

Family-based association studies

Hongyu Zhao Yale University School of Medicine, New Haven, Connecticut, USA

Over the past decade, attention has turned from positional cloning of Mendelian disease genes to the dissection of complex diseases. Both theoretical and empirical studies have shown that traditional linkage studies may be inferior in power compared to studies that directly utilize allele status. Case-control association studies, as an alternative, are subject to bias due to population stratification. As a compromise between linkage studies and case-control studies, family-based association designs have received great attention recently due to their potentially higher power to identify complex disease genes and their robustness in the presence of population substructure. In this review, we first describe the basic family-based association design involving one affected offspring with its two parents, all genotyped for a biallelic genetic marker. Extensions of the original transmission disequilibrium tests to multiallelic markers, families with multiple siblings, families with incomplete parental genotypes, and general pedigree structures are discussed. Further developments of statistical methods to study quantitative traits, to analyse genes on the X chromosome, to incorporate multiple tightly linked markers, to identify imprinting genes, and to detect gene-environment interactions are also reviewed. Finally, we discuss the implications of the completion of the Human Genome Project and the identification of hundreds of thousands of genetic polymorphisms on employing family-based association designs to search for complex disease genes.

1 Introduction

The Human Genome Project has generated a large volume of genetic markers that can be used to map genes of complex traits. In the most recent genetic maps constructed from the CEPH pedigrees,¹ Broman *et al.* mapped more than 8000 short tandem repeat polymorphisms.² In recent years it has been realized that single nucleotide polymorphisms (SNPs) have great potential in mapping complex disease genes.³ In the most recent release by the SNP consortium on 21 August 2000 at their website <http://snp.cshl.org>, there are mapped 296 990 SNPs. With the influx of these genetic markers at hand, one important issue is how to fully utilize such abundant information to most efficiently identify disease genes.

Genetic linkage studies are usually accomplished by collecting pedigrees with affected individuals and excess allele sharing among affected individuals is sought for some markers. Although this approach has been used successfully to map simple Mendelian diseases, it has not yielded consistent evidence for mapping complex disease genes. Over the last decade, attention has turned from positional cloning of Mendelian disease genes to the dissection of complex diseases. Theoretical studies³ have shown that linkage methods that utilize allele-sharing among affected relatives, e.g. the affected sib-pair method, may be inferior in power compared to studies that directly utilize allele status if dense polymorphic markers are available. These

Address for correspondence: Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA. E-mail: hongyu.zhao@yale.edu

theoretical studies have been confirmed in practice. For example, evidence from conventional linkage studies for the involvement of the insulin gene region in insulin-dependent diabetes mellitus lagged behind that from association studies.⁴

One major limitation of the case-control design is that control individuals may not be well matched to cases. A positive association can occur for three different reasons: (1) the allele itself is a cause of the disease; (2) the allele is in linkage disequilibrium with a susceptible allele at the disease gene; and (3) population subdivision and admixture may lead to disease association even in the absence of linkage. One well-known example is the immunoglobulin gene Gm for non-insulin-dependent-diabetes mellitus.⁵ Among residents of the Gila River Indian Community in Arizona, diabetes was associated with the haplotype Gm. However, this association no longer exists among ethnically homogeneous subjects. The confounding by population stratification occurred because the Gm haplotype serves as a marker for Caucasian heritage, and the risk of diabetes varies with the level of this ancestry. Population genetics studies have shown that allele frequency at some loci can vary considerably among populations contributing to the US white population.⁶

Family-based association designs offer a compromise between traditional linkage studies and case-control association studies. All methods proposed in the literature have the common feature of comparing alleles transmitted from the parents to alleles not transmitted to the affected offspring or alleles transmitted to the unaffected offspring. Recent years have seen rapid developments in statistical methods using the family-based association designs to infer allelic association and linkage with a particular allele. These developments were partly motivated by the observations made by Risch and Merikangas³ that linkage analysis is likely to succeed only for loci with relative risks in the range of four or larger but not for loci with relative risks two or less. Assuming a multiplicative model, they showed that even if one needs to test 10^6 polymorphic markers and allows for a conservative significance level of 5×10^{-8} , genes with effects as low as 1.5 could be readily detected in realistic sized samples of families using the TDT.

In this review, we first describe the basic family-based association design with one affected offspring and its two parents, all genotyped at one biallelic genetic marker. In the context of mapping genes for qualitative traits, we cover extensions of the basic transmission disequilibrium tests to multiallelic markers, families with multiple siblings, families without parents, families with only one parent, and general pedigrees. We then summarize parallel developments of statistical methods to map quantitative trait loci. Methods to detect genes on the X chromosome, to use multiple tightly linked markers, to identify genes with parent-of-origin effects, and to detect gene-environment interactions are also reviewed. We conclude this review by comparing the efficiency of different study designs and pointing to directions for future research. Because family-based association study is such an active area of research, although we have made an effort to cover all relevant topics, there is no doubt that some contributions may be missed in this review due to our inability to collect all relevant literature.

2 Qualitative traits

2.1 The simplest design: one affected child with two parents, all genotyped at one biallelic marker

Consider a biallelic marker with two alleles M and m. Genotypes for a biallelic marker are collected on the parents and their affected offspring. Genotypes from each family trio allow us to determine which of the maternal and paternal alleles are transmitted to the affected offspring and which alleles are not transmitted. Take the two marker alleles transmitted to an affected offspring to form a *case* genotype and the non-transmitted alleles to form a *control* genotype. Rubinstein *et al.*⁷ and Falk and Rubinstein⁸ proposed to examine whether allele M is present for each case genotype and its corresponding control genotype. The data can be summarized in Table 1. We adopt the notation in the literature⁹⁻¹¹ to denote the entries in this section and this part of exposition follows closely the discussion by Schaid and Sommer.¹⁰

The data in Table 1 can be analysed using a matched analysis⁹ through the matched genotype relative risk (MGRR) statistic:

$$MGRR = \frac{(B - C)^2}{B + C}$$

When the cases and controls are analysed as an unmatched design, the data can be summarized in a different form (Table 2). Note that the entries in Table 1 and Table 2 with the same notation correspond to the same observed statistic. Falk and Rubinstein⁸ proposed the genotype-based haplotype relative risk (GHRR) statistic to analyse the data in Table 2:

$$\begin{aligned} GHRR &= \frac{2n(W - Y)^2}{(W + Y)(X + Z)} \\ &= \frac{(B - C)^2}{(2A + B + C)(B + C + 2D)/2n} \end{aligned}$$

Table 1 Table for the matched analysis of transmitted and nontransmitted genotypes

Case	Control		Total
	M present	M absent	
M present	A	B	W=A+B
M absent	C	D	X=C+D
Total	Y=A+C	Z=B+D	N

Table 2 Table for the unmatched analysis of transmitted and nontransmitted genotypes

	M present	M absent	Total
Case	W	X	n
Control	Y	Z	n
Total	W+Y	X+Z	2n

The difference between the matched analysis MGRR and the unmatched analysis GHRR is the estimate of the variance of B-C. Define the relative risk (RR) as $P(\text{affected} \mid \text{presence of allele M})/P(\text{affected} \mid \text{absence of allele M})$. The value of RR can be estimated by haplotype relative risk (HRR) $WZ/(XY)$. Knapp *et al.*¹² showed that the true HRR (ignoring random sampling variation) is smaller than the true relative risk when cases and unrelated controls are used, i.e. $|HRR-1| \leq |RR-1|$. If the recombination fraction between the marker and the disease gene is 0, then $HRR = RR$ for any mode of inheritance.

Terwilliger and Ott⁹ suggested that one examine each individual allele, rather than individual genotype in the analysis. In this case, the transmission/non-transmission patterns can be summarized using Table 3.

The appropriate matched analysis for the data in Table 3 is the McNemar test, named (in this context) the transmission/disequilibrium test (TDT) by Spielman *et al.*⁴ (see also Terwilliger and Ott⁹):

$$TDT = \frac{(b - c)^2}{b + c}$$

The original intended use of the TDT was to test for linkage in cases where disease association already had been found.

We can also apply the unmatched analysis to examine the association between the transmission of alleles and the particular allele form. In this case, the data can be summarized as Table 4, and the haplotype-based haplotype relative risk (HHRR) statistic:

$$HHRR = \frac{4n(w - y)^2}{(z + y)(x + z)} = \frac{(b - c)^2}{(2a + b + c)(b + c + 2d)/4n}$$

discussed by Terwilliger and Ott⁹ can be used for statistical tests. The entries in Tables 3 and 4 with the same notation correspond to same statistic. Similar to the difference between the MGRR and the GHRR in genotype-based analysis, the difference between the TDT and the HHRR is how the variance of $b-c$ is estimated.

Table 3 Table for the matched analysis of transmitted and nontransmitted alleles

Transmitted allele	Nontransmitted allele		Total
	M	m	
M	a	b	w=a+b
m	c	d	x=c+d
Total	y=a+c	z=b+d	2n

Table 4 Table for the unmatched analysis of transmitted and nontransmitted alleles

	M	m	Total
Transmitted allele	w	x	2n
Nontransmitted allele	y	z	2n
Total	w+y	x+z	4n

Other test statistics that have been proposed to analyse family trios were discussed in detail by Schaid and Sommer.¹⁰ Note that one implicit assumption underlying these test statistics is that there is no segregation distortion at the marker locus. This assumption can be tested by comparing the transmission of a particular marker allele to affected and unaffected offspring.⁴

In an invited editorial, Spielman and Ewens¹³ discussed the validity of the TDT and the HHRR (also called the AFBAC by Thomson¹⁴) to analyse the observed family genotypes for the null hypothesis of association or the null hypothesis of no linkage if it assumed under these hypotheses that the test statistic has a chi-square distribution with one degree of freedom. The term *valid* has been used in the literature in the sense that the statistical test has the correct nominal significance level under the null hypothesis. The TDT is a valid test of linkage in structured populations, irrespective of the pedigree structures. However, the presence of multiplex sibships makes the TDT invalid as a test of association. The contingency statistic, GHRR or HHRR, is not valid, in general, as a test for association since it requires random mating in the population and no admixture for at least two generations before the sample of affected offspring is taken. Even when the contingency statistic is valid as a test of association, it is not valid as a test of linkage.

2.2 Multiple alleles and multiple sibs with two parents' genotypes available

So far we have assumed that the marker being studied has only two alleles. Markers with multiple alleles are commonly used in association and linkage studies, and many approaches have been proposed to analyse multiallelic markers in family-based association studies. Consider a disease gene with two alleles D and d, with allele frequency p_1 and p_2 , respectively. Denote the recombination fraction between them by θ . Consider a marker with k alleles, M_1, \dots, M_k . As for biallelic markers, we may construct a $k \times k$ transmission/non-transmission table, where t_{ij} in the table represents the number of parents who have genotype $M_i M_j$ and transmit M_i to the affected offspring. If the population frequency of marker allele M_i is q_i , we can define k association parameters, δ_i , as follows:

$$\delta_i = P(M_i D) - p_1 q_i, \quad i = 1, \dots, k$$

Sethuraman¹⁵ showed that the probability for each cell in the transmission/non-transmission table is

$$p_{ij} = (p_1 \pi_{11} + p_2 \pi_{12}) [m_j (m_i p_1 + \delta_i) - \theta (m_j \delta_i - m_i \delta_j)] / K \\ + (p_1 \pi_{12} + p_2 \pi_{22}) [m_j (m_i p_2 + \delta_i) + \theta (m_j \delta_i - m_i \delta_j)] / K$$

where π_{11} , π_{12} , and π_{22} are the disease penetrance for individuals carrying genotypes DD , Dd , and dd , respectively, and $K = p_1^2 \pi_{11} + 2p_1 p_2 \pi_{12} + p_2^2 \pi_{22}$ is the population prevalence of the disease. Similar expressions were used by Sham and Curtis in their logistic model,¹⁶ and by Morris *et al.*¹⁷ in their likelihood ratio tests. When $\theta = 0$

$$\frac{p_{ij}}{p_{ji}} = \frac{(p_1 \pi_{11} + p_2 \pi_{12}) [m_j (m_i p_1 + \delta_i)] + (p_1 \pi_{12} + p_2 \pi_{22}) [m_j (m_i p_2 + \delta_i)]}{(p_1 \pi_{11} + p_2 \pi_{12}) [m_i (m_j p_1 + \delta_j)] + (p_1 \pi_{12} + p_2 \pi_{22}) [m_i (m_j p_2 + \delta_i)]}$$

As shown by Zhao,¹⁸ this ratio is in fact the ratio of disease risks carried by the two alleles, M_i and M_j , respectively. Let $p_{i+} = \sum_{j=1}^k p_{ij}$ and $p_{+i} = \sum_{j=1}^k p_{ji}$. From these cell probabilities, we can obtain

$$\begin{aligned} p_{i+} &= [(p_1\pi_{11} + p_2\pi_{12})\{m_i p_1 + (1 - \theta)\delta_i\} \\ &\quad + (p_1\pi_{12} + p_2\pi_{22})\{m_i p_2 - (1 - \theta)\delta_i\}]/K \\ p_{+i} &= [(p_1\pi_{11} + p_2\pi_{12})\{m_i p_1 + \theta\delta_i\} \\ &\quad + (p_1\pi_{12} + p_2\pi_{22})\{m_i p_2 - \theta\delta_i\}]/K \end{aligned}$$

Therefore, $p_{i+} = p_{+i}$ if and only if $(1 - 2\theta)\delta = 0$, and the null hypothesis of no linkage can be tested by examining marginal homogeneity. Such a test has power only when $\delta_i \neq 0$ for at least one i .

Let $t_{i+} = \sum_{j=1}^k t_{ij}$, $t_{+i} = \sum_{j=1}^k t_{ji}$, $d_i = t_{i+} - t_{+i}$, $d' = (d_1, \dots, d_{k-1})$, $\hat{\Sigma}_{ij} = -(t_{ij} + t_{ji})$, when $i \neq j$, and $\hat{\Sigma}_{ii} = t_{i+} + t_{+i} - 2t_{ii}$. Bickeböllner and Clerget-Darpoux¹⁹ proposed the test statistic

$$W = d' \hat{\Sigma}^{-1} d$$

which has an asymptotic chi-square distribution with $k - 1$ degrees of freedom. Alternatively, transmission disequilibrium can be tested by examining whether the transmission/non-transmission table is symmetric using the following statistic

$$TDT_c = \sum_{i < j} \frac{(t_{ij} - t_{ji})^2}{t_{ij} + t_{ji}}$$

This test statistic has an asymptotic distribution with $k(k - 1)/2$ degrees of freedom.²⁰

An alternative test that tests marginal homogeneity was proposed by Spielman and Ewens¹³

$$TDT_{SE} = \frac{k - 1}{k} \sum_{i=1}^k \frac{(t_{i+} - t_{+i})^2}{t_{i+} + t_{+i} - 2t_{ii}}$$

As noted by Sham,²¹ Schaid,²² and Lazeroni and Lange,²³ the TDT_{SE} statistic may not have a chi-square distribution with $k - 1$ degrees of freedom, and tends to be anticonservative. Although this statistic is easier to compute than the W statistic, it does not account for the covariance among the $t_{i+} - t_{+i}$, $i = 1, \dots, k - 1$. In addition to the tests based on the contingency table, model-based methods have been developed to analyse multiple markers. Sham and Curtis¹⁶ proposed a logistic regression model for the probability that a particular marker allele is transmitted by a heterozygous parent. This logistic model has the following form

$$\log \frac{p_{ij}}{p_{ji}} = b_i - b_j$$

To avoid aliasing, we can set b_k to be zero. In the generalized linear model notation, this is equivalent to coding each genotype by a vector where the i th component is 1, the

j th component is -1 , and all other components are 0. This model is equivalent to the quasi-symmetry model developed by Bradley and Terry.²⁴ Using simulated data, Miller²⁵ found that the distribution of the p -values is not uniform as expected and suggested that one use Monte Carlo methods to evaluate significance levels. Models of the same form were studied by Jin *et al.*²⁶ to detect segregation distortion using randomly ascertained families. Logistic models and their extensions have also been discussed by Harley *et al.*,²⁷ Rice *et al.*²⁸ and Waldman *et al.*²⁹ In the same spirit, Sinsheimer *et al.*³⁰ developed the gamete-competition model that is applicable to general pedigrees. For family trios, their method is identical to the Sham and Curtis model. However, when there are missing genotypes, population allele frequencies are used to estimate missing genotypes, thus this procedure may be biased if there is population stratification.

Another class of model-based methods was proposed by Schaid²² using the conditional likelihoods developed by Self *et al.*³¹ Offspring genotype is modelled as a function of parental genotypes and offspring disease status as follows:

$$P(g_c | g_m, g_f, D) = \frac{P(D | g_c, g_m, g_f) P(g_c | g_m, g_f) P(g_m, g_f)}{\sum_{g^* \in G} P(D | g^*, g_m, g_f) P(g^* | g_m, g_f) P(g_m, g_f)}$$

In the above expression, D represents the event that the offspring is affected, g_c is the marker genotype of the affected offspring, g_m and g_f are the marker genotypes of the parents, and g^* is one of the four possible genotypes of the offspring conditional on parental genotypes. Assuming that $P(D | g_c, g_m, g_f) = P(D | g_c)$, the above equation reduces to

$$P(g_c | g_m, g_f, D) = \frac{r(g_c)}{\sum_{g^* \in G} r(g^*)}$$

where $r(g)$ is the relative risk of disease for genotype g , which consists of a pair of haplotypes, (i, j) . If π_g is the penetrance for genotype g , the genotype relative risk is $\pi_g = \pi_0 r(g)$ for a reference genotype. For H distinct haplotypes, $H(H+1)/2$ distinct parameters have to be defined for general disease models. The model can be simplified if we specify a certain disease model. For example, under a multiplicative model

$$\log r(g) = \log r(i, j) = \beta_i + \beta_j$$

A more general model was proposed by Clayton and Jones³² with the following form

$$h[r(i, j)] = \beta_i + \beta_j = \frac{1}{2} \{h[r(i, i)] + h[r(j, j)]\}$$

where h is an unspecified monotone increasing function. Schaid²² proposed to numerically code the marker relative risks as follows:

$$\log [r(g)] = X' \beta$$

where X is the coded vector for the observed genotype g . The null hypothesis of no association, i.e. $\beta = 0$, can be tested using the Rao efficient score statistic in the form of

$$S = U'V^{-1}U$$

where

$$U = \frac{\partial \ln L}{\partial \beta} \Big|_{\beta=0} \quad \text{and} \quad V_{ij} = -E \left\{ \frac{\partial^2 \ln L}{\partial \beta_i \partial \beta_j} \Big|_{\beta=0} \right\}$$

When the mode of inheritance is unknown, two statistics, the W statistic discussed above and the maxTDT statistic defined as

$$\text{maxTDT} = \max_i \left\{ \frac{(t_{i+} - t_{+i})^2}{t_{i+} + t_{+i} - 2t_{ii}}, \quad i = 1, \dots, k \right\}$$

were found to be generally powerful against alternatives specified by the relative risks of marker genotypes.

Using the framework put forth by Self *et al.*³¹ and Schaid,²² Lunetta *et al.*³³ considered general association models for an arbitrary phenotype Y and score statistics based on likelihoods for the distribution of Y . Their association model may include environmental factors as well as multiple genes. A major distinction between their model and previous models is that Lunetta *et al.* modelled phenotypes conditional on genotypes and adjusted for population stratification at the final step by using the appropriate permutation distributions for the offspring allele values. In our following discussion, we use n to denote the number of families, and n_i to denote the number of siblings in the i th family. For a biallelic marker with alleles M and m , when both affected and unaffected offspring are used, the score statistic proposed by Lunetta *et al.* is

$$S = \sum_{i=1}^n \sum_{j=1}^{n_i} X_{ij}(Y_{ij} - \mu) = (1 - \mu)S_A - \mu S_U$$

where S_A is the total number of M alleles transmitted to the affected offspring, S_U is the total number of M alleles transmitted to the unaffected offspring, X_{ij} is the coding for the j th sibling's genotype in the i th pedigree, and Y_{ij} is the phenotype of the j th sibling in the i th pedigree coded either 0 or 1. When μ is set to 0, $S = S_A$ and only transmissions to the affected offspring are counted. When μ is set to the population prevalence of the disease, it is identical to the T_{sib} proposed by Whittaker and Lewis,²⁰ the most powerful test under a multiplicative-genotype relative-risk disease model. When the disease is rare, $\mu \approx 0$, most information is contained in the genotypes of affected individuals. On the other hand, including the unaffected offspring may increase the power when the disease is common.

Whittemore and Tu³⁴ developed a class of likelihood-based score statistics that, in principle, can handle arbitrary family structures with arbitrary patterns of missing data. The score statistic has two components: the non-founder statistic that evaluates disequilibrium in the transmission of marker alleles from parents to offspring, and the founder statistic that compares the observed or inferred marker genotypes in the

family founders with those of controls or those of some reference population. In their model, each individual's phenotype value is defined as 1 if this person is affected, $-\psi$ if this person is unaffected, and 0 if this individual's phenotype is unknown. Each individual's genotype is coded so that $c_0 = 0$ for a homozygous mm individual, $c_2 = 1$ for a homozygous MM individual, and c_1 for a heterozygous Mm individual. When $\psi = 0$ and $c_1 = \frac{1}{2}$, their non-founder statistic is the same as the TDT. When $\psi = 0$ and $c_1 = 0$ or $c_1 = 1$, their non-founder statistic is the score statistic proposed by Schaid.²² Although they illustrated how to analyse data with incomplete parental genotypes, random mating was assumed in their formulation and this assumption may lead to bias if there is population stratification.

When there are multiple sibs in a nuclear family, the TDT is still a valid test for linkage in the presence of association. However, the TDT is no longer a valid test for association because the transmissions from a parent to its affected children are correlated if there is linkage, even if there is no association. One strategy is to randomly select one affected child from each family. Martin *et al.*³⁵ proposed alternative tests that may use data from affected sib pairs. Let s_j be the number of heterozygous parents with genotype Mm who transmit allele M to j children, their test statistic is

$$T_{sp} = \frac{(s_0 - s_2)^2}{s_0 + s_2}$$

which has an approximate chi-square distribution with one degree of freedom. Using their notation, the TDT is

$$\text{TDT} = \frac{(s_0 - s_2)^2}{\frac{1}{2}(s_0 + s_1 + s_2)}$$

The two statistics, the TDT and the T_{sp} , are related as follows:

$$\text{TDT} = T_{sp} \frac{s_0 + s_2}{\frac{1}{2}(s_0 + s_1 + s_2)}$$

and the TDT is more powerful than T_{sp} to test for linkage. Noting that $s_1 \approx s_0 + s_2$ under the null hypothesis of no linkage, Wicks³⁶ defined a family of TDT-like statistics for affected sib pairs in the form of

$$\text{TDT}(\alpha) = \frac{(s_0 - s_2)^2}{(1 - \alpha)(s_0 + s_2) + \alpha s_1}$$

Under the null hypothesis of no linkage, $\text{TDT}(\alpha)$ has a chi-square distribution with one degree of freedom. She suggested that $\text{TDT}(\alpha = 1)$ is the most powerful test in this class

$$\text{TDT}(\alpha = 1) = \frac{(s_0 - s_2)^2}{s_1}$$

The TDT permits exact calculation of p -values.³⁷ Alternatively, Morris *et al.*³⁸ des-

cribed how to use the randomization procedure to adjust for multiple comparisons and to sequentially investigate patterns of association of several individual alleles with the disease gene. The use of exact or Monte Carlo methods to calculate significance levels was also advocated by Whittaker and Thompson.³⁹

The power and sample size issues in the TDT have been investigated by many researchers. To estimate the power of the TDT, the most satisfactory approximation formulae appear to be those derived by Knapp.⁴⁰ When the TDT was compared with likelihood-based methods, Schaid⁴¹ found that for rare alleles, the TDT method is inefficient for recessive patterns of relative risks. For alleles that are not rare, assuming a multiplicative model may lead to gross underestimate of the required sample size. For common alleles, the TDT method can be very inefficient if the true genotype risks have a dominant pattern. Overall, the likelihood ratio statistic with two degrees of freedom seems to perform well across a wide range of disease models.

Simulation analyses have been performed to compare the power of a variety of tests for markers with multiple alleles. Sethuraman¹⁵ compared TDT_{SE} , W , $maxTDT$, and TDT_C . All statistics are more powerful when the mode of inheritance is recessive than when the mode of inheritance is dominant, and when the disease allele is rare than when the disease allele is common. The power of the TDT_C is much less than the other three test statistics, possibly due to the large number of degrees of freedom. Among the other three tests, the TDT_{SE} is, in general, more powerful than the W statistic. When only one marker allele is highly associated with the disease, the $maxTDT$ is very powerful, but even in such cases, the TDT_{SE} is also very powerful. Therefore, unless there is evidence that a particular marker allele is associated with the disease or the number of alleles is very large, Sethuraman suggested that there would be no real advantage in using the $maxTDT$ rather than the TDT_{SE} . The TDT_{SE} was also recommended by Kaplan *et al.*⁴²

2.3 Families without parental genotypes

Unobservable parental genotypes present problems for the TDT. Although parents may be unobservable, information about their genotypes may be contained in the genotypes of the proband and his or her siblings. Methods have been developed under the assumption of Hardy–Weinberg equilibrium,^{43–45} however, these methods may lead to biased results in the presence of population stratification.

Curtis⁴⁶ proposed to randomly select one affected offspring and then select one unaffected offspring whose marker genotype is maximally different from that of the affected offspring. This approach is unbiased although the procedure selects the most different unaffected sibling. To calculate the test statistic, each marker allele in the affected individual is compared with each marker allele in the unaffected sibling. If the alleles are identical, the comparison is ignored. If the alleles are different, then $\frac{1}{2}$ is added to T_{ij} , where i is the marker allele in the affected sibling and j is the marker allele in the unaffected sibling. For a biallelic marker, the test statistic is

$$Z_C = \frac{T_{12} - \left(\frac{N_1}{2} + N_2\right)}{\sqrt{\frac{N_1}{4} + N_2}}$$

where N_i is the number of sibships that increase the test statistic of either T_{12} or T_{21} by i .

This procedure is a valid test for both linkage and association. For multiple alleles, Curtis adopted a likelihood model similar to that of Sham and Curtis.¹⁶ However, the chi-square approximation to the likelihood ratio test distribution for this model may be poor.⁴⁷

For a marker with k alleles, Boehnke and Langefeld⁴⁸ represented the marker allele data in a $2 \times k$ contingency table in which the rows represent the disease status and the columns represent marker alleles. There are many ways to calculate entries in this contingency table; here we discuss one counting scheme that they found to have the best power overall. For this scheme, if the marker alleles in the two sibs are all different, then all four markers are counted in the table. If an affected individual and the unaffected individual have one marker allele in common, then these alleles are ignored and only the two different alleles are counted in the table. One test statistic derived from this table is

$$AC_2 = \sum_{j=1}^k \frac{(n_{1j} - n_{2j})^2}{n_{1j} + n_{2j}}$$

where n_{1j} is the count of marker allele j in the N affected sibs, and n_{2j} is the count of marker allele j in the N unaffected sibs. The permutation procedure that randomly permuted the affection statuses of the sibs was proposed to evaluate the significance level.

The sib TDT (S-TDT) developed by Spielman and Ewens⁴⁹ can analyse sibships with one or more affected sibs and one or more unaffected sibs. This test is valid as a test for linkage. To be included in the S-TDT analysis, the sibships have to meet two criteria: (1) there are at least one affected sib and one unaffected sib in each sibship in the data; and (2) the sibs must not have the same genotype. For each marker allele, Spielman and Ewens⁴⁹ defined the random variable Y_j as the number of allele M_j in the affected individuals from all sibships. The normalized statistic is

$$Z_j = \frac{Y_j - A_j}{\sqrt{V_j}}$$

and the calculations of the mean A_j and variance V_j of Y_j were discussed by Spielman and Ewens.⁴⁹ Either a permutation test or normal approximations can be used to assess the significance level. For a biallelic marker, the test statistic is Z_1 , whereas for a marker with k alleles, the test statistic they proposed is $Z_{\max} = \max|Z_j|$. Schaid and Rowland⁵⁰ noted that the S-TDT is equivalent to the conditional likelihood having log-additive effects of the marker alleles. The similarities and differences between the S-TDT and the Mantel–Haenszel test were discussed by Laird *et al.*⁵¹ and Ewens and Spielman.⁵²

For some families without parental genotype information, it may be possible to reconstruct parental genotypes from the genotypes of their offspring. In the context of the TDT, Spielman and Ewens^{13,53} suggested that one treat these reconstructed families as if parental genotypes have been typed. However, Curtis,⁵⁴ Spielman and Ewens,⁵³ and Knapp⁵⁵ noted that this procedure can introduce bias. To overcome the

potential bias in parental genotype reconstruction, Knapp⁵⁵ proposed a statistical procedure called RC-TDT (reconstruction combined TDT). He provided necessary and sufficient conditions for the observed marker genotypes in the offspring to allow for the reconstruction of the parental mating type. Similar conditions were given by Curtis,⁵⁴ but for a slightly different purpose. The mean and variance of the test statistic conditional on parental mating type being able to be reconstructed under the null hypothesis were derived. Power comparisons showed that the RC-TDT is more powerful than the S-TDT.^{55,56}

All methods described so far are based on comparing affected and unaffected sibs. Teng and Risch⁵⁷ suggested that there is additional information available in the sample from the relative frequency of the different sibship genotype constellations. Such information allows the estimate of the proportion of six mating type frequencies for a biallelic marker: $MM \times MM$, $MM \times Mm$, $MM \times mm$, $Mm \times Mm$, $Mm \times mm$, and $mm \times mm$. For the general case of r affected sibs and s unaffected sibs, they grouped the outcomes into six groups: (I) all sibs are MM ; (II) all sibs are mm ; (III) all sibs are Mm ; (IV) all sibs are either MM or Mm ; (V) all sibs are either Mm or mm ; (VI) the genotypes MM and mm appear among the sibs. However, there may be information loss in such groupings. For example, groups IV and V can be divided by the number of Mm individuals. Teng and Risch⁵⁷ derived sample size formula based on their test statistics. They found that two unaffected sibs without parents requires approximately 50% more families than when parents are available.

Using ideas similar to those in Teng and Risch,⁵⁷ Weinberg⁵⁸ proposed a likelihood-based approach to analysing families with incomplete parental information using the model discussed in their earlier work.⁵⁹ As is true for Teng and Risch's approach, the basic assumption underlying this type of analysis is that the probability of having missing information does not depend on the disease allele under study.

For extensions to families without parental genotype information, the methods discussed above are a valid test of association in the presence of linkage only if one affected sib and one unaffected sib are analysed from unrelated families. To include all available siblings from the same family, Horvath and Laird⁶⁰ developed a test of association for a candidate gene, called SDT (sibship disequilibrium test), when parental information is not available. Here we briefly describe their testing procedure. First consider a marker with two alleles M and m . For a set of siblings, let

$$m_A = (\text{Total number of allele } M \text{ among the affected})/n_A$$

$$m_U = (\text{Total number of allele } M \text{ among the unaffected})/n_U$$

where n_A and n_U are the number of affected and unaffected siblings in this family. Let $d = m_A - m_U$, b be the number of sibships for which $d > 0$, and c be the number of sibships for which $d < 0$. The SDT statistic is defined as:

$$SDT = \frac{(b - c)^2}{b + c}$$

For a marker with k alleles, we can define $d^{(j)}$ for the j th marker allele. Because

$$d^{(k)} = \sum_{j=1}^{k-1} d^{(j)}$$

we may only consider the first $k - 1$ alleles to form a vector $S' = (S^{(1)}, S^{(2)}, \dots, S^{(k-1)})$, where $S^{(j)} = \sum_{i=1}^n \text{sgn}(d_i^{(j)})$ and $d_i^{(j)}$ is the statistic $d^{(j)}$ from the i th sibship. The test statistic they proposed is

$$T = S'W^{-1}S$$

where $W_{jl} = \sum_{i=1}^n \text{sgn}(d_i^{(j)})\text{sgn}(d_i^{(l)})$. The SDT can be combined with the TDT when the data consist of a mixture of families with and without parental information, both for biallelic markers⁶⁰ and for multiallelic markers.⁶¹

In the case of discordant sib pairs, Horvath and Laird⁶⁰ also introduced a class of tests in the following form

$$T_x = \frac{(b_1 - c_1) + x(b_2 - c_2)}{\sqrt{(b_1 + c_1) + x^2(b_2 + c_2)}}$$

In the above expression, b_2 is the number of sib pairs (2,0), c_2 is the number of sib pairs (0,2), b_1 is the number of sib pairs (2,1) or (1,0), and c_1 is the number of sib pairs (1,2) or (0,1), where (i, j) denotes a sib pair with i copies of allele M in the affected sib and j copies of allele M in the unaffected sib. Within this class of tests, T_1 corresponds to the SDT, and T_2 corresponds to the test proposed by Curtis.⁵⁴

Siegmund *et al.*⁶² proposed the use of multivariate regression for correlated outcome data to analyse sibship data to test for association in the presence of linkage. Their approach uses all sib information and does not require that the exact correlation structure be specified. Let $i = 1, \dots, n$ denote the sibship, D_i denote the set of affected siblings in sibship i , n_i denote the number of affected sibs, M_i the marker genotypes, and Z_i their codings. The conditional likelihood is

$$L(\beta) = \prod_{i=1}^n \frac{\prod_{j \in D_i} \exp\{Z'_{ij}\beta\}}{\sum_{S \in C_i} \prod_{j \in S} \exp\{Z'_{ij}\beta\}}$$

where C_i denotes the set of all possible subsets for which n_i affected sibs are sampled from the i th sibship. The PHREG procedure in SAS can be used to fit the conditional logistic model, and a robust variance estimate can be used to compute a Wald test that is valid for testing association in the presence of linkage.

Finally, we note that Schaid and Rowland⁶³ proposed a general multivariate score statistic that is applicable to designs using parents as controls, sibs as controls, or unrelated individuals as controls. Their method generalizes the S-TDT of Spielman and Ewens⁴⁹ and the DAT of Boehnke and Langefeld.⁴⁸

2.4 Nuclear families with only one parent available

Curtis and Sham⁴⁶ observed that when one parent is missing, discarding parents with ambiguities may lead to higher false positive rates. For example, if the offspring of a heterozygous parent Mm is of genotype MM or mm, we can infer that the parent

transmits *M* or *m* to the offspring. If the offspring has genotype *Mm*, the family cannot be used. This method can introduce bias depending on the allele frequency in the general population. To make use of families with only one available parent, Sun *et al.*^{64,65} proposed two test statistics that lead to unbiased tests of linkage/association when only one parent is available. The first statistic, called 1-TDT, is valid if either of the two assumptions holds: (1) males and females with the same genotype have the same mating preference; and (2) father and mother are missing with the same probability given that one of them is missing. Their second test statistic is valid even when both assumptions fail. The same approach has been extended to analyse quantitative traits.⁶⁶

Wang and Sun⁶⁷ derived sample sizes required to detect linkage disequilibrium for the S-TDT and 1-TDT. Under a variety of genetic models, the sample size needed for the 1-TDT is roughly the same as the sample size needed for the S-TDT with one affected and one unaffected sibs, and is about twice the sample size needed for the TDT.

2.5 General pedigrees

One limitation of the tests discussed above is that they are not applicable to large pedigrees, although multi-generation pedigrees are routinely collected in practice. Martin *et al.*⁶⁸ proposed the pedigree disequilibrium test (PDT) to analyse linkage disequilibrium in general pedigrees. The basic idea is to collect all possible triads from informative nuclear families and all possible discordant sib pairs from informative discordant sibships in a single pedigree as a unit in the test statistic. The false positive rate is correctly controlled by using a variance estimate that is unbiased in the presence of linkage. Using the PDT can lead to substantial gains in power when extended pedigree data are available, and may increase power even without extended family information.

Rabinowitz and Laird⁶⁹ developed a general approach to estimating statistical significance levels by comparing test statistics to their distributions conditional on the minimum sufficient statistic under the null hypothesis of no linkage or the null hypothesis of no association for arbitrary genetic models, sampling strategies, and population structures. Their approach is applicable to an arbitrary pedigree structure and any pattern of missing genotypes.

3 Quantitative traits

We have so far focused our attention on qualitative traits. However, many traits are quantitative, and quantitative phenotypes contain more information than is provided by binary phenotypes. Many approaches have been developed in the last several years to analyse quantitative traits using family-based designs.

Among five statistics developed by Allison to analyse quantitative traits,⁷⁰ TDT_{Q5} was found to be preferable under a variety of genetic models. This statistic can be described as follows. For the three informative mating types (*MM* × *Mm*, *Mm* × *Mm*, and *Mm* × *mm*), two regression models can be carried out to regress the offspring phenotype value either using only parental mating types, or using both parental

mating types and allele transmission information from parents to the offspring. The TDT_{Q5} is basically an F-test that compares the fit of the two models. A similar approach developed by Xiong *et al.*⁷¹ compares the average trait values of offspring inheriting one allele versus the other from heterozygous parents. George *et al.*⁷² proposed a linear-regression approach for quantitative traits with the trait value, Y , as the dependent variable. The primary independent variable in the model is, for each individual, the transmission status X of the associated allele. In addition to X , other covariates can be incorporated into the regression model. The familial correlations among pedigree members are incorporated by means of the association model. An alternative regression model was proposed by Zhu and Elston,⁷³ and simulation studies have been performed to compare the power of these two regression approaches.⁷⁴ Assuming a random sample of individuals, Yang *et al.*⁷⁵ developed a similar approach by augmenting additional regressors in linear regression models.

An alternative method was introduced by Rabinowitz.⁷⁶ As in the discussion above, we use n to denote the number of families and n_i the number of offspring in the i th family. Denote the trait value of the j th offspring in the i th family by Q_{ij} . We further define the following index function: $Y_{ij}^{(m)} = 1/2$ (or $1/2$) if the mother in the i th family is heterozygous and transmits the M (or m) allele to the j th offspring, and $Y_{ij}^{(m)} = 0$ if the mother is homozygous. We similarly define $Y_{ij}^{(f)}$ for the father. Under the null hypothesis of no linkage between the marker locus and the quantitative trait loci, the trait value and the index functions $Y_{ij}^{(f)}$ and $Y_{ij}^{(m)}$ are conditionally independent, given the parental alleles. Thus, for any constant c , conditional on the trait values and the parental genotypes

$$s(c) = \sum_{i=1}^n \sum_{j=1}^{n_i} (Q_{ij} - c)(Y_{ij}^{(f)} + Y_{ij}^{(m)})$$

has mean 0. The test statistic proposed by Rabinowitz takes the form of $s(c)/\sigma(c)$, where $\sigma(c)$ is an estimate of the conditional variance of $s(c)$. Rabinowitz suggested one use the trait average of all the children in all the families to replace c . This approach was recently generalized to include families with missing parental information by Sun *et al.*⁶⁶ Because no assumption is made about the distribution of the trait values under this approach, the tests are valid for any types of sampling schemes based on the phenotypes of the individuals.

To use information from all available offspring, Monks and Kaplan⁷⁷ proposed three statistical tests for quantitative traits: T_{QP} , T_{QS} , and T_{QPS} . The T_{QP} uses parental genotype information and is identical to the test proposed by Rabinowitz.⁷⁶ When no parental information is available, the T_{QS} is calculated using families with at least two sibs having different genotypes. The third statistic T_{QPS} is the combination of T_{QP} and T_{QS} .

Quantitative traits also can be analysed through a likelihood framework.³² Assume the trait for a given individual has a normal distribution conditional on his/her genotype, i.e. $z \sim N(\mu_g, \sigma^2)$. Conditional on the trait value z and the parental genotypes g_m and g_f , the posterior probability that the genotype g_c is transmitted is

$$P(c|z, g_m, g_f) = \frac{\phi[(z - \mu_{g_c})/\sigma]}{\sum_g \phi[(z - \mu_g)/\sigma]}$$

where ϕ is the standard normal density function, and the sum is over all possible transmissions from the parents to the offspring. If we assume

$$h[\mu_{(i,j)}] = h(\mu) + \beta_i + \beta_j$$

the contribution of each family to the i th element of overall score statistic is

$$\frac{1}{h'(\mu)\sigma^2} (z - \mu) \left[N(i, g_c) - \frac{1}{4} \sum_g N(i, g) \right]$$

where $N(i, g_c)$ is the number of times haplotype i is transmitted from parents to the offspring.

Parallel to the developments for qualitative traits, statistical methods have been developed for quantitative traits when only genotypes from sibs are available. Allison *et al.*⁷⁸ proposed two tests, a mixed-effects model and a permutation test, to test the null hypothesis of no linkage between the marker and the trait locus using sibship data. The mixed-effects model has the following form

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where μ is a constant, the α_i are the random effects for sibship i , the β_j are the fixed effects for allele j , and the interaction terms, the $(\alpha\beta)_{ij}$, are modelled as random effects. The mix-effects model allows the straightforward inclusion of covariates and other genes. For the permutation test, the statistic is

$$S = \frac{k-1}{k} \sum_{j=1}^k \frac{[\sum_{i=1}^n (\sum_{l=1}^{n_i} Y_{il} - \mu_{ij})]^2}{\sum_{i=1}^n V_{ij}}$$

where μ_{ij} is the permutation mean of the trait value for j th allele in the i th sibship, V_{ij} is the permutation variance of the trait value for the j th allele in the i th sibship. Simulation results showed that the permutation procedure generally has greater power and, furthermore, it has the advantage of being distribution free.

Fulker *et al.*⁷⁹ extended maximum variance components procedures for mapping QTLs in sib pairs to allow for a simultaneous test of allelic association. Their model partitions association into between- and within-pairs components, and a robust test is constructed on the basis of the within-pair component. When the power of this method using simple random sibship samples was compared to the power of linkage methods, Sham *et al.*⁸⁰ found that the power of their association test is related to the QTL heritability and the square of the linkage disequilibrium measure, and the power of linkage is related to the square of the QTL heritability. Abecasis *et al.*⁸¹ further extended this approach to accommodating an arbitrary number of sibs, with or without parental genotypes.

4 Other topics

4.1 TDT for X-linked markers

Ho and Bailey-Wilson⁸² extended the TDT methods to test for linkage between X-linked markers and diseases that affect either males only or both genders. The basic statistics parallel the methods for the TDT and the S-TDT for autosomal markers, respectively extending the TDT and the S-TDT. Similarly, Horvath *et al.*⁸³ proposed two procedures: the XS-TDT and the XRC-TDT that extended the S-TDT⁴⁹ and the RC-TDT,⁵⁵ respectively. To extend the XS-TDT, they divided each family into two strata, with one stratum consisting of daughters and another stratum consisting of sons to take into account different disease prevalences in the two populations. To extend the RC-TDT to the XRC-TDT, three mating patterns were distinguished: (1) both parental genotypes are missing, (2) maternal genotype is missing, and (3) paternal genotype is missing. Simulation studies showed that the XRC-TDT is more powerful than the XS-TDT.

4.2 TDT for multiple tightly linked markers

When multiple markers are studied within a candidate region, one strategy would be to analyse each marker separately and then adjust for multiple comparisons by the Bonferroni correction. Although the Bonferroni correction is usually done, when markers are in linkage disequilibrium, such practice may be conservative. Lazzeroni and Lange²³ proposed an alternative procedure as follows. Let $T_i = t_i$ denote the test statistic at marker i , and the corresponding p -value is p_i . Under the null hypothesis, the p -value is uniformly distributed on $[0,1]$. Let H_0 denote the combined hypothesis that no transmission disequilibrium at any of the markers. The adjusted p -value is

$$\tilde{p}(p) = P[\min p_i \leq p | H_0]$$

This is the distribution function of the statistic $\min p_i(T_i)$ under H_0 evaluated at the point p . The adjusted p -value is bounded above by the usual Bonferroni correction, and it can be estimated by Monte Carlo simulations. The Monte Carlo approach to testing association for multiple markers was also discussed by McIntyre *et al.*⁸⁴ One disadvantage of this approach is that it ignores possible dependence among the markers, and such dependence may provide valuable information for linkage.

Wilson⁸⁵ extended the TDT to two markers following the likelihood ratio test proposed by Sham and Curtis.¹⁶ Sethuraman¹⁵ considered more marker configurations and developed a test similar to the TDT_{SE} , called the T^2DT_{SE} . Simulations showed that the T^2DT_{SE} is more powerful or as powerful as the likelihood ratio tests discussed by Wilson.⁸⁵ Clayton and Jones³² proposed an approach borrowed from spatial statistics by relating haplotype relative risk to haplotype similarity. All these methods assume that the haplotypes are known in the parents. Therefore, they are not applicable to data collected on nuclear families, where haplotypes in the parents may not be uniquely resolved.

As pointed out by Dudbridge *et al.*,⁸⁶ a necessary condition for haplotype ambiguity is that there is a locus for which both parents and offspring have the same heterozygous genotype and another locus for which both parents and offspring do not have the same homozygous genotype. Clayton⁸⁷ proposed that one estimate haplotype frequencies and construct a likelihood considering all possible solutions. However, his

method is not robust for population stratification. Dudbridge *et al.*⁸⁶ developed an unbiased test for individual haplotypes by calculating the correct variance for the transmission count within a family, using information from multiple siblings if they are available. However, families with ambiguous haplotypes have to be discarded from analysis, resulting in loss of information. Zhao *et al.*⁸⁸ proposed that one construct a transmission/non-transmission table by assigning families with ambiguities to compatible haplotype groups based on a set of hypothetical haplotype frequencies. They showed that under the null hypothesis of no linkage or no association, the reconstructed transmission/non-transmission table retains symmetry. Therefore, the null hypothesis of no association or no linkage can be tested by examining the symmetry of the reconstructed transmission/non-transmission table.

4.3 Parent-of-origin effects

Methods to detect parent-of-origin effects using the log-linear models were discussed by Weinberg *et al.*⁸⁹ However, Weinberg⁹⁰ later noted that the formulation was not strictly correct if the gene under study is in linkage disequilibrium with a different disease-susceptibility gene. She also pointed out the problem with alternative methods. For example, it is natural to compare the frequency of transmission from heterozygous mothers versus that from heterozygous fathers. However, if families in which both parents are heterozygous are included, they can contaminate the comparison because of statistical dependency between maternal and paternal transmissions. To remedy this problem, she proposed the following model

$$\log \left(\frac{P[M > P | \text{mating type}, C]}{P[M < P | \text{mating type}, C]} \right) = \alpha I_{(C=1)} + \beta I_{(M+P>1)} + \gamma [I_{(M+P=1)} - I_{(M+P>2)}]$$

where I is the indicator variable, M , P , and C are the number of M alleles in the mother, father, and child, respectively. The event $M > P$ corresponds to the mother carrying more variant allele than the father, and the event $M < P$ is similarly defined. The null hypothesis of no parent-of-origin effects, i.e. $\alpha = 0$, can be tested through the likelihood ratio test.

4.4 Gene-environment (GxE) interaction

In the absence of biological mechanisms for how genes and environmental factors interact, statistical models of interaction can be useful. Statistical interaction normally means that the joint effects of genetic and environmental factors cannot be added, if an additive model is assumed, or cannot be multiplied, if a multiplicative model is assumed. Statistical interactions depend on the scale of measurement.

Umbach and Weinberg⁹¹ remarked that the approach that compares the transmission rates from heterozygous parents to exposed versus unexposed probands may be invalid. For example, when there is a structured population where mating type and exposure are correlated, transmission rates can differ between exposed and unexposed triads even if exposure has no effect at all. Using the notation for M , P , and C defined above, Umbach and Weinberg proposed to use the following model to study $G \times E$ interactions

$$\log [N_{ice}] = \mu_i + \delta_i I_{\{E=1\}} + \beta_c I_{\{C=c\}} \\ + \eta_c I_{\{C=c\}} I_{\{E=e\}} + \log(2) I_{\{(M,P,C)=(1,1,1)\}}$$

where N_{ice} denotes the expected number of family trios with the i th mating type, $C = c$, and $E = e$, where E is the binary exposure variable. The null hypothesis of no interaction can be tested by setting $\eta_1 = \eta_2 = 0$. This test is equivalent to a test studied by Schaid.⁹² In the study of $G \times E$ interactions using case-parent data, one implicit assumption is that, conditional on the parental genotypes, an individual's exposure status is independent of his or her genotype at the candidate locus.

Schaid⁹² compared the required sample size to detect $G \times E$ interactions using case-parents versus that using the traditional case-control design. Witte *et al.*⁹³ considered more alternative designs, including sibling controls and cousin controls. Sibling and cousin controls are more efficient when estimating an interaction comprising a rare major susceptible gene. However, for a common gene, sibling and cousin controls are less efficient than population controls. Pseudosib controls are generally less efficient than population controls for estimating $G \times E$ interactions.

5 Conclusions

Mapping genes for complex traits presents great challenges for human geneticists. It has become clear that conventional linkage analysis as a tool for mapping disease loci is of limited potential. On the other hand, traditional case-control designs using unrelated controls may be biased in the presence of population stratification. Family-based association designs are a compromise between the above two approaches. Assuming a multiplicative disease model and focusing on families consisting of parents and two affected siblings, Risch and Merikangas³ compared the power of the TDT with the traditional affected sib-pair method that only uses allele sharing status. They concluded that genes with effects as low as 1.5 could be detected using samples of reasonable sizes. McGinnis⁹⁴ generalized their results to general disease models and reached similar conclusions. Similarly, for family association studies of quantitative traits, the TDT-type tests are at least one order of magnitude more efficient than common sib-pair tests of linkage when a candidate gene is available.⁷⁰ The power of the TDT can be further increased by recruiting families with certain structures.²⁰

Despite the great promise of family-based association designs, the primary limitation of such studies is the lack of complete linkage disequilibrium between disease gene mutation and the tested candidate allele. Tu and Whittemore⁹⁵ examined the effects of a marker allele that is not in complete linkage disequilibrium with the disease susceptible allele and of an allele frequency difference between the marker allele and the disease allele. When the frequencies of disease and marker alleles are highly discrepant, the linkage test may be more powerful if there are small departures from maximum linkage disequilibrium. Their results agreed with those made by Abel and Müller-Myhsok⁹⁶ and by Clerget-Darpoux.⁹⁷ Statistical power may be further compromised by allelic heterogeneity.⁹⁸

Although the SNP project will generate hundreds of thousands of genetic markers throughout the human genome, the linkage disequilibrium patterns are hard to predict. They are both locus specific⁹⁹ and population specific.¹⁰⁰ The success of the family-based association paradigm critically depends on our understanding of the linkage disequilibrium patterns across the human genome in worldwide populations.

Case-control association study designs have often been criticized for inducing false positives due to population stratification. However, in terms of efficiency, case-control designs are superior to family-based association tests both for qualitative traits^{101,102} and quantitative traits.¹⁰³ Recently, several articles have appeared to use genomic markers to control population stratification in the analysis of case-control data.^{104–107} These novel approaches offer great promises because they may both have greater power than family-based association designs and they may prove to be robust against potential population stratification. In addition, case-control studies have the further advantage of easy sample collection and the potential of tremendous reduction in genotyping efforts using DNA pooling.¹⁰²

Given that many statistical methods have been proposed in the last several years for family-based association studies, the performance of these methods is of great interest to human geneticists who study complex traits. Many studies have been done to compare various methods, and the relative performance clearly depends on the genetic models used in these studies. Because the mode of inheritance for complex diseases is usually unknown, methods that perform well under a wide range of models are certainly desirable. As more and more approaches are introduced in the literature, systematic comparisons are always needed to give guidelines to practitioners.

To maximize statistical power to detect linkage, Huang and Jiang¹⁰⁸ proposed a disequilibrium-maximum-likelihood-binomial test for linkage. This method appears to efficiently combine linkage information from the IBD-sharing and the allele-specific IBD sharing information. Further developments along this line are clearly warranted to develop a unified approach that simultaneously incorporates linkage and linkage disequilibrium information.

Although a theoretical framework has been laid out to study transmission disequilibrium in general pedigrees,³⁴ the test statistic proposed by Martin *et al.*⁶⁸ represents a first step to develop statistical tests to detect association that are applicable to large pedigrees. The difficulty lies in the requirement that the methods have to be robust in the presence of population stratification, which sets it apart from the methods in traditional linkage analysis that usually invoke Hardy-Weinberg equilibrium. Novel methods that can jointly analyse all pedigree members, handle both qualitative and quantitative traits, and are robust to population stratification will be of great value for the study of complex traits.

With the availability of large numbers of genetic markers in the human genome, it is already common practice to genotype many genetic markers in a candidate gene region. One difficulty of analysing multiple markers is that haplotypes may not be uniquely identified from pedigree data. Even if haplotype information is known, one potential limitation involving multiple markers is the very large number of possible haplotypes, which will likely reduce the power of the statistical tests. For a single marker with many alleles, Kaplan *et al.*¹⁰⁹ considered a general method to collapse alleles to two allelic classes that minimizes the reduction in the non-centrality

parameter for an admixed population with two founding populations. For multiple markers, Clayton and Jones³² proposed to relate haplotype relative risk to haplotype similarity as a way to reduce the number of haplotype classes in studying transmission disequilibrium. Methods similar to those proposed by Templeton and his colleagues¹¹⁰ can also be employed to reduce the complexities in examining haplotypes involving many polymorphic sites. Both theoretical and empirical studies are needed to develop and evaluate statistical methods that can reduce the complexity in the analysis of multiple markers.

Undoubtedly, the ultimate data available to human geneticists will be DNA sequences from each individual involved in each genetic study. Although such data are already on the horizon, statistical tools to efficiently use this information to map disease genes are lacking. The great challenges facing statistical geneticists in the coming years are to develop statistically powerful and computationally feasible methods to fully utilize such information, and to search for optimal study strategies to map complex disease genes in the post genome era.

Acknowledgements

The author thanks Drs Warren J Ewens, Theodore R Holford, Kenneth K Kidd, Fengzhu Sun, and an anonymous referee for their constructive comments. This work was supported in part by grants GM59507 and HD36834 from the National Institutes of Health and Research Grant FY98-0752 from the March of Dimes Birth Defects Foundation.

References

- 1 Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Program description – Center-d'Etude-du-Polymorphisme-Humain (CEPH) – collaborative genetic-mapping of the human genome. *Genomics* 1990; **6**: 575–77.
- 2 Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *American Journal of Human Genetics* 1998; **63**: 861–69.
- 3 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**(5281):1516–17.
- 4 Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 1993; **52**: 506–16.
- 5 Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3–5,13,14 and type-2 diabetes mellitus: an association in American-Indians with genetic admixture. *American Journal of Human Genetics* 1988; **43**: 520–26.
- 6 Kang AM, Palmatier MA, Kidd KK. Global variation of a 40–bp VNTR in the 3'-untranslated region of the dopamine transporter gene (SLC6A3). *Biological Psychiatry* 1999; **46**: 151–60.
- 7 Rubinstein P, Walker M, Carpenter C *et al*. Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the 'haplo-delta' (Dh) estimates in juvenile diabetes from three racial groups. *Human Immunology* 1981; **3**: 384.
- 8 Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* 1987; **51**: 227–33.
- 9 Terwilliger JD, Ott J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity* 1992; **42**: 337–46.
- 10 Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies using cases and parents. *American Journal of Human Genetics* 1994; **55**: 402–409.
- 11 Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics* 1995; **57**: 455–64.
- 12 Knapp M, Seuchter SA, Baur MP. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families.

- American Journal of Human Genetics* 1993; **52**: 1085–93.
- 13 Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association [editorial]. *American Journal of Human Genetics* 1996; **59**: 983–89.
 - 14 Thomson G. Mapping disease genes: family-based association studies. *American Journal of Human Genetics* 1995; **57**: 487–98.
 - 15 Sethuraman B. *Topics in statistical genetics*. Berkeley: University of California, 1997.
 - 16 Sham PC, Curtis D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of Human Genetics* 1995; **59**: 323–36.
 - 17 Morris AP, Whittaker JC, Curnow RN. A likelihood ratio test for detecting patterns of disease-marker association. *Annals of Human Genetics* 1997; **61**: 335–50.
 - 18 Zhao H. The interpretation of the parameters in the transmission/disequilibrium test [letter]. *American Journal of Human Genetics* 1999; **64**: 326–28.
 - 19 Bickeboller H, Clerget-Darpoux F. Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genetic Epidemiology* 1995; **12**: 865–70.
 - 20 Whittaker JC, Lewis CM. The effect of family structure on linkage tests using allelic association. *American Journal of Human Genetics* 1998; **63**: 889–97.
 - 21 Sham P. Transmission/disequilibrium tests for multiallelic loci. *American Journal of Human Genetics* 1997; **61**: 774–78.
 - 22 Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology* 1996; **13**: 423–49.
 - 23 Lazzeroni LC, Lange K. A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity* 1998; **48**: 67–81.
 - 24 Bradley RA, Terry ME. Rank analysis of incomplete block designs. 1. The method of paired comparisons. *Biometrika* 1952; **39**: 324–45.
 - 25 Miller MB. Genomic scanning and the transmission/disequilibrium test: analysis of error rates. *Genetic Epidemiology* 1997; **14**: 851–56.
 - 26 Jin K, Speed TP, Klitz W, Thomson G. Testing for segregation distortion in the HLA complex. *Biometrics* 1994; **50**: 1189–98.
 - 27 Harley JB, Moser KL, Neas BR. Logistic transmission modeling of simulated data. *Genetic Epidemiology* 1995; **12**: 607–12.
 - 28 Rice JP, Neuman RJ, Hoshaw SL, Daw EW, Gu C. TDT with covariates and genomic screens with mod scores: their behavior on simulated data. *Genetic Epidemiology* 1995; **12**: 659–64.
 - 29 Waldman ID, Robinson BF, Rowe DC. A logistic regression based extension of the TDT for continuous and categorical traits. *Annals of Human Genetics* 1999; **63**: 329–40.
 - 30 Sinsheimer JS, Blangero J, Lange K. Gamete-competition models. *American Journal of Human Genetics* 2000; **66**: 1168–72.
 - 31 Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* 1991; **47**: 53–61.
 - 32 Clayton D, Jones H. Transmission/disequilibrium tests for extended marker haplotypes. *American Journal of Human Genetics* 1999; **65**: 1161–69.
 - 33 Lunetta KL, Faraone SV, Biederman J, Laird NM. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *American Journal of Human Genetics* 2000; **66**: 605–14.
 - 34 Whittemore AS, Tu IP. Detection of disease genes by use of family data. I. Likelihood-based theory. *American Journal of Human Genetics* 2000; **66**: 1328–40.
 - 35 Martin ER, Kaplan NL, Weir BS. Tests for linkage and association in nuclear families. *American Journal of Human Genetics* 1997; **61**: 439–48.
 - 36 Wicks J. Exploiting excess sharing: A more powerful test of linkage for affected sib pairs than the transmission/disequilibrium test. *American Journal of Human Genetics* 2000; **66**: 2005–08.
 - 37 Cleves MA, Olson JM, Jacobs KB. Exact transmission-disequilibrium tests with multiallelic markers. *Genetic Epidemiology* 1997; **14**: 337–47.
 - 38 Morris AP, Curnow RN, Whittaker JC. Randomization tests of disease-marker associations. *Annals of Human Genetics* 1997; **61**: 49–60.
 - 39 Whittaker JC, Thompson DJ. Finite-sample properties of family-based association tests. *American Journal of Human Genetics* 1999; **64**: 910–15.
 - 40 Knapp M. A note on power approximations for the transmission/disequilibrium test.

- American Journal of Human Genetics* 1999; **64**: 1177–85.
- 41 Schaid DJ. Likelihoods and TDT for the case-parents design. *Genetic Epidemiology* 1999; **16**: 250–60.
 - 42 Kaplan NL, Martin ER, Weir BS. Power studies for the transmission/disequilibrium tests with multiple alleles. *American Journal of Human Genetics* 1997; **60**: 691–702.
 - 43 Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics* 1993; **53**: 1114–26.
 - 44 Schaid DJ, Li H. Genotype relative-risks and association tests for nuclear families with missing parental data. *Genetic Epidemiology* 1997; **14**: 1113–18.
 - 45 Martin RB, Alda M, MacLean CJ. Parental genotype reconstruction: applications of haplotype relative risk to incomplete parental data. *Genetic Epidemiology* 1998; **15**: 471–90.
 - 46 Curtis D, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing [letter]. *American Journal of Human Genetics* 1995; **56**: 811–12.
 - 47 Monks SA, Kaplan NL, Weir BS. A comparative study of sibship tests of linkage and/or association. *American Journal of Human Genetics* 1998; **63**: 1507–16.
 - 48 Boehnke M, Langefeld CD. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *American Journal of Human Genetics* 1998; **62**: 950–61.
 - 49 Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* 1998; **62**: 450–58.
 - 50 Schaid DJ, Rowland C. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *American Journal of Human Genetics* 1998; **63**: 1492–506.
 - 51 Laird NM, Blacker D, Wilcox M. The sib transmission/disequilibrium test is a Mantel-Haenszel test. *American Journal of Human Genetics* 1998; **63**: 1915–16.
 - 52 Ewens WJ, Spielman RS. Reply to Laird *et al.* *American Journal of Human Genetics* 1998; **63**: 1915–16.
 - 53 Spielman RS, Ewens WJ. TDT clarification [letter]. *American Journal of Human Genetics* 1999; **64**: 668.
 - 54 Curtis D. Use of siblings as controls in case-control association studies [published erratum appears in *Annals of Human Genetics* 1998; **62**: 89]. *Annals of Human Genetics* 1997; **61**: 319–33.
 - 55 Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *American Journal of Human Genetics* 1999; **64**: 861–70.
 - 56 Knapp M. Using exact p values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib transmission/disequilibrium test. *American Journal of Human Genetics* 1999; **65**: 1208–10.
 - 57 Teng J, Risch N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Research* 1999; **9**: 234–41.
 - 58 Weinberg CR. Allowing for missing parents in genetic studies of case-parent triads. *American Journal of Human Genetics* 1999; **64**: 1186–93.
 - 59 Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of 'case-parent triads'. *American Journal of Epidemiology* 1998; **148**: 893–901.
 - 60 Horvath S, Laird NM. A discordant-sibship test for disequilibrium and linkage: no need for parental data [see comments]. *American Journal of Human Genetics* 1998; **63**: 1886–97.
 - 61 Curtis D, Miller MB, Sham PC. Combining the sibling disequilibrium test and transmission/disequilibrium test for multiallelic markers. *American Journal of Human Genetics* 1999; **64**: 1785–86.
 - 62 Siegmund KD, Langholz B, Kraft P, Thomas DC. Testing linkage disequilibrium in sibships. *American Journal of Human Genetics* 2000; **67**: 244–48.
 - 63 Schaid DJ, Rowland CM. Quantitative trait transmission disequilibrium test: allowance for missing parents. *Genetic Epidemiology* 1999; **17**(Suppl 1): S307–12.
 - 64 Sun FZ, Flanders WD, Yang QH, Khoury MJ. A new method for estimating the risk ratio in studies using case-parental control design. *American Journal of Epidemiology* 1998; **148**: 902–909.
 - 65 Sun F, Flanders WD, Yang Q, Khoury MJ. Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *American Journal of Epidemiology* 1999; **150**: 97–104.
 - 66 Sun F, Flanders WD, Yang Q, Zhao H. Transmission/disequilibrium tests for quantitative traits. *Annals of Human Genetics* 2000; in press.
 - 67 Wang D, Sun F. Sample sizes for the transmission disequilibrium tests: TDT, S-

- TDT and 1-TDT. *Communications in Statistics – Theory and Methods* 2000; **29**: 1129–42.
- 68 Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *American Journal of Human Genetics* 2000; **67**: 146–54.
 - 69 Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity* 2000; **50**: 211–23.
 - 70 Allison DB. Transmission-disequilibrium tests for quantitative traits [published erratum appears in *American Journal of Human Genetics* 1997 Jun; **60**: 1571]. *American Journal of Human Genetics* 1997; **60**: 676–90.
 - 71 Xiong MM, Krushkal J, Boerwinkle E. TDT statistics for mapping quantitative trait loci. *Annals of Human Genetics* 1998; **62**: 431–52.
 - 72 George V, Tiwari HK, Zhu X, Elston RC. A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *American Journal of Human Genetics* 1999; **65**: 236–45.
 - 73 Zhu X, Elston RC. Transmission/disequilibrium test for quantitative traits. *Genetic Epidemiology* 2001; **20**: 57–74.
 - 74 Zhu X, Elston RC. Power comparison of regression methods to test quantitative traits for association and linkage. *Genetic Epidemiology* 2000; **18**: 322–30.
 - 75 Yang Q, Rabinowitz D, Isasi C, Shea S. Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits. *Human Heredity* 2000; **50**: 227–33.
 - 76 Rabinowitz D. A transmission disequilibrium test for quantitative trait loci. *Human Heredity* 1997; **47**: 342–50.
 - 77 Monks SA, Kaplan NL. Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. *American Journal of Human Genetics* 2000; **66**: 576–92.
 - 78 Allison DB, Heo M, Kaplan N, Martin ER. Sibling-based tests of linkage and association for quantitative traits. *American Journal of Human Genetics* 1999; **64**: 1754–63.
 - 79 Fulker DW, Cherny SS, Sham PC, Hewitt JK. Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics* 1999; **64**: 259–67.
 - 80 Sham P, Cherny SS, Purcell S, Hewitt JK. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics* 2000; **66**: 1616–30.
 - 81 Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* 2000; **66**: 279–92.
 - 82 Ho GY, Bailey-Wilson JE. The transmission/disequilibrium test for linkage on the X chromosome. *American Journal of Human Genetics* 2000; **66**: 1158–60.
 - 83 Horvath S, Laird NM, Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. *American Journal of Human Genetics* 2000; **66**: 1161–7.
 - 84 McIntyre LM, Martin ER, Simonsen KL, Kaplan NL. Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. *Genetic Epidemiology* 2000; **19**: 18–29.
 - 85 Wilson SR. On extending the transmission/disequilibrium test (TDT). *Annals of Human Genetics* 1997; **61**: 151–61.
 - 86 Dudbridge F, Koeleman BP, Todd JA, Clayton DG. Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *American Journal of Human Genetics* 2000; **66**: 2009–12.
 - 87 Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *American Journal of Human Genetics* 1999; **65**: 1170–77.
 - 88 Zhao H, Zhang S, Merikangas KR, Wildenaur D, Sun F, Kidd KK. Transmission/disequilibrium tests for multiple tightly linked markers. *American Journal of Human Genetics* 2000; **67**: 936–46.
 - 89 Weinberg CR, Wilcox AJ, Lie RT. 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. *American Journal of Human Genetics* 1998; **62**: 969–78.
 - 90 Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *American Journal of Human Genetics* 1999; **65**: 229–35.
 - 91 Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. *American Journal of Human Genetics* 2000; **66**: 251–61.
 - 92 Schaid DJ. Case-parents design for gene-environment interaction. *Genetic Epidemiology* 1999; **16**: 261–73.

- 93 Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *American Journal of Epidemiology* 1999; **149**: 693–705.
- 94 McGinnis RE. Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT). *Annals of Human Genetics* 1998; **62**: 159–79.
- 95 Tu IP, Whittemore AS. Power of association and linkage tests when the disease alleles are unobserved. *American Journal of Human Genetics* 1999; **64**: 641–9.
- 96 Abel L, Muller-Myhsok B. Maximum-likelihood expression of the transmission/disequilibrium test and power considerations [letter]. *American Journal of Human Genetics* 1998; **63**: 664–67.
- 97 Clerget-Darpoux F, Babron MC, Bickeboller H. Comparing the power of linkage detection by the transmission disequilibrium test and the identity-by-descent test. *Genetic Epidemiology* 1995; **12**: 583–88.
- 98 Slager SL, Huang J, Vieland VJ. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genetic Epidemiology* 2000; **18**: 143–56.
- 99 Clark AG, Weiss KM, Nickerson DA *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics* 1998; **63**: 595–612.
- 100 Kidd JR, Pakstis AJ, Zhao H *et al.* Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus (PAH) in a global representation of populations. *American Journal of Human Genetics* 2000; **66**: 1882–99.
- 101 Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. *Proceedings of the National Academy of Sciences of the United States of America* 1998; **95**: 11389–93.
- 102 Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Research* 1998; **8**: 1273–88.
- 103 van den Oord EJCG. A comparison between different designs and tests to detect QTLs in association studies. *Behavioral Genetics* 1999; **29**: 245–56.
- 104 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
- 105 Bacanu SA, Devlin B, Roeder K. The power of genomic control. *American Journal of Human Genetics* 2000; **66**: 1933–44.
- 106 Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* 1999; **65**: 220–28.
- 107 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *American Journal of Human Genetics* 2000; **67**: 170–81.
- 108 Huang J, Jiang Y. Linkage detection adaptive to linkage disequilibrium: the disequilibrium maximum-likelihood-binomial test for affected-sibship data. *American Journal of Human Genetics* 1999; **65**: 1741–59.
- 109 Kaplan NL, Martin ER, Morris RW, Weir BS. Marker selection for the transmission/disequilibrium test, in recently admixed populations. *American Journal of Human Genetics* 1998; **62**: 703–12.
- 110 Templeton AR, Boerwinkle E, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in drosophila. *Genetics* 1987; **117**: 343–51.