function of genotype; either likelihood-ratio or score tests are used to test the hypothesis of no association.

One type of likelihood method for case–parent trios creates ‘pseudo-controls’ using the non-transmitted alleles<sup>62</sup>, and constructs a conditional logistic regression likelihood; under a log-additive relative-risk model, the likelihood-ratio test of this approach is equivalent to the transmission disequilibrium test (TDT). The approach has been extended to haplotypes<sup>25</sup>, gene–environment interactions and gene–gene interactions<sup>26,63</sup>.

A second approach uses multinomial likelihoods. The likelihood for a case–parent trio is factored as:

$$L = L_c L_p \quad (1)$$

$L_c$  is the probability density of the child’s genotype conditional on the parents’ genotype and the child’s disease status, and  $L_p$  is the probability density of the parental genotypes, given the child’s disease status. With parental data, all information on association is contained in  $L_c$  and likelihood-ratio tests based on  $L_c$  are optimal<sup>64–66</sup>.

Score tests are generally more popular than likelihood-ratio tests. They can easily be extended to accommodate multiple offspring, including unaffected, without the need for distributional assumptions under the alternative<sup>47</sup>. Score tests based on the multinomial model have been generalized to encompass complex phenotypes<sup>47</sup>.

The family-based association test (FBAT) statistic is also a score test under general assumptions about the distribution of the offspring phenotypes<sup>61,67</sup>. For trios, the score test based on  $L_c$  and the FBAT are identical. The approaches diverge in the treatment of missing parental data. Here, the FBAT approach replaces conditioning on parents by conditioning on the sufficient statistics  $S$ , which maintains the independence of the test from allele frequencies estimates. The likelihood approach estimates the probabilities of parental genotypes from the likelihood,  $L_p$ , and averages  $E(X|P)$  over the estimated distribution of parental genotypes for probands whose parents are missing. Likelihood approaches are more efficient, but the efficiency gain relies on the assumption that their parents’ genotypes can be estimated unbiasedly. Using Poisson likelihoods, extensions of the multinomial-model approach also incorporate parental imprinting, gene–environment interaction and quantitative phenotypes<sup>68–74</sup>.

A popular likelihood approach for the quantitative transmission disequilibrium test assumes the trait follows a normal distribution, with the mean depending linearly on  $X$ <sup>11,32,33</sup>. Inferences are based on the normal likelihood for phenotype given genotype, rather than genotype given phenotype. A correction for population substructure is made by incorporating  $E(X|P)$  into the mean model (equation 2). Because the approach requires the correctness of the likelihood function, the likelihood ratio test can be sensitive to distributional assumptions and ascertainment conditions; the model does not incorporate excess variation, which can arise in the presence of population admixture and can lead to anti-conservative tests<sup>13</sup>.

Likelihood-based approaches offer the possibility of more sophisticated tests, for example, nested models, and can be more efficient because they incorporate the between-family information. However, model-based validation and screening is much more easily carried out in the FBAT approach, using the between-family information for screening, and the within-family component for testing.

**Population stratification**

The presence in a population of distinct strata or groups that show limited inbreeding; they might have different disease rates and distinct allele-frequency distributions. Failure to control for the stratification can invalidate tests of association.

**Linkage disequilibrium**

(LD). This occurs when alleles at two different loci are associated in a population because of tight linkage.

**Haplotype**

A set of alleles at different loci that are present together on the same chromosome.

**Phase**

The arrangement of alleles at multiple loci on homologous chromosomes. For example, in a diploid individual with genotype  $Aa$  at one locus and genotype  $Bb$  at another locus, possible linkage phases are  $BA/ba$  or  $Ba/bA$ , where ‘/’ separates the two homologous chromosomes.

**Covariance**

A measure of association between two variables that characterizes the tendency for the two variables to co-vary around their mean in a systematic way.

**Informative families**

Families that make a contribution to the FBAT test; that is, those with at least one heterozygote parent, or sibships with at least two distinct genotypes.

**Nuisance parameters**

Parameters that are not the primary focus of a statistical analysis, but for which misspecification might lead to biased results, for example, allele frequency in association tests.

**Sufficient statistics**

A data reduction function that retains all information about an unknown parameter; they are used to remove the dependence of a test on nuisance parameters that are unknown or difficult to model.

**Testing for association with multiple offspring in the presence of linkage.** It is common to use FBATs for fine mapping under a linkage peak. In some cases, it might be feasible to use the same data set first for testing linkage and then for association, using additional markers in the linked region. In this context, the null hypothesis of no association in the presence of linkage is appropriate. If only families with a single offspring are used, the TDT remains a valid test. However, transmissions from the same parent to multiple offspring will be correlated because of patterns of identity-by-descent (IBD). Therefore, tests of association that treat multiple offspring as independent are not valid<sup>16</sup>.

The full distribution of transmission to multiple offspring depends on the unknown recombination fraction between the marker and the proposed DSL<sup>17</sup>, as well as observed traits. However, IBD status among sibs forms the sufficient statistics for estimating the recombination fraction, so conditioning on observed patterns of IBD will result in a distribution for transmissions to multiple offspring that does not depend on the recombination parameter. When the parents are genotyped, such a distribution can be constructed using permutation: for each heterozygous parent, the values of their two alleles can be permuted independently, while fixing the allele that

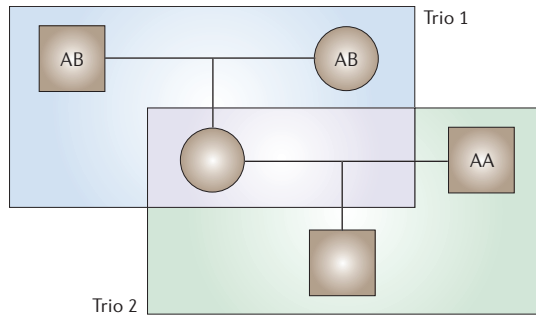
is transmitted to each child. Therefore, A and B will be transmitted equally often, but the observed patterns of IBD will remain fixed. This approach can also be extended to situations in which parents are missing. Although the resulting distributions are simple to obtain<sup>18</sup>, a simpler approach of estimating the variance of the correlated transmissions empirically is more commonly used<sup>16,19</sup>.

The effect of recombination on correlation between transmissions to sibs is often modest, and conditioning on IBD reduces the size of the possible outcome space, resulting in a loss of information. In practice, however, the loss is usually small, except in cases of large nuclear families without parental genotypes. Therefore, our recommendation is to test the null hypothesis of no association and no linkage, unless the same sample that was used to test association has previously been used to demonstrate linkage in the region.

**Missing parents.** Missing parents present an obvious difficulty for the TDT and can be common in studying disorders that occur in later life. There have been numerous proposals for extending the TDT to handle this problem; several likelihood approaches are discussed in BOX 3. Other approaches<sup>16,20–22</sup> compensate for missing parents by comparing genotypes in affected

Box 4 | The use of pedigrees in family-based designs

By conditioning on all founders, the family-based association test (FBAT) can be extended readily to incorporate pedigrees. In relation to the pedigree shown to the right, the table shows the informativeness of conditioning on founder genotypes, as opposed to conditioning only on parents, separately for each trio, when computing transmissions to the two offspring. When the genotype probability is 1, the family is non-informative by itself. The informativeness ratio is the ratio of the number of informative transmissions when the extended pedigree is analysed versus that when the trios are analysed independently. With two heterozygote parents (see figure), the top trio in the pedigree above gives maximal information, but the bottom trio is only informative when the mother is AB. Computing the distribution of the 2 offspring based on the founders' genotypes only (that is, the genotypes of the two grandparents), we can increase the number of informative transmissions by 100%, relative to treating them as independent trios.



Offspring genotypes		Probability of genotype			Informativeness ratio
Trio 1	Trio 2	Based on both trios	Based on trio 1 only	Based on trio 2 only	
AA	AA	¼	¼	1	2
AB	AB	¼	½	½	1
AB	AA	¼	½	½	1
BB	AB	¼	¼	1	2

Confounding

A measure of the association between a disease and a risk factor is distorted because other variables, associated with both the disease and the risk factor, are not controlled for in the calculation of the measure of association.

Likelihood-ratio tests

A class of statistical tests obtained by comparing the likelihood statistic under the alternative hypothesis to the likelihood under the null hypothesis.

Score tests

A class of statistical tests that are derived from a likelihood model and are generally easier to compute than likelihood-ratio tests.

Identity-by-descent

(IBD). An allele shared by two related individuals is said to be identical-by-descent if the allele is inherited from the same common ancestor.

offspring and unaffected sibs. Conditioning approaches for missing parents are described in two studies<sup>10,23</sup>; the first study conditions on being able to reconstruct both parental genotypes, and the second conditions on the sufficient statistic for missing parents. Missing-parent designs are generally less efficient than trios when disease prevalence is low, but discordant-sib trios compensate well for missing parents when disease prevalence is high (FIG. 1). Discordant-sib pairs (DSPs) can also be more efficient than parents with high prevalence and larger-effect sizes<sup>14</sup>. Note that DSPs only require genotyping two individuals as opposed to three, and can therefore be cost effective as well as powerful.

**Haplotypes.** When testing for association with candidate genes, it is common to genotype multiple SNPs within the gene. Testing each SNP separately leads to multiple testing issues, and will not be efficient when the SNPs are in high LD. One alternative is to test for over- or under-transmission of haplotypes, rather than individual SNPs. However, phase cannot always be determined from the available marker-genotype data. Conditioning on the sufficient statistics for parental phase can be used to form a distribution for haplotype transmissions that does not rely on estimating phase<sup>24,25</sup>. Likelihood approaches<sup>26</sup> for haplotype analysis allow exploration of marker interactions

using nested models. However, in general, extensions for unknown-phase haplotypes are more difficult in the likelihood approach<sup>27</sup>.

Extensions of the TDT involving phenotypes

**Extension to samples of affected and unaffected offspring.** With common diseases or disorders, incorporating unaffected sibs can provide additional information<sup>28,29</sup>. Using a multiplicative genetic model for a common disease, Whitaker and Lewis<sup>30</sup> showed that the power of a test that is equivalent to the FBAT can be maximized by setting the offset  $\mu$  to the population prevalence of the disease. With  $Y$  defined as 1 or 0 for affected/unaffected, this yields coded traits of  $(1 - \mu)$  and  $(-\mu)$ . In the absence of an ascertainment condition (meaning offspring are not selected into the study on the basis of their trait) this optimality holds under any genetic model<sup>31</sup>.

**Extension to quantitative traits.** The most commonly used complex phenotypes in genetic association studies are quantitative traits. In generalizing the TDT to handle these traits, it is important to note that offspring now provide both phenotypic and genotypic variation. This offers the possibility of reversing the role of phenotype and genotype — that is, treating the phenotype as the random response and the genotype as the fixed predictor. Ordinary linear regression of  $Y$  on  $X$  can then be used to test for association, giving equal weight to trios with the same  $X$ , no matter if there are 0, 1 or 2 heterozygote parents. This can introduce bias in the presence of population substructure, so to circumvent this the linear-regression model of  $Y$  on  $X^{11}$  can be modified to fit:

$$E(Y) = m + a_w * (X - E(X|P)) + a_b * E(X|P) \tag{2}$$

Here,  $a_w$  measures within-family correlation between phenotype and genotype, and is therefore similar to the FBAT statistic;  $a_b$  measures the between-family — or between-population — correlation. With random population samples and in the absence of any population substructure, the two coefficients should be approximately the same. Several popular approaches, such as the quantitative TDT (QTDT) likelihood approach<sup>32</sup>, are based on model equation 2 or extensions of it<sup>27,33</sup>, sometimes making the additional assumption of normality for the phenotypic distribution. We refer to such approaches as model-based, as the validity of the inference generally requires that the model holds; otherwise levels of type-1 error are not maintained below a suitable threshold in the presence of population substructures<sup>13</sup>.

In contrast, the FBAT approach continues to condition on the trait and parental genotypes, and is non-parametric in the sense that no model or distributional assumptions for the trait are required. The power is optimized with unselected subjects by setting  $\mu$  to the population mean. When this is unknown, the observed phenotype data can be used to determine the offset  $\mu$ <sup>34</sup>.

With quantitative traits, the power of any test will depend on the size of the genetic effect as well as the variation in the trait. Therefore, it is useful to minimize the degree of extraneous variation in the data (owing to

**Permutation**

An approach in which the actual data are randomized many times to generate a distribution of outcomes, so that the fraction of observations with values that are more extreme than the outcome that is observed with the real data reflects the statistical significance.

**Outcome space**

Set of all possible genotype configurations for a specific pedigree that are plausible under Mendelian transmissions, and consistent with the sufficient statistics for parental genotype.

**Discordant sibs**

A family design for testing association that uses a case and his/her unaffected sib.

**Nested models**

A sequence of statistical models, each specifying a different hypothesis, such that each model in the sequence contains one more factor than the preceding model. Nested models are often used to test for the presence of interactions between two or more risk factors.

**Multiplicative genetic model**

A genetic model for penetrance functions that assumes the relative risk for disease given two alleles is the square of the relative risk for disease given only one allele.

**Linear regression**

A statistical method used to test and to describe the linear relationship between two or more variables.

**Type-1 error**

The probability that the null hypothesis is falsely rejected.

**Intermediate phenotypes or endophenotypes**

Measured biological variables intermediate between genotype and external phenotype that can indicate susceptibility to, or manifest as early signs of, a wide range of diseases or disorders.

**Imputed**

A statistical method for handling missing data which replaces the missing values by estimated values.

random error, environmental factors or measurement error). Offspring characteristics, such as age, sex, race and smoking, can be used as covariates in a regression analysis to reduce extraneous variation and improve power<sup>35</sup>. Adjustment in the model-based setting by inclusion of covariates in model equation 2 is straightforward. In the FBAT approach, the quantitative traits are first regressed on the covariates and the offset  $\mu$  is then set to the phenotype that is predicted by the regression model,  $\hat{Y}$ , for each offspring. The coded trait  $T$  is therefore given by the residual ( $Y - \hat{Y}$ ).

Extensive power considerations indicate that both the non-parametric quantitative FBAT approach and the QTD have optimal power for a quantitative phenotype for which no ascertainment condition has been imposed<sup>32,36</sup>. However, with highly ascertained samples, such as discordant-sib designs, quantitative traits should be converted to dichotomous variables and analysed as such<sup>36</sup>.

**Extension to complex phenotypes**

*The challenge of complex phenotypes.* Appropriate modelling of phenotypic information is important, particularly for complex diseases. For example, in asthma<sup>37,38</sup>, chronic obstructive pulmonary disease<sup>39,40,41</sup> and attention-deficit hyperactivity disorder<sup>42</sup>, the definition of the affection status is a binary phenotype that aggregates information from a variety of complex phenotypes. The definition can vary between studies, which should be taken into account when unaffected probands are incorporated in the test statistic. In addition to such dichotomous phenotypes, we might have multiple and/or repeated measures that characterize disease, as well as multiple covariates that might influence the phenotype. Assuming small genetic-effect sizes for complex diseases, the use of intermediate phenotypes or endophenotypes in the association analysis can enhance the statistical power.

The FBAT approach has been extended in several ways to handle more complex modelling issues. For example, REF. 43 describes an approach for multiple traits in which  $T$  is a vector. Time-to-onset versions of the FBAT test have been described, using various codings for the trait<sup>44–46</sup>. Likelihood approaches have also been developed for time-to-onset<sup>47</sup>. In principle, it is straightforward to handle multiple or repeated measures that use likelihood models that are based on an appropriate extension of model equation 2 to adjust for population substructure. In practice, however, the analysis of complex traits has to deal with two major statistical obstacles. The first is the appropriate modelling of the complex phenotypes. In addition to incorporation of covariates, this requires attention to various other modelling factors, including which aspects of the phenotypes are optimal for analysis (for example, average or change over time; early versus late time-to-onset; combinations or clusters of multiple traits), the need to account for environmental correlation and the need to handle missing phenotypes. Methods that are akin to model validation that allow hypothesis generation and model selection to be carried out independently of model testing offer one solution to this problem. The second problem relates to multiple-hypothesis testing: regardless

of model complexity, it is usually desirable to consider a variety of different combinations of phenotypes, genetic models and markers.

*A general approach to complex phenotypes.* In view of the multiple-comparison and model-selection issues that arise in modelling complex traits, it is useful to have a mechanism for exploratory model development and/or screening, followed by an independent confirmatory step. Here, we consider a general approach to screening, model selection and/or hypothesis generation that is based on separating family data into two independent partitions that correspond to the population information and the within-family information. This allows one part of the data to be used for model building and selection, and the other part for confirmatory testing.

The full distribution for family data consists of a joint distribution for all offspring phenotypes,  $Y$ , all offspring genotypes,  $X$ , and all parental genotypes,  $P$  (or more generally, the sufficient statistics for parental or founder genotypes,  $S$ ). The joint distribution is partitioned into two independent parts:

$$P(Y, X, S) = P(X|Y, S)P(S, Y) \quad (3)$$

Model building, hypothesis generation and screening can be based solely on  $S$  and  $Y$ , so that subsequent hypothesis testing using any test statistic with a distribution that is based on  $P(X|S, Y)$  will be independent of the selected model. Note that equation 3 simplifies further if it is assumed that there is no linkage, as  $P(X|Y, S)$  can then be replaced by  $P(X|S)$ .

To illustrate this approach, consider testing a quantitative phenotype with a single marker. To use a population-based approach, two studies<sup>48,49</sup> proposed a 'conditional-mean model' that was obtained by setting the  $a_w$  to zero in model equation 2:

$$E(Y) = m + a_b * E(X|S) \quad (4)$$

Note that for doubly homozygous parents,  $X = E(X|S)$ ; otherwise, we can think of  $X$  as missing if parents are informative, and  $E(X|S)$  replaces the missing  $X$ . In effect, equation 4 defines a population regression in which some values of  $X$  are imputed using parental information (or the sufficient statistics for parental information if parents are missing); generalization to pedigrees and haplotypes is immediate. Because the regression uses only  $Y$  and  $S$ , all the statistics are statistically independent of any FBAT statistic that is subsequently computed by equation 1. Model equation 4 can be used for any choice of coded genotype, any number of phenotypes and any number of markers. Model selection for confirmatory testing using FBAT can be based on  $p$ -values for testing the null hypothesis that the between-family or between-population correlation ( $a_b$ ) is zero<sup>48,49</sup>. Alternatively, the estimated  $a_b$  can be used to compute the conditional power of the FBAT statistic. The conditional power calculation depends on the effect size, as well as the observed parental genotypes and traits<sup>48,49</sup>. In general, selection that is based on the conditional power is preferable<sup>50</sup>.

**Bonferroni or Hochberg corrections**  
 Statistical methods, proposed by Bonferroni and Hochberg, for controlling type-1 error (false positives) in the presence of multiple testing.

This basic approach has been extended to handle longitudinal and repeated measures (FBAT PC)<sup>35</sup> and multivariate data<sup>51</sup> by using the screening stage to select optimal linear combinations of traits for subsequent testing. Jiang *et al.*<sup>46</sup> proposed a method to determine the genetically relevant age range for time to onset, which is particularly useful for diseases in which an early onset indicates a strong genetic component, whereas a late onset might be attributable to environmental effects.

**Genome-wide association studies**

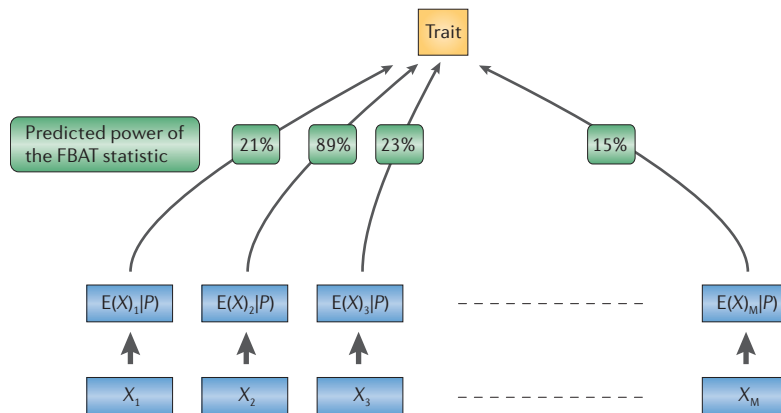
The main scientific obstacle in genome-wide association studies is the hundreds of thousands of SNPs and potential statistical tests that can be computed, resulting in numerous hypothesis-testing issues. To avoid this problem, multi-stage designs have been proposed for case-control studies<sup>8,52</sup>; the number of genotyped SNPs is reduced in each stage of the design, so that genome-wide significance is achieved step by step. The screening approach for family studies that is described

above extends readily to a genome-wide association study with quantitative traits<sup>50</sup> (BOX 5). With family-based designs, the screening procedure uses all families, even the 'non-informative' ones.

Assuming moderate- to low-effect sizes, simulation studies indicate that if a true DSL, or an SNP in LD with a DSL, is included in the data set, it is sufficient to select the highest 10 or 20 SNPs based on the power estimates for further testing, and retain high power for the overall procedure. By contrast, procedures that rank the SNPs based on *p*-values require the selection of many more SNPs to ensure that the true DSL is selected<sup>50,53</sup>. The advantage of family-based screening is that the same data set is used for the screening step and the testing step. This means only one sample needs to be recruited, and replication in other studies serves the purpose of generalizing a significant finding to other populations. Although the screening step relies on population-based analysis and is consequently susceptible to confounding by undetected population substructures, simulation studies indicate that the ranking of the SNPs is relatively well maintained<sup>50</sup>. The strategy has been successfully applied to a 100k SNP scan for obesity in families from the **Framingham Heart Study**. A new candidate gene for body-mass index was discovered that would have been missed by standard approaches (for example, the Bonferroni or Hochberg corrections for multiple testing<sup>54,55</sup>). Using the same genetic model, the finding was replicated in four independent studies, including cohort, case-control and family-based samples<sup>56</sup>.

The promise of whole-genome association scans offers great expectations for genetic association studies. Most projections agree that large samples of individuals will be necessary to separate the wheat from the chaff in these large-genome scans<sup>2,8,50</sup>, no matter what the design. Although it is inescapable that large samples from existing cohort or case-control studies that do not include data on relatives are generally much easier to obtain than large numbers of suitable families, such approaches carry the risk of increased numbers of false results owing to heterogeneity between studies and undetected, subtle population substructures<sup>3</sup>. We believe that the innovative use of the population information contained in family data for screening and hypothesis generation, which allows the establishment of genome-wide significance in just one modestly sized study<sup>56</sup>, coupled with their robustness to population substructure make these family studies competitive. In addition, with approaches for handling pedigrees with missing founders, family data that have already been collected for linkage studies can, in many cases, be recycled for association.

Box 5 | Using family-based designs in whole-genome association studies



The conditional-mean model<sup>35,48-50</sup> can be used to minimize the multiple-testing problem. Here, we take the example of 1 quantitative trait and *M* SNPs. In the first step, which is shown in the figure, the conditional-mean model specifies a linear regression of the phenotype, *Y*, on the expected SNP marker scores, *E*(*X*|*P*) or *E*(*X*|*S*), conditional on the parental genotypes (*P*) or the sufficient statistic (*S*), respectively<sup>11</sup>. The true-offspring genotype is treated as missing. The observed phenotypes and expected marker scores are used to estimate the conditional-mean model. The power depends on the observed parental genotypes and the effect size that is estimated from this model.

In the second step, as illustrated in the table, the *K* SNPs with the highest power estimates are tested for association with the family-based association test (FBAT) statistic at a Bonferroni-adjusted significance level of  $\alpha/K$  where  $\alpha$  denotes the overall-significance level. Because only *K* of the original *M* SNPs have been selected for testing, it is only necessary to adjust for *K* comparisons instead of *M*.

Power rank	Estimated power of FBAT statistic	SNP	<i>p</i> -value of FBAT statistic
1	0.92	3	0.90
2	0.89	100	0.20
3	0.85	25	0.00001
...	...	...	...
<i>K</i>	0.70	53	0.20

**Outlook — the future of family designs**

Although we have outlined several reasons why we feel that family-based designs are useful, there are features that can make them less attractive than their population-based counterparts. One feature is the sensitivity to genotyping errors<sup>57-59</sup>, which can lead to false inferences as the test distribution depends on the assumption that parental genotypes are correct. In the population-based setting, non-differential genotyping

errors will only make tests conservative under the null hypothesis, but with family based tests, random genotyping errors can inflate the false-positive rate, sometimes substantially<sup>8</sup>. This issue will become less important with time, given the constant improvements in genotyping technology.

Population-based samples also have the advantage in that their analysis can largely be implemented by standard commercial software packages, whereas with

family-based designs, the development of software for most methods beyond the TDT has been home-grown. As availability of commercial software with adequate support increases, this will greatly enhance the productivity of family designs.

These future advances will add to the attractiveness of family-based methods, which should prove particularly valuable in light of their important advantages for genome-wide association studies.

1. Risch, N. & Merikangas, K. The future of genetics studies of complex human diseases. *Science* **273**, 1516–1517 (1996).  
**Shows that genome-wide association scans based on trios have greater power than genome-wide linkage scans based on affected sib pairs.**
2. 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).
3. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nature Genet.* **36**, 388–393 (2004).
4. McGinnis, R. General equations for Pt, Ps, and the power of the TDT and the affected-sib-pair test. *Am. J. Hum. Genet.* **67**, 1340–1347 (2000).
5. McGinnis, R., Shifman, S. & Darvasi, A. Power and efficiency of the TDT and case-control design for association scans. *Behav. Genet.* **32**, 135–144 (2002).
6. Zollner, S. *et al.* Evidence for extensive transmission distortion in the human genome. *Am. J. Hum. Genet.* **74**, 62–72 (2004).
7. Ott, J. Statistical properties of the haplotype relative risk. *Genet. Epidemiol.* **6**, 127–130. (1989)  
**Demonstrates the need for linkage and association under the alternative hypothesis for a family-based test.**
8. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
9. Lazerzeroni, L. C. & Lange, K. A conditional inference framework for extending the transmission/disequilibrium test. *Hum. Hered.* **48**, 67–81 (1998).
10. Rabinowitz, D. & Laird, N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* **50**, 211–223 (2000).  
**Generalization of the TDT for general pedigrees, missing parents and arbitrary phenotypes using the approach of conditioning on the sufficient statistic.**
11. Fulker, D. W. *et al.* Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259–267 (1999).  
**Forms the basis of the likelihood approaches for quantitative traits in family-based studies with correction for admixture.**
12. Cox, D. R. & Hinkley, D. V. *Theoretical Statistics* 18–23 (Chapman and Hall, London, 1974).
13. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet.* **38**, 203–208 (2006).
14. Laird, N. *et al.* in *Respiratory Genetics* (eds Silverman, E. *et al.*) 27–46 (Hodder Arnold, Boston, 2005).
15. Weinberg, C. R. Studying parents and grandparents to assess genetic contributions to early-onset disease. *Am. J. Hum. Genet.* **72**, 438–447 (2003).
16. Martin, E. R., Kaplan, N. L. & Weir, B. S. Tests for linkage and association in nuclear families. *Am. J. Hum. Genet.* **61**, 439–448 (1997).
17. Thompson, G. Mapping disease genes: family-based association studies. *Am. J. Hum. Genet.* **57**, 487–498 (1995).
18. Schneider, K., Laird, N. & Corcoran, C. Exact family-based association tests for biallelic data. *Genet. Epidemiol.* **29**, 185–194 (2005).
19. Lake, S. L., Blacker, D. & Laird, N. M. Family-based tests of association in the presence of linkage. *Am. J. Hum. Genet.* **67**, 1515–1525 (2000).
20. Curtis, D., Miller, M. B. & Sham, P. C. Combining the sibling disequilibrium test and transmission/disequilibrium test for multiallelic markers. *Am. J. Hum. Genet.* **64**, 1785–1786 (1999).
21. Horvath, S. & Laird, N. M. A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am. J. Hum. Genet.* **63**, 1886–1897 (1998).
22. Spielman, R. S. & Ewens, W. J. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450–458 (1998).
23. Knapp, M. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am. J. Hum. Genet.* **64**, 861–870 (1999).
24. Horvath, S. *et al.* Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet. Epidemiol.* **26**, 61–69 (2004).
25. Dudbridge, F. Pedigree disequilibrium tests for multilocus haplotypes. *Genet. Epidemiol.* **25**, 115–121 (2005).
26. Cordell, H. J., Barratt, B. J. & Clayton, D. G. Case/pseudocontrol 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.* **26**, 167–185 (2004).
27. Purcell, S., Sham, P. & Daly, M. J. Parental phenotypes in family-based association analysis. *Am. J. Hum. Genet.* **76**, 249–259 (2005).
28. Whittaker, J. C. & Lewis, C. M. Power comparisons of the transmission/disequilibrium test and sib-transmission/disequilibrium-test statistics. *Am. J. Hum. Genet.* **65**, 578–580 (1999).
29. Lange, C. & Laird, N. Analytical sample size and power calculations for a general class of family-based association tests: dichotomous traits. *Am. J. Hum. Genet.* **71**, 575–584 (2002).
30. Whittaker, J. C. & Lewis, C. M. The effect of family structure on linkage tests using allelic association. *Am. J. Hum. Genet.* **63**, 889–897 (1998).
31. Lange, C. & Laird, N. M. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genet. Epidemiol.* **23**, 165–180 (2002).
32. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
33. Gauderman, W. J. Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genet. Epidemiol.* **25**, 327–338 (2003).
34. Lunetta, K. L. *et al.* Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am. J. Hum. Genet.* **66**, 605–614 (2000).
35. Lange, C. *et al.* A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat. Appl. Genet. Mol. Biol.* **3**, 17 (2004).
36. Lange, C., DeMeo, D. L. & Laird, N. M. Power and design considerations for a general class of family-based association tests: quantitative traits. *Am. J. Hum. Genet.* **71**, 1330–1341 (2002).
37. Weiss, S. T. The origins of childhood asthma. *Monaldi Arch. Chest Dis.* **49**, 154–158 (1994).
38. Weiss, S. T. Epidemiology and heterogeneity of asthma. *Ann. Allergy Asthma Immunol.* **87** (1 Suppl. 1), 5–8 (2001).
39. Silverman, E. K. *et al.* Familial aggregation of severe, early-onset COPD: candidate gene approaches. *Chest* **117** (5 Suppl. 1), 273S–274S (2000).
40. Demeo, D. L. *et al.* The *SERPINE2* gene is associated with chronic obstructive pulmonary disease. *Am. J. Hum. Genet.* **78**, 253–264 (2005).
41. Celedon, J. C. *et al.* The transforming growth factor- $\beta$ 1 (TGFB1) gene is associated with chronic obstructive pulmonary disease (COPD). *Hum. Mol. Genet.* **13**, 1649–1656 (2004).
42. Todd, R. Genetics of attention deficit/hyperactivity disorder: are we ready for molecular genetic studies? *Am. J. Med. Genet.* **96**, 241–243 (2000).
43. Lange, C. *et al.* A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* **4**, 195–206 (2003).
44. Mokliatchouk, O., Blacker, O. & Rabinowitz, D. Association tests for traits with variable age at onset. *Hum. Hered.* **51**, 46–53 (2001).
45. Lange, C., Blacker, D. & Laird, N. M. Family-based association tests for survival and times-to-onset analysis. *Stat. Med.* **23**, 179–189 (2004).
46. Jiang, H. *et al.* Family-based association test for time-to-onset data with time-dependent differences between the hazard functions. *Genet. Epidemiol.* **30**, 124–132 (2005).
47. Shih, M. C. & Whittemore, A. S. Tests for genetic association using family data. *Genet. Epidemiol.* **22**, 128–145 (2002).
48. Lange, C. *et al.* Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am. J. Hum. Genet.* **73**, 801–811 (2003).
49. Lange, C. *et al.* PBAT: tools for family-based association studies. *Am. J. Hum. Genet.* **74**, 367–369 (2004).
50. 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).  
**Demonstrates that the multi-testing problem can be handled at a genome-wide level in family-based association tests.**
51. Lasky-Su, J. *et al.* Family-based association analysis of a statistically derived quantitative trait for ADDO reveals an association in DRD4 with inattentive simony in AD individuals. *Am. J. Med. Genet. B Neurophysiatr. Genet.* **138B**, 57–58 (2005).
52. Thomas, D., Xie, R. & Gebregziabher, M. Two-stage sampling designs for gene association studies. *Genet. Epidemiol.* **27**, 401–414 (2004).
53. Zaykin, D. V. & Zhivotovskiy, L. A. Ranks of genuine associations in whole-genome scans. *Genetics* **171**, 813–823 (2005).
54. Rosner, B. *Fundamentals of Biostatistics* 5th edn 527–530 (Duxbury, Boston MA, 1995).
55. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802 (1988).
56. Herbert, A. *et al.* A common genetic variant 10 kb upstream of *INSIG2* is associated with adult and childhood obesity. *Science* (in the press).
57. Gordon, D. *et al.* A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur. J. Hum. Genet.* **12**, 752–761 (2004).
58. Gordon, D. *et al.* A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am. J. Hum. Genet.* **69**, 371–380 (2001).
59. Gordon, D. & Ott, J. Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac. Symp. Biocomput.* 18–29 (2001).
60. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).  
**Proposed the original idea of the TDT.**
61. Laird, N. M., Horvath, S. & Xu, X. Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* **19** (Suppl. 1), S36–S42 (2000).

62. Self, S. *et al.* On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* **47**, 53–61 (1991).
63. Cordell, H. J. & Clayton, D. G. 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.* **70**, 124–141 (2002).
64. Schaid, D. J. General score tests for associations of genetic markers with disease using cases and their parents. *Genet. Epidemiol.* **13**, 423–449 (1996). **Shows how the TDT can be derived as a score statistic from a multinomial likelihood model.**
65. Clayton, D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am. J. Hum. Genet.* **65**, 1170–1177 (1999). **The first haplotype-analysis paper to use a likelihood approach.**
66. Whittemore, A. S. & Tu, I. P. Detection of disease genes by use of family data. I. Likelihood-based theory. *Am. J. Hum. Genet.* **66**, 1328–1340 (2000). **Generalized Schaid's-likelihood approach to handle missing parents, multiple offspring and incorporate founders into the test statistic.**
67. Horvath, S., Xu, X. & Laird, N. M. The family based association test method: strategies for studying general genotype–phenotype associations. *Euro. J. Hum. Gen.* **9**, 301–306 (2001).
68. Weinberg, C. R., Wilcox, A. J. & Lie, R. T. 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.* **62**, 969–978 (1998).
69. Weinberg, C. R. Allowing for missing parents in genetic studies of case-parent triads. *Am. J. Hum. Genet.* **64**, 1186–1193 (1999).
70. Weinberg, C. R. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am. J. Hum. Genet.* **65**, 229–235 (1999).
71. Umbach, D. M. & Weinberg, C. R. The use of case-parent triads to study joint effects of genotype and exposure. *Am. J. Hum. Genet.* **66**, 251–261 (2000).
72. Kistner, E. O. & Weinberg, C. R. Method for using complete and incomplete trios to identify genes related to a quantitative trait. *Genet. Epidemiol.* **27**, 33–42 (2004).
73. Kistner, E. O. & Weinberg, C. R. A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents. *Genet. Epidemiol.* **29**, 155–165 (2005).
74. Kistner, E. O., Infante-Rivard, C. & Weinberg, C. R. A method for using incomplete triads to test maternally mediated genetic effects and parent-of-origin effects in relation to a quantitative trait. *Am. J. Epidemiol.* **163**, 255–261 (2006).
76. Witte, J. S., Gauderman, W. J. & Thomas, D. C. Asymptotic bias and efficiency in case–control studies of candidate genes and gene–environment interactions: basic family designs. *Am. J. Epidemiol.* **149**, 693–705 (1999).

## Acknowledgements

This work was supported by the National Institute of Mental Health and the National Heart, Lung and Blood Institute, USA. We would like to thank C. Garcia for with help with the preparation of this manuscript.

## Competing interests statement

The authors declare no competing financial interests.

## FURTHER INFORMATION

**FBAT homepage:** <http://www.biostat.harvard.edu/~fbat/default.html>  
**Framingham Heart Study:** <http://www.framingham.com/heart/>  
**Human Genome Project:** [http://www.ornl.gov/sci/techresources/Human\\_Genome/home.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)  
**Nan Laird's homepage:** <http://www.hsph.harvard.edu/facres/lrd.html>  
**PBAT homepage:** <http://www.biostat.harvard.edu/~clange/default.html>  
**Access to this links box is available online.**

**Online summary**

- Either population-based or family-based designs can be used in gene-association studies. Population-based designs use unrelated individuals; family-based designs use probands and their relatives, typically either parents or siblings.
- Genetic-association studies face the obstacles of population substructures and multiple testing.
- Family-based designs are favoured because they are robust against confounding due to population substructures and test both linkage and association.
- Case-control designs are preferred for the relative ease of data collection. They have modest power advantages, depending on the prevalence of the disease.
- Family-based designs can be extended to incorporate pedigrees and complex phenotypes.
- Screening tools are available for family-based designs that allow the multiple-testing problem, which is an important issue in whole-genome association studies, to be handled.

**Biographies**

Nan Laird is Professor of Biostatistics at the Harvard School of Public Health, Boston, Massachusetts, USA. Her main interests include family studies in genetics, missing data, longitudinal studies and genetic studies in psychiatric disorders, asthma and chronic obstructive pulmonary disease. She is a fellow of the American Statistical Association and the International Statistical Institute, and the recipient of the FN David and the Janet Norwood awards.

Christoph Lange is an Assistant Professor of Biostatistics at the Harvard School of Public Health, Boston, Massachusetts, USA, and an Assistant Professor of Medicine at Harvard Medical School. His main interests include family studies in genetics, causal inference and generalized linear models in asthma, chronic obstructive pulmonary disease, psychiatric disorders and obesity. He received a Ph.D. in Statistics from the University of Reading, UK. He also holds Master's Degrees in Mathematics and Computer Science from the University of Augsburg, Germany, and in Biostatistics from the University of Hasselt, Belgium.

**ToC Blurb**

Although they are sometimes overlooked, family-based designs provide important advantages for detecting genetic associations in studies of complex disease. In particular, they provide a means of overcoming the problems that arise when multiple hypotheses are tested in genome-wide association studies.

**Online links**

**FBAT homepage:** <http://www.biostat.harvard.edu/~fbat/default.html>

**Framingham Heart Study:** <http://www.framingham.com/heart/>

**Human Genome Project:** [http://www.ornl.gov/sci/techresources/Human\\_Genome/home.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)

**Nan Laird's homepage:** <http://www.hsph.harvard.edu/facres/lrd.html>

**PBAT homepage:** <http://www.biostat.harvard.edu/~clange/default.html>