

**Transmission Disequilibrium Methods
for Family-Based Studies**

Daniel J. Schaid
Technical Report #72
July, 2004

Correspondence to: Daniel J. Schaid, Ph.D., Harwick 775, Division of Biostatistics Mayo Clinic/Foundation, Rochester, MN 55905, USA. E-mail: schaid@mayo.edu

Introduction

The study of the association of genetic markers with complex traits has generated a wide range of statistical methods, particularly those that are based on transmission-disequilibrium. Understanding the relationships of new and old ideas has been a real challenge, since it seems that a cottage-industry has emerged in the production of new methods for transmission-disequilibrium. This technical report provides a review of methods in this area through approximately 1999. It was originally intended for this report to be a book chapter, although the complete book was never completed. None the less, this review is fairly comprehensive through 1999, and will serve as a good basis for future updated editions.

Association and Linkage Disequilibrium

The genetic variants related to a disease can be evaluated in multiple ways, such as by functional cloning (identification of the defective protein first, which then leads to the defective gene), positional cloning (using genetic markers and genome-wide screens), or evaluation of candidate genes (genes with known function as likely candidates related to the disease). Association studies - often used to evaluate candidate genes - typically are based on the case-control study design with unrelated subjects. The finding of a greater frequency of a marker allele among affected subjects compared to unaffected subjects is suggestive of a functional relationship of the marker with disease. But, because genetically unrelated subjects do not provide information on inheritance, any observed association can be attributed to both genetic and non-genetic causes. Relevant genetic causes are either that the genetic marker plays a functional role in the disease process, or that the marker and disease susceptibility loci are in close proximity on the same chromosome and their alleles are statistically associated at the population level, due to common ancestry. The statistical dependence of alleles at two loci has been called linkage disequilibrium, gametic disequilibrium, and allelic association. To avoid ambiguity, allelic association shall be used to represent the dependence of alleles at two loci for any reason, but linkage disequilibrium will be reserved for the more restrictive case when both genetic linkage exists (i.e., the meiotic recombination fraction, θ , is less than $\frac{1}{2}$), and the alleles at the two loci are associated. Linkage disequilibrium (LD) is typically measured as the deviation of the relative frequency of a haplotype from its equilibrium value. If $P(D)$ and $P(M_1)$ represent the frequencies of alleles D and M_1 at the susceptibility and marker loci, respectively, and $P(D-M_1)$ represents the frequency of the $D-M_1$ haplotype, then a measure of LD is $\delta = P(D-M_1) - P(D)P(M_1)$. Linkage disequilibrium occurs when a new allele at a locus arises by mutation on a unique chromosomal background, characterized by the pattern of alleles at other loci on the same chromosome. This unique chromosomal pattern is transmitted *en bloc* from generation to generation, until recombination between loci randomly shuffles alleles, diminishing the association of alleles between loci. Although LD diminishes over generations due to recombination, the loci that are closest to the locus with the mutant allele are less likely to experience recombination events, hence maintaining the strongest association. If LD

exists, it can offer valuable information to refine the location of disease susceptibility genes, taking advantage of few ancestral recombinations.

However, other non-genetic causes of association, such as population stratification, admixture, or small sample variation, can mislead epidemiologic investigations of disease etiology. Causal inferences based on case-control studies can be difficult to interpret, particularly when cases and controls are sampled from a heterogeneous population. If a population is stratified according to multiple ethnic subgroups (i.e., without random mating among the various subgroups), then any disease that occurs at a higher frequency in a subgroup will be positively associated with any allele that happens to be more frequent in that group. A classic example is the association of haplotype of the Gm marker with non-insulin-dependent diabetes mellitus (NIDDM) among American Pima Indians. This group has both Indian and Caucasian ancestors, and the amount of admixture varies among subjects. The prevalence of NIDDM increases as the amount of Indian heritage increases, and certain Gm haplotypes are associated with Indian ancestry. Ignoring the admixture of subjects leads to a strong association of Gm haplotypes with NIDDM, yet stratifying on an index of Indian heritage (reported heritage of the parents' grandparents) removes this confounding. For further details of this example, see Khoury, et al. (Khoury et al. 1993, p. 133-135) and Knowler, et al. (Knowler et al. 1988). Another example for which population stratification has been speculated to play a role is the association of the *A1* allele of the dopamine *D2* receptor gene with alcoholism (Parsian et al. 1991; Gelernter et al. 1993; Pato et al. 1993). Meta-analyses have shown that this association varies significantly among different populations, and that the variation of the *A1* allele has been greater among different studies than between alcoholics and controls. This type of bias has motivated investigators to match cases and controls by their neighborhood of residence, or by ethnic background. However, it is difficult to measure ethnicity. Ethnicity encompasses not only common ancestry, but also shared social, religious, environmental, and cultural backgrounds. Furthermore, ethnicity is not the same as nationality. Although country of origin of parents, or grandparents, of cases and controls is commonly used as an index of ethnicity, this is still a crude measure, which may not capture all the genetic variation in a heterogeneous country (Senior and Bhopal 1994). Furthermore, in heterogeneous populations like the United States, it is not unusual to find fewer than four grandparents from the same country, leading to cases and controls with mixed nationality. Because of these difficulties in defining an appropriate control group, the relatives of cases may be an ideal source of controls that are matched on genetic ancestry.

Haplotype Relative Risk

Falk and Rubinstein (Falk and Rubinstein 1987) proposed using the parental alleles not present in the affected child (i.e., the case) as a control sample. This novel method requires measurement of the genetic marker for the case and both of its parents, in order to compare the frequency of carriers of a particular allele between cases and the controls. For example, let M_1 denote the allele of interest, and M_2 denote all other alleles. If an affected child has the genotype M_1/M_2 , and its parents have genotypes M_1/M_2 and

M_2/M_2 , then the alleles not transmitted from these parents to the child are used to create a pseudo-control. Since an M_2 allele is not transmitted from each parent, the artificial “genotype” for the pseudo-control is M_2/M_2 . Then, the frequency of cases that carry the M_1 allele can be compared with the same frequency among the pseudo-controls by creating a 2x2 table, with rows representing the cases and controls, and columns indicating carrier status. Traditional methods can be used to analyze this 2x2 table, such as Pearson’s chi-square contingency statistic and the usual cross-product odds ratio. This method, known as the Haplotype Relative Risk (HRR), will produce unbiased estimates of the conventional relative risk under certain conditions. First, the usual assumption of a rare disease is needed for the odds ratio to approximate the relative risk. But, it is also necessary for the two non-transmitted parental alleles to be representative of a random sample of alleles from the population at large. This requires no inbreeding among parents, no correlation of parental phenotypes, single ascertainment of families, no differential fertility of the disease phenotypes, and no recombination between the disease and marker loci ($\theta = 0$). When recombination occurs ($\theta > 0$), the non-transmitted alleles are not a random sample from the population, causing the HRR to underestimate the relative risk for the susceptibility locus (Knapp et al. 1993). Without linkage ($\theta = 0.5$), HRR shrinks to 1. It is worthwhile to emphasize that the relative risk from a case-control study with unrelated subjects is different from 1 whenever there is allelic association ($\delta \neq 0$), no matter what causes the association, but HRR differs from 1 only when there is both linkage and linkage disequilibrium. The null hypothesis that HRR is equal to 1 is equivalent to the compound null hypothesis $\delta(1 - 2\theta) = 0$, which mathematically states that absence of linkage ($\theta = 0.5$) or absence of association ($\delta = 0$), or both, will cause the expected value of HRR to be unity.

Note that the HRR method is used to compare the frequencies of genotypes (grouping homozygous and heterozygous carriers), so that N cases are compared with N pseudo-controls, for a total sample size of $2N$. But, under the null hypothesis of no allelic association, the transmitted and non-transmitted alleles of each parent are independent, and the transmission of alleles is independent between parents (Ott 1989). This implies that all 4 alleles of the parents of a case are independent, so that it is possible to use a total sample size of $4N$ alleles, instead of $2N$ genotypes, for more powerful statistical testing. This is achieved by first creating a 2x2 table of allele counts, where the rows classify the parental alleles as transmitted and non-transmitted, and the columns indicate the presence or absence of the allele of interest, and then applying Pearson’s contingency statistic, which has a chi-square distribution with 1df (Terwilliger and Ott 1992). This approach, which uses as controls those parental alleles never transmitted to one or more affected offspring, has been referred to as *affected family-based controls* (AFBAC) (Thomson 1995). Although this method leads to a valid test of allelic association, it is not a valid method to test for linkage in the presence of association, because when allelic association exists - say due to population stratification, the transmitted and non-transmitted alleles are no longer independent, even when the loci are not linked ($\theta = 0.5$) (Ott 1989). This dependence of alleles is not taken into account when using Pearson’s contingency statistic, which can lead to inflated Type I error rates when using this method as a test for linkage (Ewens and Spielman 1995).

Transmission/Disequilibrium Test

To develop a test specific for linkage in the presence of association, Spielman et al. (Spielman et al. 1993) considered the transmission of marker alleles from heterozygous parents to their affected child. Each pair of alleles from a heterozygous parent is considered a matched pair, and these alleles are scored as transmitted or non-transmitted to the affected child, as illustrated in Table 1 for the evaluation of allele M_1 versus all other alleles, M_2 . When there is no LD, the two parental alleles have an equal chance of being transmitted to the affected child. In contrast, the presence of LD distorts this equality, so that the frequency of allele M_1 will differ between transmitted and non-transmitted alleles. But, because we condition on the marker genotype of a parent, say M_1M_2 , the transmission of these two alleles are perfectly negatively correlated, because knowing which allele is transmitted gives information on which allele is not transmitted. McNemar's chi-square statistic for matched pairs, which accounts for this correlation, offers a valid method of comparison. Based on the notation in Table 1, McNemar's chi-square statistic, called the transmission/disequilibrium test (Spielman et al. 1993), is $TDT = (b - c)^2 / (b + c)$. For a large number of discordant pairs, $(b + c)$, this statistic has an approximate chi-square distribution with 1 df. Alternatively, for small sample sizes, exact probability values can be computed based on the binomial distribution with sample size $(b + c)$ and probability $1/2$. As with traditional matched studies in epidemiology, the ratio $r = b / c$ estimates a relative risk, which in this case is the relative risk per allele. Note that homozygous parents (cells a and d in Table 1) do not contribute information, and hence are not used for the TDT . An advantage of the transmission disequilibrium method is that it does not require a genetic model for disease transmission, which can be difficult to specify for complex traits. Furthermore, several advantages are gained by conditioning on parental marker genotypes: the influence of non-genetic associations due to population structure is eliminated; allele frequencies are not required; any dependence of parental marker genotypes due to non-random mating (e.g., assortative mating) is eliminated.

The fact that the TDT can detect linkage only in the presence of association can be understood as follows. When sampling a diseased child and its parents, the linkage phase of the parents is unknown. Linkage phase refers to which alleles at the disease and marker loci occur together on each of the two haplotypes of a parent. For example, if a

Table 1. Counts of transmitted/non-transmitted marker allele pairs for $2N$ parents of N affected children: two marker alleles M_1 and M_2 .

Transmitted Allele	Non-transmitted Allele		Total
	M_1	M_2	
M_1	a	b	$a+b$
M_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2N$

parent has the disease genotype D/N and marker genotype M_1 / M_2 , the D can occur on a haplotype with either M_1 or M_2 , resulting in the parent's linkage phase as either $D-M_1 / N-M_2$ or $D-M_2 / N-M_1$. A recombinant child under one phase is non-recombinant under the other phase. Now, first consider when linkage exists, but there is no allelic association. In this case, the parent's two linkage phases are equally likely, so approximately one-half of these types of parents will have the $D-M_1$ haplotype, and one-half the $D-M_2$ haplotype. Without linkage phase information, and a single affected child, there is no information to classify the child's unobserved meiotic event as recombinant or non-recombinant. This will cause the parental marker alleles to appear to segregate randomly to their children, and hence the TDT statistic will not detect a signal. In contrast, linkage information would exist if multiple affected children were sampled, and allele sharing within families were evaluated. Now consider when there is no linkage, but population association exists. Because of the association, the parent's two phases are not equally likely, but since there is no linkage, the recombinant and non-recombinant meioses are equally likely, implying random segregation of the parental marker alleles, and hence the TDT will not detect a signal. It is only when both linkage ($\theta < 0.5$) and allelic association ($\delta \neq 0$) coexist that the apparent transmission of alleles from heterozygous parents will deviate from the Mendelian 1:1 chance segregation. For these reasons, the TDT is used to test the compound null hypothesis $H_0: \delta(1 - 2\theta) = 0$. Note that when sampling families with a single affected child, the parameters for linkage (θ) and LD (δ) are completely confounded, meaning that we cannot obtain separate estimates of them. Nonetheless, the TDT can be used either to test for linkage in the presence of association, or to test for association in the presence of linkage, which can be useful for fine localization of disease genes once linkage to a chromosomal region is detected.

A limitation of the TDT method is that it does not use a control group, but rather relies on Mendelian segregation (i.e., 1:1) of marker alleles under the null hypothesis. If the segregation of marker alleles in a random person is distorted from Mendelian transmission for any reason, then results from the TDT can be misleading. An example of segregation distortion is provided by *Drosophila melanogaster*, in which the Segregation distorter gene causes segregation distortion in males but not in females, due to the influence of this gene on sperm dysfunction (Hartl and Hiraizumi 1976). In humans, evidence of segregation distortion has been detected for a genetic marker near the insulin gene (Eaves et al. 1999). Another example of deviation from Mendelian segregation would be if a marker were in LD with a gene that causes early fetal death. The segregation of such a marker will deviate from Mendelian segregation, so that the TDT can lead one to wrongly conclude that there is linkage disequilibrium of this marker with other traits. To avoid this problem, Mendelian segregation of markers can be tested on a random sample of offspring. If Mendelian segregation is questionable, the frequency of transmission of alleles from heterozygous parents can be compared between affected and unaffected siblings, because under the null hypothesis, the segregation of parental alleles does not depend on affection status, even if Mendelian segregation does not hold. This analysis can be conducted by creating a 2x2 table, with the rows representing affected and unaffected offspring, and the columns the transmitted alleles (Spielman et al. 1993), although care should be taken to perform an analysis stratified on

sibships whenever population heterogeneity is believed to exist (Schaid and Rowland 1998b).

Even when a marker segregates according to Mendelian probability, using information from both affected and unaffected children can sometimes increase the power over using only affected children. This is because when the penetrance of the underlying gene is relatively large, the marker allele that is frequently transmitted to affected offspring will be less frequently transmitted to unaffected offspring. Using this logic, a pooled *TDT* statistic that uses affected and unaffected children can be created by simply reversing the roles of which alleles are scored as transmitted and non-transmitted for the affected and unaffected children. For example, if a parent has alleles M_1 and M_2 and among affected offspring M_1 is scored as transmitted (M_2 not transmitted), then among unaffected offspring M_2 would be scored as transmitted (M_1 not transmitted). (Lazzeroni and Lange 1998). However, caution should be taken when using unaffected offspring, because if the penetrance is low, the transmission of alleles to unaffected offspring will not deviate much from Mendelian segregation. This can cause the unaffected children to contribute little to the numerator of the *TDT*, yet still contribute to the variation in the denominator of the *TDT*, leading to diminished power compared to using only affected children (Boehnke and Langefeld 1997).

Multiple Marker Alleles

When there are $K > 2$ alleles for a given marker locus, the *TDT* can be applied by comparing the transmission of each allele versus all others combined, resulting in K statistical comparisons. The largest of these K statistics, denoted *maxTDT*, can be used as a summary statistic. To avoid inflated Type-I error rates by multiple testing, the Bonferroni correction can be applied, which tends to be conservative, or empirical p-values can be computed (Schaid 1996; Morris et al. 1997). Empirical p-values are easy to compute: 1) create a new genotype for each affected child by randomly choosing, with equal probability, an allele from each of its parents; 2) recompute the statistic for this new dataset; and 3) repeat this process a large number of times to enumerate the distribution of the test statistic. The fraction of simulated statistics greater than or equal to the observed statistic is the empirical p-value. The advantage of using the *maxTDT* is its simplicity, and it can be more powerful than other methods when only one marker allele is in strong LD with a disease allele (Schaid 1996). But, when LD does not favor a single allele, this approach can lose information by grouping together alleles that are positively and negatively associated with disease.

Several alternative methods that consider all K alleles simultaneously are based on expanding the 2x2 table of transmitted/non-transmitted alleles to a $K \times K$ table that distinguishes the different alleles – see Table 2. When the analysis is restricted to heterozygous parents, the diagonal elements of Table 2, n_{ii} , are all zero; otherwise the statistical methods remove these non-informative contributions from homozygous parents. Under the null hypothesis, Table 2 is expected to be fully symmetric. That is, if a parent is heterozygous for alleles i and j , then the number of times that this parent

Table 2. Counts of transmitted/non-transmitted marker allele pairs for $2N$ parents of N affected children: K marker alleles, labeled as $M_1, M_2 \dots M_K$.

Transmitted Allele	Non-transmitted Allele				Row Total
	M_1	M_2	...	M_K	
M_1	n_{11}	n_{12}	...	n_{1K}	$n_{1\bullet}$
M_2	n_{21}	n_{22}	...	n_{2K}	$n_{2\bullet}$
...
M_K	n_{K1}	n_{K2}	...	n_{KK}	$n_{K\bullet}$
Column Total	$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet K}$	$n_{\bullet\bullet} = 2N$

transmits allele i should equal, on average, the number of times that this parent transmits allele j . This expectation can be written as $E[n_{ij}] = E[n_{ji}]$, where n_{ij} are the cell counts in Table 2, and $E[\]$ denotes large sample expectation. To test for complete symmetry, a chi-square statistic can be applied, $T_{c-sym} = \sum_{i < j} (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji})$, which has an

approximate chi-square distribution. The degrees of freedom are the number of types of heterozygous parents, (i.e., the number of terms for which $(n_{ij} + n_{ji}) > 0$), which can be as large as $K(K-1)/2$. However, this approach has weak power to detect important LD, because of the many degrees of freedom. Also, it may not be unusual for some of the n_{ij} values in Table 2 to be small, resulting in poor approximation of p-values by the chi-square distribution. Permutation methods can be used to circumvent the use of the chi-square distribution (Lazzeroni and Lange 1998), but the complete symmetry hypothesis is still likely to have weak power. As a rough guide, simulations suggest that the test of complete symmetry has little power when the number of marker alleles exceeds 5 (Cleves et al. 1997).

A more powerful approach than testing for complete symmetry is to test for marginal homogeneity, which states that the total number of times that an allele is transmitted is expected to equal the total number of times it is not transmitted. In terms of the row and column totals in Table 2, this implies that the differences in the marginal totals, $d_i = n_{i\bullet} - n_{\bullet i}$, are expected to be zero. However, these differences are correlated, which must be taken into account by computing the covariance matrix for the vector of differences. A chi-square statistic to test marginal homogeneity can be computed as $T_{m-sym} = d'V^{-1}d$, where d denotes the vector of differences, d' its transpose, V the covariance matrix of d , and V^{-1} a generalized inverse of the matrix V . A generalized inverse can be useful to evaluate the instability of the matrix V when attempting to compute its inverse, due to unexpected dependencies, and to determine the appropriate degrees of freedom; the degrees of freedom is no larger than $(K-1)$, but can be smaller, depending on the pattern of heterozygous parents (Schaid 1996). For example, if the data set is composed of only parental mating types $1/2 \times 3/4$, then even though there are 4 alleles, the degrees of freedom is 2 and not 3. This is because we can only compare the

preferential transmission of allele 1 versus 2, with one df, and the preferential transmission of allele 3 versus 4, with one df. Several authors have considered similar multivariate chi-square statistics, but differed regarding how the covariance matrix is computed: V can be computed in its most general form allowing the alternative hypothesis to be true, which is called Bhapkar's statistic (Jin et al. 1994), or by assuming marginal homogeneity to be true, which is called Stuart's statistic (Bickeböller and Clerget-Darpoux 1995). Stuart's statistic is more restrictive, requiring fewer terms to compute the covariance matrix V , making it more stable and probably more powerful. It can also be shown that Stuart's statistic is the score statistic for a genetic model in which the alleles act in a multiplicative fashion on the genotype odds ratios, which is appealing from the view of logistic regression modeling (Schaid 1996). But, because these methods require inversion of a matrix, the even more restrictive assumption that $n_{ij} + n_{ji}$ is constant for all values of i and j in Table 2 makes the inversion algebraically feasible, resulting in

$$T_{SE} = [(K-1) / K] \sum_i (n_{i\bullet} - n_{\bullet i})^2 / (n_{i\bullet} + n_{\bullet i} - 2n_{ii}),$$

as proposed by Spielman and Ewens (Spielman and Ewens 1996). This statistic, which removes the contribution from homozygous parents by the term $-2n_{ii}$, is analogous to adding the individual *TDT* statistics to create a total summary statistic, and is simple to compute. It generally has an approximate chi-square distribution with $(K-1)$ degrees of freedom, although there are situations where the degrees of freedom will actually be fewer (see above example), making the statistic conservative. One advantage of the more complicated Stuart's statistic is that the appropriate degrees of freedom can be computed by numerical methods that examine the linear dependencies inherent in the covariance matrix. But, for sparse data, the chi-square approximation can be invalid, with an inflated rate of false-positive findings when the number of marker alleles is small, yet conservative when the number of marker alleles is large. Alternatively, either simulated or exact methods to compute p-values tend to be more reliable for small sample sizes (Cleves et al. 1997).

Multiple affected siblings can be included in analyses simply by treating each affected offspring and its parents as independent trios, provided that the affected siblings are independent, given their marker genotypes and their parents' marker genotypes. This independence is true when there is no linkage, but it is not true when linkage exists. So, it is valid to include multiple affected siblings in tests for linkage when association exists, but it is not valid to include multiple affected siblings in tests for association when linkage exists. More discussion of these issues is provided in a later section regarding sibling controls.

Although the chi-square statistics presented in this section are simple to compute, other likelihood-based methods of assessing transmission/disequilibrium are worth discussing because of the powerful statistical tools and intuitive understanding they provide. Two different logistic regression models are presented in the next two sections - an allele transmission model and a genotype model with pseudo-sib controls.

Logistic Allele Transmission Model

The allele transmission model considers the pairing of each heterozygous parent with its affected child. If both parents are heterozygous, then each parent is paired with the affected child (i.e., two pairs). The two alleles of a heterozygous parent are considered a matched pair, and conditional logistic regression methods can be used to model the probability of their transmission. Let p_{ij} denote the probability that allele i is transmitted and allele j is not transmitted for a parent with genotype ij . The corresponding conditional logistic regression model considers which parental allele is transmitted, conditional on the parent's genotype, and can be written as

$$\log(p_{ij} / p_{ji}) = \alpha_i - \alpha_j + \beta_{ij}, \quad \text{for } i < j.$$

Because only one allele is transmitted, $p_{ij} + p_{ji} = 1$. The parameter α_i represents the log odds ratio for the main effect of allele i on the transmission probability. When $\alpha_i > 0$, the corresponding allele is preferentially transmitted, indicating a positive LD with disease. A strong negative association also indicates LD, but the marker allele may not be causative. Because these are odds ratio parameters, the number of α_i parameters that can be independently estimated is $(K-1)$. So, out of the total $K(K-1)/2$ degrees of freedom used to test complete symmetry, $(K-1)$ can be used to evaluate the main effects of alleles. The remaining degrees of freedom ($K^2 / 2 - 3K / 2 + 1$) can be used to evaluate the β_{ij} interaction parameters. The interaction parameter β_{ij} represents the situation when the odds that allele i is transmitted depends on which allele is not transmitted. Jin et al. (Jin et al. 1994) provide a more thorough discussion of this fully parameterized logistic regression model. Under the null hypothesis of complete symmetry, all α_i and β_{ij} parameters are zero, so that a likelihood ratio statistic can be constructed to test this hypothesis.

When a marker is highly polymorphic, with many alleles, the number of interaction terms is large, diminishing the power of the global likelihood ratio statistic. Alternatively, it can be assumed that there are no interactions, equivalent to assuming that the effects of alleles are multiplicative on the genotype odds ratios, and a likelihood ratio statistic can be used to test whether the main effects, α_i , are all zero. This restricted likelihood ratio statistic has an approximate chi-square distribution with $(K-1)$ df. Two methods to implement this restricted likelihood ratio statistic for marginal homogeneity have been proposed. One is based on the summary counts provided by Table 2, and called the *ETDT* (extended *TDT*, to $K > 2$) (Sham and Curtis 1995). Another equivalent method uses widely available software for conditional logistic regression for matched pairs (Harley et al. 1995). Each heterozygous parent and affected child contributes a matched pair of observations. The "case", with dependent variable $y=1$, is the transmitted allele, and the "control", with dependent variable $y=0$, is the non-transmitted allele. Then, covariates are created to compare particular types of alleles between cases and controls. For each case and control, the i^{th} allele is scored 1 or 0, according to whether it is present or absent. Because odds ratios are computed, one allele value is considered the "baseline". This coding will generate a covariate vector of length $(K-1)$ for each observation. An example of this coding is presented in Table 3 for an affected child with

genotype 1/4, maternal genotype 1/2, and paternal genotype 3/4. This example assumes that some allele value larger than 4 is used as the baseline. It is worthwhile to recognize that not only can this model be fit with software for conditional logistic regression, but also software for standard (unconditional) logistic regression. The trick to use standard logistic for matched pairs is to reduce each pair to a single observation by creating a new covariate from the covariate difference between cases and controls (e.g., $x = x_{case} - x_{control}$), creating the outcome $y=1$ for all pairs, and then fit the logistic model without an intercept. It is worthwhile to note that the score test from this logistic regression model is the same as Stuart's chi-square statistic to test marginal homogeneity. Hence, Stuart's statistic and the likelihood ratio statistic for the main effects of allele transmission are likely to have similar power, particularly for small values of α_i .

A strength of logistic regression is the ability to evaluate heterogeneity in allelic transmission to affected offspring. Heterogeneity can be caused by multiple factors, such as heterogeneity in risk associated with different marker alleles, parent-of-origin effects, a mixture of genetic and sporadic cases in the sample, or gene x environment interaction. To assess allelic heterogeneity, a step-wise selection process can be used to sequentially remove from the logistic model any marker alleles whose risk does not differ statistically from that of the baseline allele, and hence create a more parsimonious model. This can be helpful to group alleles with similar effects. Heterogeneity can also be caused by interactions, such that the strength of genetic evidence for particular alleles varies according to other covariates, such as age at disease onset, affection status of parents, strength of family history, ethnic background of the family, or exposure to environmental factors. To evaluate interactions, simply include an interaction term between the non-genetic covariate and the covariate representing the allele(s) of interest. For example, if A_i codes the presence/absence of a particular allele i among the transmitted and non-transmitted alleles, and E codes the presence/absence of an environmental exposure, then the logistic regression model that includes the interaction parameter β_{int} is

$$\log(p_{ij} / p_{ji}) = \alpha_i A_i + \beta_{int} (A_i E). \quad (1)$$

Table 3. Coding for logistic allele transmission modeling for a child with genotype 1/4, maternal genotype 1/2 and paternal genotype 3/4.

		Allele:			
		1	2	3	4
maternal contribution:	“case”	1	0	0	0
	“control”	0	1	0	0
paternal contribution:	“case”	0	0	0	1
	“control”	0	0	1	0

case = transmitted allele

control = non-transmitted allele

Harley et al. (Harley et al. 1995) used this approach to evaluate the differential effect of gender of the affected child on the transmission probability of particular alleles, as well as the differential effect of gender of the parent. They also considered the interaction of alleles from two marker loci by this approach. This method has also been used to evaluate interactions of genes with race, family history, maternal smoking and type of cleft for children with oral clefts (Maestri et al. 1997), with significant evidence of interaction between maternal smoking and markers near transforming growth factor genes. The availability of step-wise selection and model diagnostics available with most conditional logistic regression software packages makes this general approach of logistic transmission modeling a very powerful analytic method.

Although logistic allele transmission modeling offers a number of critical advantages, it does have several limitations. First, the main effects of covariates specific to the child cannot be estimated, because these factors are identical for each matched pair of parental alleles. In a sense, the case and control alleles are perfectly matched on all factors except the allele of interest. Second, because we are modeling transmitted alleles, and not genotypes, the logistic model in expression (1) requires the assumption of multiplicative effects of alleles, which can lead to dramatic odds-ratios for cases that are homozygous for the “high-risk” allele and exposed to an environmental factor. To see this, note that $r_i = e^{\alpha_i}$ is the odds ratio for cases that are unexposed to the environmental factor and have one copy of allele i . By the assumed multiplicative effects, unexposed homozygous i/i cases have an odds ratio of r_i^2 . The effect of exposure on allele i is $r_e = e^{\beta_{int}}$, so that exposed cases with one copy of allele i have an odds ratio of $r_i r_e$, and exposed homozygous cases have an odds ratio of $r_i^2 r_e^2$. This approach can lead to bias if the assumption of multiplicative effects is not true. Alternative odds ratio models that do not require the assumption of multiplicative allelic effects are discussed in the next section. Third, there is an implicit assumption that for each pairing of the diseased child and one of its parents, the environmental factor and the genotype of the child are independent, conditional on the genotype of the parent. Further discussion of this type of conditional independence is given in the next section.

Logistic Genotype Model with Pseudo-sib Controls

Some of the limitations of logistic allele transmission modeling can be overcome by considering the genotypes of both parents simultaneously, providing more general models of genotype odds ratios. This method takes advantage of the fact that when the disease is rare, the distribution of the genotypes of the cases depends on the genotype relative risks, as long as we condition on the genotypes of the parents. Using the results outlined by Self et al. (Self et al. 1991), let g_c be the marker genotype of a case, and g_m and g_f be those of the case's mother and father. The genotype of the case can be coded numerically using the vector Z , offering many options to model the genotype odds ratios. For example, odds ratios specific to each genotype can be modeled by the use of indicator variables in the vector Z that indicate which genotype the case possesses; dominant

effects can be modeled by indicating whether the case possesses either one or two copies of the allele of interest; recessive effects can be addressed by indicator variables for homozygous genotypes; multiplicative effects of alleles can be modeled by counting the number of alleles (e.g., 0, 1, or 2) of a particular type (Schaid 1996). See Table 4 for examples of these coded vectors.

Self et al. (Self et al. 1991) have shown that conditional on the genotypes of the parents, and given that offspring are sampled because they have disease, the contribution of a diseased case and its parents to the likelihood is

$$L = \frac{\exp(Z_i \gamma)}{\sum_{g_j \in G} \exp(Z_j \gamma)}, \quad (2)$$

where Z_i is the numerically coded genotype for the case, γ is a vector of log-odds ratios which approximate the log genotype relative risks, G is the set of the four possible offspring genotypes that the parents can produce, and g_j represents one of these four genotypes with its corresponding numerical code given by Z_j . For example, if M_1 and M_2 are the two alleles of the mother of the case, and F_1 and F_2 those of the father, then $G = \{M_1F_1, M_1F_2, M_2F_1, M_2F_2\}$. Note that the likelihood in equation (2) is the same form as a conditional logistic regression model such that each case is matched to three hypothetical siblings. This method has been referred to as "pseudo-sib" controls. Standard software for either conditional logistic regression or Poisson regression (Weinberg et al. 1998) can be used to fit the logistic genotype model.

The general likelihood with pseudo-sib controls simplifies when there are only two alleles at the marker, resulting in three genotypes and two odds ratio parameters. For this

Table 4. Examples of numerical coding of genotypes when there are 3 alleles, labeled 1, 2, and 3.

Genotype	Genotype indicators ^a	Dominant allele indicators ^b		Recessive allele indicators ^b		Allele counts ^b	
		2	3	2	3	2	3
<i>1/1</i>	0 0 0 0 0	0	0	0	0	0	0
<i>1/2</i>	1 0 0 0 0	1	0	0	0	1	0
<i>1/3</i>	0 1 0 0 0	0	1	0	0	0	1
<i>2/2</i>	0 0 1 0 0	1	0	1	0	2	0
<i>2/3</i>	0 0 0 1 0	1	1	0	0	1	1
<i>3/3</i>	0 0 0 0 1	0	1	0	1	0	2

a) Genotype *1/1* is used as baseline for odds-ratios

b) Allele *1* is used as baseline for odds-ratios

situation, hypotheses regarding the odds ratios can be tested with score tests (Schaid and Sommer 1993) and parameters can be estimated by maximum likelihood methods (Knapp et al. 1995); alternatively, simpler, but less efficient non-iterative methods can be used to estimate odds ratios (Khoury 1994; Flanders and Khoury 1996; Sun et al. 1998).

An advantage of this conditional likelihood framework is that the baseline risk can vary from family to family, or from case to case, yet the genotype odds ratios remain unbiased as long as the measured marker genotypes have a multiplicative effect on the baseline risk. An interpretation of this is that the affected case and its pseudo-sib controls are perfectly matched for all factors not related to the marker locus. Note that this logistic regression method compares the frequency of the genotypes of the cases relative to their expected Mendelian frequency; deviations from Mendelian expectations are captured by the genotype odds ratio parameters, γ . When there is no association between the marker genotypes and disease, i.e., $\gamma=0$, the offspring genotype frequencies are determined by Mendelian segregation frequencies. As typical with many applications of the logistic regression model, the odds ratios estimated by the conditional logistic genotype likelihood are good approximations of the genotype relative risks whenever the disease is rare. When the disease is not rare, the odds ratios are biased toward unity, because it is implicitly assumed that the pseudo-sib controls would not have been diseased. If the population baseline rate of disease is known, this information can be used to correct the likelihood to obtain unbiased estimates of genotype relative risks (Witte et al. 1999).

Although the perfect matching of the case and pseudo-sibs prevents assessment of the main effects of non-genetic covariates, it is possible to use the conditional logistic model to evaluate interaction of the marker genotypes with environmental covariates. To do so simply requires the inclusion of an interaction covariate (i.e., the product of the genetic marker covariate and the environmental covariate) in the logistic regression model. But, the validity of this method requires that the genotype and environmental covariate are independent, given the parents' genotypes. This independence of gene and environmental factors is similar to that required for the study of gene x environment interaction in the case-only study design (Khoury and Flanders 1996), although in our situation the assumed independence is less restrictive, because we require independence *conditional* on the parental genotypes. An excellent example (Thomas 2000) is the study of the interaction between the *BRCA1* susceptibility gene for breast cancer and oral contraceptive use. Because the decision to use oral contraceptives may depend on a woman's family history of breast cancer, and family history of breast cancer is associated with the *BRCA1* gene, family history can cause confounding by inducing a population association between the *BRCA1* carrier status and oral contraceptive use. Now, consider a family-based study in which cases are selected because they have breast cancer. If oral contraceptive use among cases depends on the family history of breast cancer among their relatives, then whether independence between *BRCA1* and oral contraceptive use is true depends on the type of relative (ascendant versus descendent). If the relatives are ascendants (e.g., mother, grand-mother, sister, aunt, niece), then by conditioning on the genotypes of the case's parents, the case's genotype and the genotypes of the ascendant relatives are independent, which then causes independence between the case's genotype and oral contraceptive use. Hence, for this example it would be valid to use logistic

regression to evaluate interaction. In contrast, if family history is based on descendents, such as daughters of the case, then the case's genotype and oral contraceptive use are not independent, making logistic regression invalid. The dependence is induced through a series of dependencies: the genotype of the case influences the genotype of the daughter, which influences the disease status of the daughter, which influences the mother's decision to use oral contraceptives. This emphasizes the need for careful evaluation of the assumption of conditional independence between genotype and environmental exposures.

Comparison of Logistic Models: Allele Transmission versus Genotype

Up to now we have discussed two logistic models: 1) the logistic allele transmission model with case and control defined as the transmitted and non-transmitted alleles, respectively, and 2) the logistic genotype model that considers genotype odds ratios and uses pseudo-sib controls. These two methods condition on different information, causing subtle differences in their statistical properties. If the maternal and paternal genotypes are associated with each other in the general population, which can occur when there is population stratification or assortative mating, the transmission of maternal and paternal alleles are not independent of each other. But, conditioning on the joint genotypes of the parents causes the transmission of parental alleles to be independent (Lazzeroni and Lange 1998). This is the advantage of the logistic genotype model with pseudo-sib controls. In contrast, the logistic allele transmission model requires the transmission of maternal and paternal alleles to be independent, which may not be true. If independence fails, then the variance estimates of the regression parameters will be invalid, leading to incorrect confidence intervals.

However, the logistic allele transmission model is a special case of the logistic genotype model whenever the effects of marker alleles on the genotype odds ratio are multiplicative. To see this, consider a parent with marker genotype ij , a second parent with marker genotype kl , and an affected child with genotype ik . For the genotype odds ratio model, let γ_{ik} be the log odds-ratio for the genotype ik . Then, the likelihood contribution for this trio is

$$L = \frac{e^{\gamma_{ik}}}{e^{\gamma_{ik}} + e^{\gamma_{il}} + e^{\gamma_{jk}} + e^{\gamma_{jl}}}. \quad (3)$$

Now, if allelic effects are multiplicative (i.e., log-additive), then $\gamma_{ik} = \alpha_i + \alpha_k$, where α_i represents the main effect of allele i . After substituting this additive model into expression (3), the likelihood factors into the allelic transmission contributions from each parent,

$$L = \left(\frac{e^{\alpha_i}}{e^{\alpha_i} + e^{\alpha_j}} \right) \left(\frac{e^{\alpha_k}}{e^{\alpha_k} + e^{\alpha_l}} \right) = L_1 L_2,$$

where L_1 and L_2 are the likelihood contributions to the logistic allele transmission model from the first and second parents, respectively. This factorization of the likelihood

emphasizes that the transmission of maternal and paternal marker alleles to an affected child are independent of each other whenever the effects are multiplicative on the genotype odds ratios. A special case occurs under the null hypothesis (i.e., either $\theta = .05$ or $\delta = 0$), which emphasizes that it is valid to treat maternal and paternal allelic transmissions as independent when performing transmission/disequilibrium testing (Whittaker et al. 1997). But, when linkage ($\theta < .05$) and association both exist, parental contributions are not independent unless the effects of the marker alleles are multiplicative. This multiplicative assumption can be tested by including interaction parameters in the logistic genotype odds ratio model. However, for highly polymorphic markers, there can be many genotypes, leading to many pair-wise interaction terms, and a large number of degrees of freedom to test the multiplicative assumption, and consequently weak power.

A potential disadvantage of many of the transmission disequilibrium methods, including the logistic allele transmission model, is that they treat heterozygous parents as independent. If a child and both of its parents have the same heterozygous genotype, then methods that consider parents independently can lose power, because the transmission of a marker allele from one parent cancels the non-transmission of the same allele from the other parent. In contrast, the logistic genotype model with pseudo-sib controls offers more flexible modeling of genetic effects by considering the joint transmission of both parental alleles, and any deviation of the offspring genotype counts from Mendelian expectation influences the genotype odds ratios. When there are only two alleles at the marker locus, say alleles M_1 and M_2 , there are three genotypes M_1/M_1 , M_1/M_2 , and M_2/M_2 . If allele M_1 is considered the high-risk allele, and genotype M_2/M_2 the baseline, there are two genotype odds ratios, $r_{1,1}$ and $r_{1,2}$. If it can be assumed that these two odds ratios depend on a single parameter that captures the effect of the high-risk allele, then a likelihood ratio statistic with 1 df can be used to test the null hypothesis. Some examples of such odds ratio models, with the dependent parameter denoted γ , are dominant ($r_{1,1} = r_{1,2} = \gamma$), recessive ($r_{1,1} = \gamma$; $r_{1,2} = 1$), or multiplicative ($r_{1,1} = \gamma^2$; $r_{1,2} = \gamma$). Alternatively, the two odds ratios can be estimated without restrictions, leading to a likelihood ratio statistic with 2 df. Simulations have shown that this unrestricted likelihood ratio statistic is fairly robust for different genetic mechanisms, and can offer greater power than the *TDT*, particularly for recessive effects (Schaid 1998; Weinberg et al. 1998). However, this general approach loses its appeal when evaluating a highly polymorphic marker, due to the large number of possible genotypes, and hence large number of degrees of freedom. If there exists *a priori* knowledge of the likely genetic mode of inheritance of the disease, then this information can be used to create a more powerful statistical test. For example, when the true mode of inheritance is recessive, and the marker genotypes are coded appropriately to allow for this, the score statistic from the logistic genotype model can be substantially more powerful than the various "*TDT* methods" that consider the independent transmission of parental alleles (Schaid 1996). But, without this prior knowledge, it seems most reasonable to first evaluate the transmission of alleles, either by logistic allelic transmission modeling, or by chi-square statistics for marginal homogeneity. Then, after statistically significant alleles are identified, more refined models using logistic genotype regression models with

pseudo-sib controls can be used to further evaluate the influence of the various alleles, and their interaction with non-genetic factors, on the genotype odds ratios.

In summary, an advantage of the allele transmission model is that it can be used even when one parent is missing (provided bias is avoided - see section below on Missing Parents), whereas the logistic genotype model requires both parents. But, the allele transmission model can fit only multiplicative allele effects, causing discard of triads that have the same heterozygote genotype. The logistic genotype model is more flexible, allowing for genotype odds ratio models that differ from multiplicative effects. A few additional points to consider for logistic regression modeling are that sparse data may require exact logistic regression methods, and when evaluating many marker alleles by stepwise methods, forward selection may be more reliable than backward selection because of dependencies among alleles.

Relationship of Marker and Susceptibility Genotype Odds Ratios

Although we have focused on the effects of the marker locus, independence of parental contributions can also be stated in terms of the underlying susceptibility locus. Assuming Hardy-Weinberg equilibrium, random mating, and equal recombination rates for males and females, independence of parental transmissions occurs when the penetrance of the susceptibility genotype is multiplicative – that is, when the penetrance for genotype ij has the form $f_{ij} = \sqrt{f_{ii}} \sqrt{f_{jj}}$ (Knapp et al. 1993; Bickeböllner and Clerget-Darpoux 1995; Whittaker et al. 1997). A special case is a recessive disease without phenocopies (Ott 1989). The relationship between the marker genotype odds ratio and the susceptibility genotype odds ratio is complex when the marker is not the causative gene. The marker genotype odds ratios depend on the frequencies of the parents' two-locus haplotypes, where the haplotypes are composed of the marker and susceptibility alleles. These haplotype frequencies in turn depend on the amount of LD and the frequency of the susceptibility allele. The marker genotype odds ratios also depend on the probability that recombination occurs between the two loci, and on the penetrance of the underlying susceptibility genotypes. It can be shown that the marker genotype odds ratios are closer to unity than the susceptibility genotype odds ratios whenever recombination between the marker and susceptibility loci occurs, or when LD is not complete (Schaid 1996). By complete LD, we mean that a susceptibility allele and marker allele only occur together on the same haplotype.

Imprinting: Modeling Parent-of-origin Effects

There is a large body of evidence that the expression of some genes depends on the parent from whom they were inherited, so that the disease risk to a child depends not only on the child's genotype, but also on whether the inherited allele originated from the mother or father (Hall 1990). This type of deviation from Mendelian inheritance has been called imprinting. Although imprinting may affect an individual's phenotype, the

process does not appear to cause a permanent alteration in the DNA, but rather a modification of its expression. Methylation may be one of the molecular mechanisms involved in causing parent-of-origin effects, but other still unknown mechanisms are likely to be involved (Hall 1990).

The logistic genotype model can account for parent-of-origin effects by including covariates in the logistic regression model that indicate whether the alleles were transmitted from the mother or father. This can be accomplished by first choosing a method to code the child’s genotype (see examples in Table 4). Then, for alleles that are not considered the baseline allele, covariates can be created to indicate whether the alleles originated in the mother or the father. Extending the example given in Table 4 for which genotype indicators were used, we present in Table 5 the possible coded vectors when the parent-of-origin is included in the logistic model. Note that the child’s genotype is ordered according to the maternal/paternal origin of alleles. This does not affect the numerical coding of the child’s genotype, but it does affect the coding of the indicator variables for the maternal and paternal origin. Using this setup, conditional logistic regression with pseudo-sib controls can be used. One limitation is when the origin of parental alleles cannot be assigned, which occurs when the parents and the child all have the same heterozygous genotype. In this case, we can view the parental origin of alleles as missing data, and use the expectation-maximization (EM) algorithm to estimate the parameters (Weinberg et al. 1998). Standard conditional logistic regression software can be adapted to implement this EM approach (Horton and Laird 1998). Caution is warranted when recombination can occur between the marker and susceptibility loci (i.e.,

Table 5. Examples of numerical coding of a child’s genotype to account for parent-of-origin effects.

Genotype ordered by origin (maternal/paternal)	Genotype indicators ^a	Maternal Origin Indicators ^b		Paternal Origin Indicators ^b	
		2	3	2	3
<i>1/1</i>	0 0 0 0 0	0	0	0	0
<i>1/2</i>	1 0 0 0 0	0	0	1	0
<i>2/1</i>	1 0 0 0 0	1	0	0	0
<i>1/3</i>	0 1 0 0 0	0	0	0	1
<i>3/1</i>	0 1 0 0 0	0	1	0	0
<i>2/2</i>	0 0 1 0 0	1	0	1	0
<i>2/3</i>	0 0 0 1 0	1	0	0	1
<i>3/2</i>	0 0 0 1 0	0	1	1	0
<i>3/3</i>	0 0 0 0 1	0	1	0	1

a) Genotype *1/1* is used as baseline for child’s genotype odds-ratios

b) Allele *1* is used as baseline for parent-of-origin odds-ratios

the marker is not causative, but rather in LD with the underlying susceptibility locus). When recombination is possible, the transmission of parental alleles are not statistically independent between two heterozygous parents unless the marker allele effects are multiplicative on the odds ratio. This dependence can invalidate the likelihood ratio statistic when testing for parent-of-origin effects, in the presence of known transmission disequilibrium. More refined methods for testing parent-of-origin effects can be found elsewhere (Weinberg 1999b).

Parental Genotype Effects

For some disorders, the parent's genotype may influence the risk to the child, regardless of which alleles were transmitted from the parents to the child. This can occur, for example, when the mother's genotype influences the risk of birth defects for the fetus, due to the influence of the maternal genotype on the environment of the fetus. It has been speculated that the risk of neural tube defects in a child could be due to the mother's genotype (van der Put et al. 1996). One way to evaluate the influence of maternal genotype is to obtain genotype information from her parents, in order to evaluate whether the transmission of grandparental alleles to the mother deviate from Mendelian transmission (e.g., apply the *TDT* to the triad of mother and her parents) (Mitchell 1997). Although this approach requires only that the marker be transmitted in Mendelian fashion, it can be difficult to obtain samples from the grandparents. An alternative approach is to evaluate the relative frequency of all three genotypes of the case and its parents. If a mother's genotype influences the disease risk to her child, but the father's genotype and the child's genotype do not influence the child's risk, then we would expect to observe parental matings such that the mother is more likely to have the high-risk genotype than the father, yet the transmission of alleles from parents to child are random. If there were no maternal effect, the maternal and paternal genotype frequencies would be symmetric within each mating type. Although this approach uses information *between* families by considering the relative frequency of the different triads, and the logistic genotype model uses information *within* families by considering parental transmission of alleles, the two approaches can be combined by using Poisson regression to simultaneously estimate the effects of the mother's genotype, the child's genotype, and parent-of-origin of alleles (Weinberg et al. 1998; Wilcox et al. 1998). Note that the information for parental genotype effects comes from the asymmetric frequencies of maternal and paternal genotypes within the different mating type strata. It should be recognized that any factors that cause asymmetry of parental genotypes within the mating type strata will be attributed to the influence of parental genotypes on the child's disease risk, even if this is not true. For example, if a person's survival to parenthood depends on both gender and their genotype, then this can cause unequal maternal and paternal genotype frequencies within the different parental mating type strata. So too can mate selection that depends on particular combinations of genotypes, such as when mothers that come from a population with a high frequency of a particular genotype tend to marry fathers that come from a different population with a low frequency of the same genotype. Hence, the application of these methods requires attention to possible biases.

Missing Parents

The greatest limitation of transmission disequilibrium methods is the need for parental genotypes. A number of methods have been developed to account for missing parents, such as analyzing only the sub-sample of complete data, reconstructing the missing parental genotype(s), using “informative” parent-child pairs, enumeration of all possible genotypes for the missing parents along with their corresponding relative probabilities, or even using unaffected siblings as controls - in place of the parental information. Each of these approaches will be discussed. But, it should be recognized that without parents, there is often a loss of information, and more assumptions are required in order to avoid bias.

Excluding cases when one or both parents are missing results in loss of information, although the statistical test remains valid. This exclusion can also result in a set of cases that are not representative of the total sample, particularly if the reason parents are missing is related to their genotypes (Witte et al. 1999). To prevent information loss, it is tempting to include families with a single observed parent whenever the parent-child pair appears “informative” for the *TDT*. However, there is a potential bias that must be avoided. To illustrate this bias, consider a marker with two alleles, *A* and *B*, and suppose that the observed parent has marker genotype *AB*. The affected child can have either the *AA* or *AB* genotype. When the affected child has the *AB* genotype, it is not clear whether the *A* or *B* allele was transmitted from the observed parent, so this parent-child pair is uninformative. When the affected child has the *AA* genotype, the observed parent must have transmitted the *A* allele and not the *B* allele, giving the impression that this “informative” pair can be used in analyses. But, if the *B* allele is rare in the population, the missing parent most likely has the *AA* genotype. For this situation, an observed *AB* parent transmits the *B* allele to a case with the *AB* genotype (declared uninformative), and an *A* allele to a case with the *AA* genotype. If there is no association, these two transmission scenarios are equally likely, yet discarding uninformative pairs will selectively remove meioses for which allele *B* was transmitted, making it appear that allele *A* is preferentially transmitted, and hence biasing the *TDT*. To avoid this kind of bias, the only type of child-single-parent pairs that should be used in analyses are those for which the parent and the child are both heterozygotes, yet of different genotypes (Curtis and Sham 1995). Application of this rule for markers with two alleles indicates that all cases with one observed parent should be excluded.

When one or both parents are frequently missing, alternative methods are needed to avoid discarding much of the data. One approach is to reconstruct the missing parental genotype(s) using only unaffected siblings, and then apply the usual transmission disequilibrium methods to the affected offspring. When one parent is missing, its genotype can be reconstructed whenever there are two alleles among the offspring that did not originate in the observed parent. When both parents are missing, both parental genotypes can be reconstructed whenever four alleles are observed among their offspring such that these four alleles are not identical-by-descent (IBD). For example, although offspring with *AA* and *BB* genotypes represent 2 alleles similar in state, they actually

represent 4 alleles that are not IBD, implying that both parents must have the *AB* genotype. Note, however, that if affected offspring are used to reconstruct the missing parental genotypes, and then these parental genotypes are scored for transmission to the same affected offspring, then this circular algorithm does not have the same statistical properties as when the parents are actually observed. To see this, consider the example with both parents missing and the offspring with genotypes *AA* and *BB*. Given these two offspring genotypes, both parents must have the *AB* genotype. Under the null hypothesis, these two offspring genotypes are equally likely, which implies that the mean and variance of the number of *A* alleles in the affected offspring are both 1 [the mathematical proof is provided as follows: mean = $(2_{AA}) \cdot .5 + (0_{BB}) \cdot .5$; variance = $(2_{AA} - 1)^2 \cdot .5 + (0_{BB} - 1)^2 \cdot .5$; where the subscript indicates the genotype which is counted for *A* alleles, and the fractions are probabilities of the different genotypes]. In contrast, when both parents are known to be *AB*, we do not need to condition on the types of offspring that they have - because we do not need to use offspring to reconstruct the parents. In this case, the mean of the number of *A* alleles is the same as when parents were reconstructed, but the variance is half as large [mathematical proof: mean = $(2_{AA}) \cdot .25 + (1_{AB}) \cdot .5 + (0_{BB}) \cdot .25 = 1$; variance = $(2_{AA} - 1)^2 \cdot .25 + (1_{AB} - 1)^2 \cdot .5 + (0_{BB} - 1)^2 \cdot .25 = .5$]. If parental reconstruction were ignored, then the Type-I error of the statistical test would be inflated for this example. Knapp et al. (Knapp 1999) provide further details on how to correct the mean and variance of the statistical methods when parental genotypes are reconstructed from their affected offspring, as well as how to combine these corrected statistics with both the usual *TDT* when both parents are available and the *S-TDT* statistic (discussed below) for when sib-controls are used because parental genotypes could not be reconstructed.

The above approach to parental reconstruction requires that there is only one genotype per parent consistent with the offspring's genotypes, otherwise there would be ambiguity regarding transmission of alleles. An alternative method that accounts for this ambiguity is to consider all possible genotypes of the missing parent(s), given the information available for the nuclear families, and then perform some kind of averaging over the missing data in order to estimate parameters and test hypotheses. The available information to impute the missing genotypes might include the genotypes of an observed parent, the affected child, and perhaps additional offspring. Although it may not be difficult to enumerate all possible genotypes for missing parents (e.g., using genotype elimination algorithms (Lange and Goradia 1987)), it can be difficult to reliably estimate the relative probabilities of these enumerated genotypes. One approach is to assume a homogeneous randomly mating population in Hardy-Weinberg equilibrium, and use marker allele frequencies and Mendelian segregation probabilities to compute the probabilities for the enumerated parental genotypes (Martin et al. 1998). But, several concerns arise. First, this method provides a valid test of association, but not linkage, because it is sensitive to any cause of association, including population stratification. Recall that population stratification is often the prime motivator to consider parental controls in the first place! Although this approach can be used for families that have already demonstrated linkage in order to evaluate whether association exists as well, there is still a problem because siblings are not independent when linkage exists. Second, the analyses are now dependent on allele frequencies, which leads to biased tests and

parameter estimates if these frequencies are misspecified (e.g., underestimation of allele frequencies can lead to falsely inflated information on LD). Third, the above method to compute the probabilities of the missing parental genotypes is based on the assumption of no association of the marker alleles with disease, which can diminish power. Although statistical methods can account for these latter two difficulties by simultaneous estimation of both allele frequencies and relative risk parameters (Schaid and Li 1997), the assumption of a homogeneous population remains a problem.

To avoid the tenuous assumption of a homogeneous random mating population, an EM algorithm can be used. If there were no missing parents, the likelihood for the observed data can be maximized to jointly estimate the parental mating type probabilities and the genotype odds ratios, yet without requiring Hardy-Weinberg genotype proportions and random mating. When some parents are missing, the two-step iterative EM procedure begins with a set of initial values for the parameters, and uses them to fractionate the observed data according to the expected counts for the missing data; these expected counts are then used to compute updated estimates of the parental mating type probabilities and the genotype odds ratios. Iteration continues until convergence. To see how this works, suppose that there are two alleles A and B . Without considering order of the parents, there are 6 mating types ($AA \times AA$, $AA \times AB$, $AA \times BB$, $AB \times AB$, $AB \times BB$, and $BB \times BB$); denote the probability of the i^{th} mating type as p_i . If a mother and child both have genotype AA , yet the father is missing, then the father can have genotype AA or AB , implying two possible mating types: $AA \times AA$ and $AA \times AB$. Since we know that there are only these two possibilities, the probability of the first mating type is $p_1 / (p_1 + p_2)$, and the probability of the second mating type is $p_2 / (p_1 + p_2)$. These mating type probabilities are used to fractionate the observed mother-child pair into two components, which are then used for the EM procedure. This approach can be implemented by use of a log-linear model that includes a parameter for each parental mating type, and captures much of the missing information when there are only two marker alleles (Weinberg 1999a). When there are more than two marker alleles, the number of possible parental mating types increases dramatically - for K alleles, the number of mating types is $[K^2(K+1)^2 + 2K(K+1)] / 8$. This can cause sparse data, making it difficult to reliably estimate all the parameters representing the mating type probabilities. Hence, it may be necessary to compare each allele versus all others by performing a series of bi-allelic analyses. The advantage of the EM approach is that it uses all the data, and could even include cases without both parents, if these are not frequent. But, the EM approach does require specialized software.

The most general method to test for linkage in the presence of association that can account for a mixture of types of families (trios, nuclear families, sibships, etc.), some of which may have missing genetic data among the parents, is provided by a method called Family Based Association Tests (FBAT) (Rabinowitz and Laird 1999; Laird et al. 2000). This approach evaluates the correlation of the trait and genetic marker for an offspring, but computes this correlation by conditioning on appropriate information within each family to avoid biases caused by population stratification.

It is worth noting that a non-iterative approach to estimating the relative risk parameters and their variances has been proposed, which also accounts for when a single parent is missing (Sun et al. 1998). This differs from the *TDT* method, in two ways: 1) two relative risks are estimated for a bi-allelic marker (relative risks for heterozygous and homozygous carriers), and 2) observed parents that are homozygous are included when the other parent is missing, which allows for the possibility that the missing parent could be heterozygous. This approach does not require Hardy-Wienberg equilibrium nor random mating. Although this method offers non-iterative simple formulas, it is not as efficient as likelihood methods.

Sib Controls

For diseases that initiate at older ages, it is nearly impossible to obtain parental genotypes, and a large amount of work is required to obtain a sufficient number of offspring to reconstruct the missing parental genotypes. For example, in order to have at least a 50% chance to reconstruct both parents' genotypes (i.e., determine whether they are heterozygous carriers for a particular allele of interest) at least 5 offspring are needed when both parents have the same heterozygous genotype, at least 4 offspring are needed when both parents are different heterozygotes but share one allele in common, and at least 3 offspring are needed when both parents are different heterozygotes and share no alleles in common (using results in Table 1 from (Knapp 1999)). A reasonable alternative to using parental genotype data is to use unaffected siblings as controls. Note that the use of true sib controls differs from the concept of pseudo-sib controls discussed above. Pseudo-sib controls are a theoretical construct that represent Mendelian expectations based on measured parental genotypes, and hence there is no random variation for the pseudo-sib control genotypes. Rather, the random variation is associated with the case's genotype. In contrast, true sib controls have two sources of variation. First, the genotypes of sib controls represent a realization of a random process – the random segregation of alleles from their parents. Second, because parents are missing, there is additional variation due to the different random genotypes the parents could possess, which are consistent with their offspring's genotypes. This second source of variation causes correlation of alleles among siblings, even under the null hypothesis. But, this correlation can be taken into account by a matched analysis that compares the allele frequencies between affected offspring and their matched unaffected siblings, as discussed in the next section.

Some important considerations when choosing sib controls are: 1) older sib controls may be better than younger, because younger sib controls may not remain free of disease until the age at which the case was diagnosed; 2) secular trends may confound older siblings, because siblings cannot be matched on birth dates; 3) recall bias may occur whenever the time interval from exposure to interview differs among siblings (Witte et al. 1999). Alternatively, cousins could be selected as controls, which may make it easier to match closer on birth date and gender. But, cousins do not provide absolute protection against the effects of population stratification, because cousins each have one parent that did not descend from a common ancestor. Furthermore, as in any case-control study design,