Am. J. Hum. Genet. 66:576–592, 2000

Removing the Sampling Restrictions from Family-Based Tests of Association for a Quantitative-Trait Locus

S. A. Monks[1] and N. L. Kaplan[2]

[1]Department of Biostatistics, University of Washington, Seattle, and [2]Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

Summary

One strategy for localization of a quantitative-trait locus (QTL) is to test whether the distribution of a quantitative trait depends on the number of copies of a specific genetic-marker allele that an individual possesses. This approach tests for association between alleles at the marker and the QTL, and it assumes that association is a consequence of the marker being physically close to the QTL. However, problems can occur when data are not from a homogeneous population, since associations can arise irrespective of a genetic marker being in physical proximity to the QTL—that is, no information is gained regarding localization. Methods to address this problem have recently been proposed. These proposed methods use family data for indirect stratification of a population, thereby removing the effect of associations that are due to unknown population substructure. They are, however, restricted in terms of the number of children per family that can be used in the analysis. Here we introduce tests that can be used on family data with parent and child genotypes, with child genotypes only, or with a combination of these types of families, without size restrictions. Furthermore, equations that allow one to determine the sample size needed to achieve desired power are derived. By means of simulation, we demonstrate that the existing tests have an elevated false-positive rate when the size restrictions are not followed and that a good deal of information is lost as a result of adherence to the size restrictions. Finally, we introduce permutation procedures that are recommended for small samples but that can also be used for extensions of the tests to multiallelic markers and to the simultaneous use of more than one marker.

Received March 11, 1999; accepted for publication November 23, 1999; electronically published February 16, 2000.

Address for correspondence and reprints: Dr. S. A. Monks, Department of Biostatistics, University of Washington, P. O. Box 357232, Seattle, WA 98195-7232. E-mail: steph@biostat.washington.edu

0002-9297/2000/6602-0026$02.00

Introduction

The transmission/disequilibrium test (TDT) introduced by Spielman et al. (1993) has become a popular family-based test of linkage between a marker and a susceptibility locus. The attractiveness of the TDT is a result of both its validity in structured populations and its power, which can be much greater than that of conventional linkage tests. The TDT has been extended to multiallelic markers (Bickeboller and Darpoux 1995; Sham and Curtis 1995; Schaid 1996; Spielman and Ewens 1996), to families without parental genotype information (Curtis 1997; Boehnke and Langefeld 1998; Monks et al. 1998; Schaid and Rowland 1998; Spielman and Ewens 1998), and to quantitative traits (Allison 1997; Rabinowitz 1997; Schaid and Rowland 1999). Although the TDT and its extensions that use parental information were designed as tests of linkage in the presence of association, they are also tests of association in the presence of linkage, for samples of unrelated parent/child trios. The advantage of the TDT, in this context, is that it is not sensitive to population stratification in the parental population, which can be a problem for the usual case-control test. For extensions to families without parental-genotype information, the TDT is a valid test of association in the presence of linkage only if samples contain unrelated sibships with exactly two children (one that is affected and one that is unaffected, for tests involving a susceptibility locus). If samples contain larger sibships, then these tests are valid only as tests of linkage.

Tests of association are often used for a candidate gene. Also, once a chromosomal region has been designated, through use of a linkage test, as being of interest, association tests done with the use of markers in the region may be useful for further localization of the susceptibility locus or QTL, since association is thought to exist in human populations for small distances—typically, <2 cM. When the TDT and its extensions are used to test for association, data sets must contain unrelated families of minimal size (one child for families with parental genetic information and two children for families without this information). If families with arbitrary numbers of children have been sampled, then strategies

must be used to reduce the data set. One is always reluctant to discard data, because of the probable loss of power; however, if methods are not used to adjust for the correlation between siblings that results from the marker and the QTL being linked, then the false-positive rate for tests of association will be unknown and will be larger than expected.

Martin et al. (1997) generalized the TDT as a test of association in the presence of linkage for a susceptibility gene, for families with an arbitrary number of affected children and with available parental marker–genotype information. More recently, Horvath and Laird (1998), using sibships of arbitrary size, developed a test of association for a susceptibility gene when parental data are missing. In both cases, the authors used the family as the independent sampling unit, to avoid the elevated false-positive rate caused by correlation between sibs. In the present study, we extend the ideas of Martin et al. (1997) and Horvath and Laird (1998), and we propose three tests that can be used for quantitative-trait data and that use information from all children. These tests are valid tests of linkage and association, regardless of the number of children sampled. Throughout the present study, we assume that linkage is present and focus on the test for association in the presence of linkage. For the T_{QP} test, we use genotype information for parents and for all of their children, whereas, for the T_{QS} test, we use genotypes for all siblings (with no parental information). Finally, for the T_{QPS} test, we use a combination of these types of family information. Similar to its treatment in the study by Martin et al. (1997) and Horvath and Laird (1998), the family is treated as the independent sampling unit in all three tests.

In the Methods section below, we introduce each of the three test statistics and derive their distributional properties. Specifically, we show that the test statistics are asymptotically standard normal under the null hypothesis of no linkage or no association. We derive their distribution on the basis of the alternative hypothesis of linkage and association, and, assuming that there is Hardy-Weinberg equilibrium at the marker and at the QTL, we provide formulas to be used for sample-size calculations. Additionally, permutation procedures are given that are recommended for small samples but that can also be used for extensions of the three tests to multiallelic markers and to the simultaneous use of more than one marker.

We then compare the T_{QP} and T_{QS} tests with two other nonparametric tests. The first test, which was introduced by Rabinowitz (1997), uses parental-transmission information. The second test, which was developed by Allison et al. (1999), is a permutation test for which only sibship information is used. We will denote these tests as T_R and T_A, respectively. Through simulation, we demonstrate that the false-positive rate increases for both of these tests, when nonminimal families are used. We also compare the power of our tests, in which information from all children is used, to the power of the T_R and T_A tests, in which data sets composed of families of minimal size are used. These comparisons demonstrate the validity of our tests as well as what is gained by use of this additional information. We then provide evidence that our permutation procedures for T_{QP} and T_{QS} are valid. Finally, we demonstrate that the validity of the T_{QP} and T_{QS} tests is not affected by population stratification.

Methods

Notation

Consider a diallelic-marker locus with alleles A_1 and A_2, with population frequencies p_1 and p_2, and a diallelic QTL with alleles Q_1 and Q_2, with population frequencies q_1 and q_2. Let θ denote the recombination fraction between the marker locus and the QTL. Association—also referred to as linkage disequilibrium—is measured with the use of the disequilibrium coefficient for A_1 and Q_1, denoted as D, where $D = Pr(A_1Q_1) - p_1q_1$ (Weir 1996). Define μ_{rs} as the trait mean for individuals with QTL genotype Q_rQ_s, and define σ_E^2 as the trait-distribution variance within one QTL genotype class. The parameter σ_E^2 represents all phenotypic variation not attributable to the QTL. We will use the parameterization in which the trait mean of the homozygotes is centered at zero, so that $\mu_{11} = a$, $\mu_{12} = d$, and $\mu_{22} = -a$, where $a > 0$. The phenotypic variance resulting from the QTL can be written as $\sigma_G^2 = 2q_1q_2[a + d(q_2 - q_1)] + (2q_1q_2d)^2$. As a measure of the amount of variation caused by the QTL, we use broad-sense heritability, denoted as H^2, which is the proportion of the phenotypic variance caused by the QTL—that is, $H^2 = \sigma_G^2/(\sigma_G^2 + \sigma_E^2)$ (Falconer and Mackay 1996).

Suppose that there are F families indexed by i. Let the t_i children in the ith family be indexed by j. Unless otherwise noted, all families have the same number of children—that is, $t_i = t$ for all families. Let Y_{ij} denote the trait value for the jth child in the ith family, and let $\bar{Y}$ denote the mean over all children in all families, as computed by $\bar{Y} = \frac{1}{F}\Sigma_{i=1}^{F}\frac{1}{t_i}\Sigma_{j=1}^{t_i}Y_{ij}$. Let X_{iM}^* (X_{iF}^*) = 1 if the mother (father) is heterozygous at the marker locus, and let X_{iM}^* (X_{iF}^*) = 0 otherwise; denote $X_{iM}^* + X_{iF}^*$ as h_i. Let X_{ijM} (X_{ijF}) indicate whether marker allele A_1 was transmitted to the jth child by the mother (father), and define $\bar{X}_i$ as the mean of the t_i values of $X_{ijM} + X_{ijF}$. See figure 1 for an example of notation.

The null hypothesis is H_o:*no linkage or no association,* whereas the alternative hypothesis is H_a:*linkage and association*. The part of the hypotheses concerning linkage is straightforward; however, the part concerning asso-

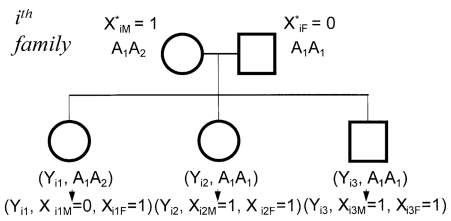

Figure 1 Example of notation for a family with three children

ciation requires further detail, because of the possibility of population stratification. If there is no population stratification, then the null hypothesis is that there is no linkage or association in the parental population. Alternatively, if there is population stratification, then the null hypothesis is that there is no linkage or association in any of the subpopulations from which parental chromosomes might originate. For clarity of presentation, it is assumed, throughout the present study, that the parental population is homogeneous; however, computations can be extended to a more-complicated population structure. An example is given in Appendix A. For a more-detailed discussion, the reader is referred elsewhere (Ewens and Spielman 1995). When the hypotheses are expressed in terms of parameters, we have $H_o : \theta = .5$ or $D = 0$ versus $H_a : \theta < .5$ and $D \neq 0$. In the present study, we assume that the marker and the QTL are in tight linkage and that only the test of association is of interest. For our theoretical derivations and simulations, we assume that the marker and the QTL are both in Hardy-Weinberg equilibrium. This assumption is made for computational convenience and is required only for power calculations.

Test Statistics

We begin with the case for which parental-genotype information is available. If there is no association between the marker allele A_1 and the QTL allele Q_1, then what is transmitted, at the marker locus, to a child from a marker-heterozygous parent neither affects that child's quantitative trait nor is related to anything that affects the quantitative trait, regardless of whether the marker and QTL are linked. One measure of the marker/QTL relationship is the covariance between the quantitative trait and a variable representing transmissions at the marker locus (Rabinowitz 1997). Consider a family with one parent that is heterozygous at the marker locus—that is, $h_i = 1$. Without loss of generality, suppose that the mother is heterozygous. For each family, the covariance between X_{ijM} and Y_{ij} should be zero under the null hypothesis. Since the expectation of X_{ijM} is .5, an estimate of the covariance is $\frac{1}{t}\Sigma_{j=1}^{t}(Y_{ij} - \bar{Y})(X_{ijM} - .5)$. Likewise, for a family with $h_i = 2$, the covariance between $X_{ijM} + X_{ijF}$ and Y_{ij} should be zero. The expectation of $X_{ijM} + X_{ijF}$ is 1; therefore, an estimate of the covariance is $\frac{1}{t}\Sigma_{j=1}^{t}(Y_{ij} - \bar{Y})(X_{ijM} + X_{ijF} - 1)$. For the ith family, these estimates of covariance can be written, in our notation, as random variable U_i:

$$U_i = \frac{1}{t}\sum_{j=1}^{t}(Y_{ij} - \bar{Y})[X^*_{iM}(X_{ijM} - .5) + X^*_{iF}(X_{ijF} - .5)] \ .$$

Suppose that our data set contains families with at least one parent that is heterozygous at the marker locus; in this instance, for a random family, we have U_i corresponding to $h_i = 1$ or $h_i = 2$. In Appendix B, it is shown that the expectation of U_i for such families is

$$E(U_i) = \frac{D(1 - 2\theta)[a + d(q_2 - q_1)]}{4p_1p_2(1 - p_1p_2)} \ . \qquad (1)$$

Hence, we see that the expectation of U_i will be zero, if at least one of the following four scenarios is true:

1. The marker and QTL are unlinked (i.e., $\theta = 0.5$).
2. There is no association between the marker and the QTL (i.e., $D = 0$).
3. The QTL has no effect on the trait (i.e., $a = d = 0$).
4. $a + d(q_2 - q_1)$ is 0 without $a = d = 0$.

Scenario (4), although possible, is not a likely scenario. Scenario (3) contradicts the definition of a QTL so that only scenarios (1) and (2) are pertinent. Furthermore, scenarios (1) and (2) jointly compose the null hypothesis; therefore, discounting scenario (4), the null hypothesis will be true if and only if $E(U_i) = 0$. Although we have assumed, throughout this discussion, that the population is homogeneous—that is, there is no stratification—it is not difficult to show that, if there is population stratification, the expectation of U_i will be zero if there is no association or no linkage within each subpopulation (see Appendix A).

If alleles A_1 and Q_1 are positively associated—that is, if $D > 0$—then, since Q_1 causes high quantitative-trait values (because $a > 0$), we would expect the covariance to be positive. In other words, high trait values will often occur with transmissions of allele A_1, and, therefore, the expectation of U_i will be positive. Similarly, if the A_1 and Q_1 alleles are negatively associated ($D < 0$), we would expect the covariance to be negative. From equation (1), we can see that this is, indeed, often true; however, it is possible to see a negative covariance resulting from a positive association or a positive covariance resulting from a negative association. Nonetheless, in all of these situations, the expectation of U_i will not be equal to zero.

It is thus reasonable to construct a test of association and linkage, on the basis of U_i. We will denote the number of families with one heterozygous parent as F_1 and

the number of families with two heterozygous parents as F_2. When the family is noted as the independent unit, a reasonable test statistic is

$$\frac{\bar{U}}{\sqrt{\mathrm{Var}_0(U_i)/(F_1 + F_2)}}\ , \tag{2}$$

where $\bar{U}$ is the mean of the $F_1 + F_2$ values of U_i and where $\mathrm{Var}_0(U_i)$ is the variance of the random variable U_i, under the null hypothesis. Knowledge of the underlying genetic model would be needed to compute $\mathrm{Var}_0(U_i)$ exactly. This type of information will rarely be available; therefore, an estimate must be obtained from the data. Since the expectation of U_i is 0 under the null hypothesis, an estimate of $\mathrm{Var}_0(U_i)$ is given as $s^2_{U0} = \Sigma_{i=1}^{F_1+F_2} U_i^2/(F_1 + F_2)$. It is noted that the U_i do not have to be identically distributed, as long as their expectation is 0. An example of such a situation arises when sampling families with information on different numbers of children. Using this estimate in our statistic from equation (2), we have

$$T_{QP} = \frac{\bar{U}}{\sqrt{s^2_{U0}/(F_1 + F_2)}} = \frac{\sum_{i=1}^{F_1+F_2} U_i}{\sqrt{\sum_{i=1}^{F_1+F_2} U_i^2}}\ .$$

Under the alternative hypothesis, the expectation of U_i will be nonzero, and the estimate of variance will no longer be correct. If the number of families sampled is large, then the distribution of T_{QP} can be approximated by a normal random variable with a nonzero mean and with unit variance multiplied by a factor that is a function of $\mathrm{Var}_A(U_i)$, which is the variance of U_i under the alternative hypothesis, and the expectation of s^2_{U0} under the alternative hypothesis:

$$\begin{aligned} T_{QP} &= \frac{\bar{U}}{\sqrt{s^2_{U0}/(F_1 + F_2)}} \\ &= \frac{\bar{U}}{\sqrt{\mathrm{Var}_A(U_i)/(F_1 + F_2)}} \times \sqrt{\frac{\mathrm{Var}_A(U_i)}{s^2_{U0}}} \\ &\approx \frac{\bar{U}}{\sqrt{\mathrm{Var}_A(U_i)/(\tilde{F}_1 + \tilde{F}_2)}} \times \sqrt{\frac{\mathrm{Var}_A(U_i)}{E(s^2_{U0})}}\ , \end{aligned}$$

where $\tilde{F}_1$ and $\tilde{F}_2$ are the expected values of F_1 and F_2, respectively. That is,

$$\tilde{F}_1 = 4p_1p_2(1 - 2p_1p_2)F \tag{3}$$

and

$$\tilde{F}_2 = 4p_1^2p_2^2F\ . \tag{4}$$

From this approximation, we can compute the power for a given model as well as the average sample size needed to achieve a specified power. Let

$$W_{QP} = \frac{\bar{U}}{\sqrt{\mathrm{Var}_A(U_i)/(\tilde{F}_1 + \tilde{F}_2)}} \tag{5}$$

and let

$$\gamma_{QP} = \sqrt{\frac{\mathrm{Var}_A(U_i)}{E(s^2_{U0})}}\ ,$$

so that $T_{QP} \approx W_{QP} \times \gamma_{QP}$. Substituting equations (1), (3), and (4) into the expectation of W_{QP} from equation (5), we get

$$\begin{aligned} E(W_{QP}) &= \frac{E(U_i|h_i = 1 \text{ or } h_i = 2)}{\sqrt{\mathrm{Var}_A(U_i)/(\tilde{F}_1 + \tilde{F}_2)}} \\ &= \frac{\frac{D}{4p_1p_2(1-p_1p_2)}(1 - 2\theta)[a + d(q_2 - q_1)]}{\sqrt{\mathrm{Var}_A(U_i)/[4p_1p_2(1 - 2p_1p_2)F + 4p_1^2p_2^2F]}} \\ &= \frac{\sqrt{F}D(1 - 2\theta)[a + d(q_2 - q_1)]}{\sqrt{\mathrm{Var}_A(U_i)}\sqrt{4p_1p_2(1 - p_1p_2)}}\ . \end{aligned}$$

Suppose that we are interested in testing the alternative hypothesis that there is positive association ($D > 0$) between the marker allele A_1 and the QTL allele that causes high trait values, and suppose that we are assuming that this corresponds to the expectation of U_i being positive. If we let z_α denote the value such that $Pr(Z \geq z_\alpha) = \alpha$, where Z is a standard normal random variable, then power is given by

$$\begin{aligned} Pr(T_{QP} \geq z_\alpha) &\approx Pr(W_{QP} \times \gamma_{QP} \geq z_\alpha) \\ &= Pr\left(W_{QP} \geq \frac{z_\alpha}{\gamma_{QP}}\right) \\ &\approx Pr\left[Z \geq \frac{z_\alpha}{\gamma_{QP}} - E(W_{QP})\right]\ . \end{aligned}$$

Given a marker and QTL model, along with the type I error rate (α), power can be computed by use of a standard normal distribution.

Of greater interest is the calculation of the sample size needed for a specified power of $1 - \beta$—that is, calcu-

lation of F for which $Pr(T_{QP} \geq z_\alpha) = 1 - \beta$. When the above approximation is used, F must satisfy:

$$z_{1-\beta} = \frac{z_\alpha}{\gamma_{QP}} - E(W_{QP}) \ . \tag{6}$$

Only $E(W_{QP})$ is a function of F. Solving equation (6) for F, we get:

$$F = \left(\frac{z_\alpha}{\gamma_{QP}} - z_{1-\beta}\right)^2 \left\{\frac{4p_1p_2(1 - p_1p_2)\mathrm{Var}_A(U_i)}{D^2(1 - 2\theta)^2[a + d(q_2 - q_1)]^2}\right\} \ .$$

It is noted that, although t does not appear explicitly in the formula for F, it will affect the variance of U_i and the expectation of s^2_{U0}. Formulas for these are not shown; however, a program that calculates these quantities —along with power or sample size—is available from the authors (University of Washington School of Public Health and Community Medicine Biostatistics).

Next, consider the case for which no parental-genotype information is available. Since inference of parental genotypes in an unknown mixture of populations is not straightforward, we chose not to infer parental genotypes. We instead used informative families to indicate the presence of at least one parent that is heterozygous at the marker. A family is informative if there are at least two children with differing marker genotypes. The probability of an informative family with t children is

$$Pr(\text{info}) = 4p_1p_2(1 - 2p_1p_2)\left(1 - \frac{1}{2^{t-1}}\right) + 4p_1^2p_2^2\left(1 - \frac{1}{2^{2t-1}} - \frac{1}{2^t}\right) \ .$$

We define the following random variable for the ith family: $V_i = \frac{1}{t}\Sigma_{j=1}^{t}(Y_{ij} - \bar{Y})(X_{ijM} + X_{ijF} - \bar{X}_i)$. This is analogous to U_i with $X^*_{iM}(X_{ijM} - .5) + X^*_{iF}(X_{ijF} - .5)$ replaced by an estimate. Conditional on the family being informative, the expectation of V_i is

$$E(V_i|\text{info}) = \frac{1}{Pr(\text{info})}\left(\frac{t-1}{t}\right)D(1 - 2\theta)[a + d(q_2 - q_1)] \ .$$

See Appendix C for the derivation. It is interesting to note that

$$E(V_i|\text{info}) = \frac{t-1}{t} \times \frac{4p_1p_2(1 - p_1p_2)}{Pr(\text{info})} \times E(U_i|h_i = 1 \text{ or } h_i = 2) \ ,$$

and, so, as t increases, the expected value of V_i approaches that of U_i. Following the same reasoning used in the construction of T_{QP}, we define a statistic on the basis of an estimate of the variance of V_i under the null hypothesis. Without loss of generality, suppose that the first F_I of the F families sampled are informative. Under the null hypothesis, the expectation of V_i is zero, so that an estimate of the null variance is $s^2_{V0} = (\Sigma_{i=1}^{F_I}V_i^2)/F_I$. Let $\bar{V}$ be the mean of the V_i for the F_I informative families. Our statistic is

$$T_{QS} = \frac{\bar{V}}{\sqrt{s^2_{V0}/F_I}} = \frac{\sum_{i=1}^{F_I} V_i/F_I}{\sqrt{(1/F_I)\left(\sum_{i=1}^{F_I} V_i^2/F_I\right)}} = \frac{\sum_{i=1}^{F_I} V_i}{\sqrt{\sum_{i=1}^{F_I} V_i^2}} \ .$$

T_{QS} is asymptotically standard normal under the null hypothesis. As was the case for U_i, under the alternative hypothesis, the expectation of V_i will be nonzero, and, for large samples, an approximation can be used to compute power and sample size. Denote the variance of V_i, under the alternative hypothesis, as σ^2_{VA}, and denote the expected number of informative families as $\tilde{F}_I = FPr(\text{info})$. Then the approximate power of the T_{QS} test, for a test of positive association with type I error α, is given by

$$Pr\left[Z \geq z_\alpha\sqrt{\frac{E(s^2_{V0})}{\sigma^2_{VA}}} - E(V_i|\text{info})\sqrt{\frac{\tilde{F}_I}{\sigma^2_{VA}}}\right] \ ,$$

where Z is a standard normal random variable. For power equal to $1 - \beta$, we have a required sample size of

$$F = \left(z_\alpha\sqrt{\frac{E(s^2_{V0})}{\sigma^2_{VA}}} - z_{1-\beta}\right)^2\left(\frac{t}{t-1}\right)^2 \times \left\{\frac{\sigma^2_{VA}Pr(\text{info})}{D^2(1 - 2\theta)^2[a + d(q_2 - q_1)]^2}\right\} \ .$$

It is straightforward to combine the two types of family data. Let F_P represent the number of families for which there is parental-genotype information and in which at least one of the parents is heterozygous. Let F_S represent the number of informative families that do not

have parental-genotype information. Define the following test statistic as

$$T_{QPS} = \frac{\left(\sum_{i=1}^{F_P} U_i + \sum_{i=1}^{F_S} V_i\right)/(F_P + F_S)}{\sqrt{[1/(F_P + F_S)]^2 (F_P s_{U0}^2 + F_S s_{V0}^2)}}$$

$$= \frac{\sum_{i=1}^{F_P} U_i + \sum_{i=1}^{F_S} V_i}{\sqrt{\sum_{i=1}^{F_P} U_i^2 + \sum_{i=1}^{F_S} V_i^2}} \ .$$

Our use of this statistic combines the two types of families by giving more weight to the most-sampled type. As before, T_{QPS} is asymptotically standard normal under the null hypothesis and will have the same properties as T_{QP} and T_{QS}. Following the preceeding work for T_{QP} and T_{QS}, it is straightforward to compute either the power for a given model or the sample size required for a specified power. The only additional information needed for these computations is what fraction of the families will (or will not) have parental information.

Permutation Procedures

For small samples, permutation procedures can be used to determine significance levels for the T_{QP}, T_{QS}, and T_{QPS} tests. These procedures can also be used to determine either the significance of extensions to multiallelic markers or extensions that utilize more than one marker.

Under the null hypothesis, the probability that a heterozygous parent transmits marker allele A_1 to a child with trait value Y is equal to .5. Thus, if the mother is heterozygous, then, for child j with trait value Y_{ij}, X_{ijM} is equally likely to be 0 or 1. If there is only one child, then a permutation procedure can be based on random assignment of X_{ijM} as being equal to 0 or 1 with equal probability. Complications arise when more than one child in the family has been sampled. These complications are a result of linkage between the marker and the QTL. In the presence of linkage, children with shared marker alleles will have similar quantitative traits, even in the absence of association. This can be taken into account by simultaneous randomization of X_{ijM} (and, similarly, of X_{ijF}), for heterozygous parents across the sibship. Consider the X_{ijM} for family i. Given the vectors $\mathbf{T_{iM}} = [X_{i1M}, X_{i2M}, \cdots, X_{itM}]'$ and $1 - \mathbf{T_{iM}} = [1 - X_{i1M}, 1 - X_{i2M}, \cdots, 1 - X_{itM}]'$, a permutation procedure can be constructed by randomization between $\mathbf{T_{iM}}$ and $1 - \mathbf{T_{iM}}$. It is easy to show that this procedure is equivalent to randomization of the sign of U_{iM}. An analogous procedure holds for U_{iF} with vectors $\mathbf{T_{iF}}$ and $1 - \mathbf{T_{iF}}$. If the parental contributions to U_i cannot be determined, then the sign of U_i can instead be randomized.

A similar scenario arises when parental-genotype information is not available. First, consider the following form of V_i

$$V_i = \frac{1}{t}\sum_{j=1}^{t} (Y_{ij} - \bar{Y})(X_{ijM} + X_{ijF} - \bar{X}_i)$$

$$= \frac{1}{t}\sum_{j=1}^{t} (Y_{ij} - \bar{Y})(X_{ijM} - \bar{X}_{iM})$$

$$+ \frac{1}{t}\sum_{j=1}^{t} (Y_{ij} - \bar{Y})(X_{ijF} - \bar{X}_{iF}) \ ,$$

where $\bar{X}_{iM}$ ($\bar{X}_{iF}$) is the fraction of $\bar{X}_i$ contributed by the mother (father). Noting that V_i will be nonzero only if at least one of the parents is heterozygous, suppose that only the mother is heterozygous. Randomization between the $\mathbf{T_{iM}}$ and $1 - \mathbf{T_{iM}}$, for the mother, will be valid under the null hypothesis. While we cannot determine these vectors, this randomization is equivalent to randomization of the sign of V_i. An analogous case exists if only the father is heterozygous. If both parents are heterozygous, then there are four equally likely permutations:

1. $\mathbf{T_{iM}}$ and $\mathbf{T_{iF}}$
2. $\mathbf{T_{iM}}$ and $1 - \mathbf{T_{iF}}$
3. $1 - \mathbf{T_{iM}}$ and $\mathbf{T_{iF}}$
4. $1 - \mathbf{T_{iM}}$ and $1 - \mathbf{T_{iF}}$.

Since we do not know the values for these vectors, we cannot randomize among the four permutations; however, we can randomize between permutations 1 and 4. This is again equivalent to randomization of the sign of V_i.

This results in a unified permutation procedure. Given U_{iM}, U_{iF}, U_i, or V_i for a family, a permutation procedure is based on randomization of the sign of the observed value. We suggest use of a Monte Carlo approximation to measure significance. Details are given elsewhere (Monks et al. 1998).

Simulation Parameters

We considered 18 QTL models. Heritability (H^2) was equal to .1, .3, or .5. The QTL allele Q_1 had a population frequency of .1 or .5. We studied QTLs with additive, dominant, and recessive modes of inheritance. By setting these parameters, the means for the trait distributions for individuals with QTL genotypes Q_1Q_1, Q_1Q_2, and Q_2Q_2 were uniquely determined. Conditional on an individual's genotype, the trait distribution is normal, with the appropriate mean and variance, σ_E^2, equal to 1.

Marker allele A_1 had a population frequency of .5 or .8 and was completely linked to the QTL. The disequilibrium coefficient was set to 0 for simulations under

Table 1

Estimates of Significance Level for the T_R Test

			Estimates of Significance Level for[a]						
			T_R						T_{QP}
H^2	$Pr(Q_1)$	$Pr(A_1)$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 6$
.1	.1	.5	.010	.011	.011	.012	.014	.014	.010
.1	.1	.8	.010	.011	.011	.012	.013	.014	.010
.1	.5	.5	.010	.011	.012	.012	.013	.014	.010
.1	.5	.8	.010	.011	.012	.012	.013	.014	.010
.3	.1	.5	.010	.012	.015	.018	.021	.024	.010
.3	.1	.8	.010	.012	.015	.018	.021	.024	.010
.3	.5	.5	.010	.012	.015	.018	.021	.024	.010
.3	.5	.8	.010	.012	.015	.018	.021	.024	.010
.5	.1	.5	.010	.014	.018	.024	.029	.034	.010
.5	.1	.8	.010	.014	.018	.024	.029	.034	.010
.5	.5	.5	.010	.014	.019	.023	.028	.034	.010
.5	.5	.8	.010	.014	.019	.024	.028	.034	.010

[a] Estimates are based on 1,000,000 simulated samples of 500 families with t children, for a QTL with additive mode of inheritance. Estimates for the T_{QP} test, with $t = 6$, are also given.

the null hypothesis, and it was set to its maximum value for simulations under the alternative hypothesis.

We derived formulas for the power of the T_{QP} and T_{QS} tests. The powers for the T_R test with $t = 1$ and for the T_A test with $t = 2$ can be computed from the formulas for the T_{QP} and T_{QS} tests, respectively. For estimates of significance level, we simulated 1,000,000 data sets, to achieve precision to three decimal places. For estimates of the significance level for the permutation procedures for T_{QP} and T_{QS}, we used 10,000 simulated data sets with 99 permutations. For estimates of significance level within a stratified population, we also used 10,000 simulated data sets. All estimates correspond to a one-sided test of positive association, with a significance level of $\alpha = .01$.

Results

Significance Levels of T_R *and* T_A *for Increasing* t

Table 1 contains estimates of the significance level for the T_R test, for a QTL with additive mode of inheritance. Estimates are based on 500 families that all have the same number of children t. We show estimates for $t = 1, \cdots, 6$. For $t = 1$, the T_R test is valid as a test of association, and, so, our estimates are equal to the actual significance level of 0.01. As t increases, the actual level of significance increases as a result of the nonvalidity of the T_R test. The increase becomes more extreme as heritability, H^2, increases. In particular, consider the QTL/marker model with $Pr(Q_1) = .1$ and $Pr(A_1) = .5$, for sibships of size $t = 6$. The estimate of significance level at $H^2 = .1$ is .014, which is much less than the estimate for $H^2 = .3$, which is .024. When heritability is .5, the estimate of significance level, .034, is even larger. Estimates of significance are also given, for our T_{QP} test, for families with $t = 6$. All estimates are equal to .01 and thus support the validity of the T_{QP} test. The results for a QTL with dominant and recessive modes of inheritance were similar to those presented (data not shown).

Table 2 contains estimates of the significance level for the T_A test, for a QTL with additive mode of inheritance. Estimates are based on 500 families, all of which have the same number of children. We give estimates for $t = 2, \cdots, 6$. For $t = 2$, the T_A test is a valid test of association, as our estimates confirm. As t increases, the actual level of significance also increases, as a result of the nonvalidity of T_A. As was the case for T_R, this increase becomes greater as heritability increases. For the same model mentioned above, estimates of significance are .014, .024, and .037 for heritability of .1, .3, and .5, respectively. Estimates of significance are also given, for our T_{QS} test, for families with $t = 6$; in all cases, the estimates are equal to .01. The results for a QTL with dominant and recessive modes of inheritance were similar to those presented (data not shown).

The amount of increase from the expected level of significance, for the T_R and T_A tests, will depend on more than heritability. The QTL model, marker model, number of children, expected level of significance (in this case, .01), and sample size will all affect the increase. Unfortunately, we will not usually know the parameters of our model and, therefore, will not know the impact of nonvalidity. All that can be stated is that the level of significance will be larger than that which is expected.

Comparison of the T_{QP} *and* T_R *Tests*

We have established that the T_{QP} test is a valid test of association, regardless of the number of children in the

Table 2

Estimates of Significance Level for the T_A Test

			Estimates of Significance Level for[a]					
			T_A					T_{QS}
H^2	$Pr(Q_1)$	$Pr(A_1)$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 6$
.1	.1	.5	.010	.011	.012	.013	.014	.010
.1	.1	.8	.010	.011	.012	.013	.014	.010
.1	.5	.5	.010	.011	.012	.013	.014	.010
.1	.5	.8	.010	.011	.012	.012	.014	.010
.3	.1	.5	.010	.014	.017	.020	.024	.010
.3	.1	.8	.010	.014	.017	.020	.024	.010
.3	.5	.5	.010	.014	.017	.020	.024	.010
.3	.5	.8	.010	.013	.017	.020	.024	.010
.5	.1	.5	.010	.017	.024	.031	.037	.010
.5	.1	.8	.009	.017	.023	.030	.037	.010
.5	.5	.5	.010	.017	.024	.030	.037	.010
.5	.5	.8	.010	.017	.024	.030	.037	.010

[a] Estimates are based on 1,000,000 simulated samples of 500 families with t children, for a QTL with additive mode of inheritance. Estimates for the T_{QS} test, with $t = 6$, are also given.

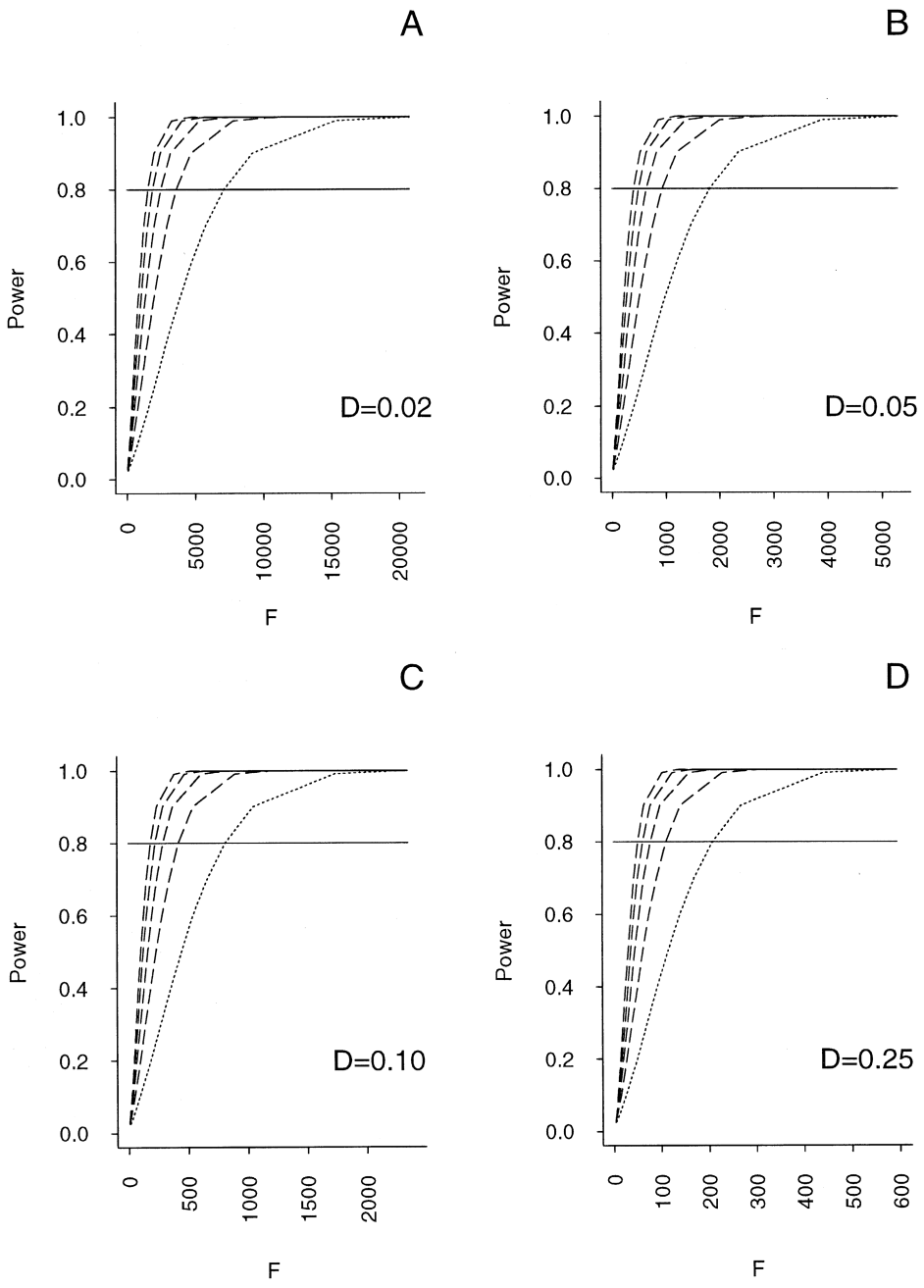

Figure 2 Estimated power for the T_{QP} and T_R tests of association, for a QTL with additive mode of inheritance and heritability equal to .1. For figures *A–D*, $\{Pr(A_1),Pr(Q_1)\}$ was set equal to {0.8,0.1}, {0.5,0.1}, {0.8,0.5} and {0.5,0.5}, respectively. The T_{QP} (*dashed lines*) test is based on samples of families with two, three, four, or five children (power curves are indicated by increasing power with t), whereas the T_R (*dotted line*) test is based on families with one child. A solid line indicates power equal to 0.8.

family that have been sampled (see the Methods section above). Although it is expected that the T_{QP} test based on information for all children will be a more-powerful test of association than will the T_R test based on information for a single child, to what extent this will occur is unknown. To investigate this, we compared the power of the T_{QP} test done with samples of families with t children with the power of the T_R test done with samples of families with one child. Figure 2 contains power curves for the QTL/marker models with additive mode of inheritance and heritability equal to .1, for the T_R test and for the T_{QP} test, with t = 2–5. It is noted that the T_{QP} statistic reduces to the T_R statistic when t = 1. The four models provide power comparisons over a range

of values for the disequilibrium coefficient. Specifically, figure 2*A–D* shows disequilibrium coefficients equal to .02, .05, .1, and .25, respectively. Panels *A–D* demonstrate that, as the disequilibrium increases, the sample size required for reasonable power decreases. However, despite the level of disequilibrium, the relationship between the T_{QP} and T_R tests is clear. Significant information is lost by use of only families of minimal size—that is, families where $t = 1$. As an example, consider figure 2*B*, for which $Pr(A_1) = .5$ and $Pr(Q_1) = .1$. For 80% power, the T_{QP} test requires 404 families with five children, whereas the T_R test requires 1,813 families or, with respect to genotyping, 2,828 genotypes, compared with 5,439 genotypes, respectively. From a planning standpoint, more than four times as many singleton families, compared with families with five children, need to be collected for 80% power. While it is clear that a good deal of power is gained by use of more than one child per family, it is also apparent that, with each additional child used, there is a diminishing gain in power. The largest gain in power is obtained by use of two children per family rather than one. The increase in power gained from use of three—rather than two—children is also substantial; however, information gained from an increase beyond three children per family continues to diminish. For model 2C, the number of families required for 80% power are 808, 417, 287, 221, and 182 for families with one, two, three, four, and five children, respectively. Approximately half as many families with $t = 2$ are needed, compared with families with $t = 1$. This is a sizable decrease, compared with the ~20% fewer families required with families with five children compared with those with four children. Although most of the gain in power is derived from increasing t from 1 to 2 and from 2 to 3, it is striking to note that, by use of five children—rather than just one child—per family, only 23% as many families are needed.

Comparison of the T_{QS} *and* T_A *Tests*

We also compared the T_{QS} test, in which sibships of size three, four, and five were used, with the T_A test, in which sibships of size two were used. The tests were compared with regard to the QTL/marker models in figure 2 (results not shown). Table 3 contains the values of the ratio F_t/F_2, where F_t is the number of families required for the T_{QS} test, with samples of sibships of size t, to have 80% power, and where F_2 is the number of families required for the T_A test to have 80% power. Ratios are given for the four models of figure 2, with $t = 2–5$. The T_{QS} statistic is equal to the T_A test statistic, when $t = 2$. Conclusions reached from these results are identical to those achieved when T_{QP} is compared with T_R: considerable information is obtained by use of more than two children per family, and the largest gains in power result from an increase from $t = 2$ to $t = 3$ or from $t = 3$ to $t = 4$. We see that, for each of the models, approximately half as many families are required for T_{QS} with $t = 3$ as are required for T_A. Approximately one-third as many families are required for T_{QS} with $t = 4$ as are required for T_A. Both of these decreases in required sample size are considerable; however, we again see that the increase in power and, therefore, the decrease in sample size, diminish as t increases.

Table 3

Sample-Size Ratios for the T_{QS} and T_A Tests, at 80% Power

	Ratio[a] for Models in Figure 2			
t	A	B	C	D
2	1.000	1.000	1.000	1.000
3	.506	.510	.511	.522
4	.343	.348	.350	.364
5	.261	.268	.270	.286

[a] Ratio of the sample size required for the T_{QS} test, with samples of families with t children, to achieve 80% power to the sample size required for the T_A test, with samples of families with two children. Models correspond to those of figure 2.

Comparison of the T_{QP} *and* T_{QS} *Tests*

The question arises, with the use of any family-based test, as to how much information is gained by genotyping of parents. To answer this question, we sampled families with a constant number of children t, with $t = 2–6$. We calculated the number of families, F_P, required for the T_{QP} test to have power equal to 80%. We then calculated the number of sibships, F_S, needed for the T_{QS} test to have 80% power. Table 4 contains the F_P/F_S ratio for the 12 marker/QTL models with an additive mode of inheritance for $t = 2–6$. Now the statistics T_{QP} and T_{QS} are composed of random variables U and V, respectively. Since V is an estimate of U, the statistic T_{QS} is expected to be more variable than is T_{QP} and, thus, is expected to result in a less-powerful test. The F_P/F_S ratio should therefore be <1. As t increases, the estimate V will improve, and, so, the ratio should increase to one. The results support this. For each of the 12 models, the ratio is smallest for $t = 2$, and it increases with t. Consider, as an example, the model with $H^2 = .1$, $q_1 = .5$, and $p_1 = .5$. For $t = 2$, the ratio is .524. In other words, the T_{QP} test requires only 52.4% of the families that are required for the T_{QS} test. However, if families with six children were sampled, the T_{QP} test would require 86.3% of the families required for the T_{QS} test, thus bringing into question how much effort should be given toward collection of parental genotypes. A few ratios were >1.

Table 4

Sample-Size Ratios for the T_{QP} and T_{QS} Tests, at 80% Power

			Sample-Size Ratio[a]				
H^2	$Pr(Q_1)$	$Pr(A_1)$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
.1	.1	.5	.526	.705	.795	.849	.885
.1	.1	.8	.522	.699	.788	.841	.876
.1	.5	.5	.524	.700	.792	.850	.863
.1	.5	.8	.525	.707	.795	.847	.881
.3	.1	.5	.580	.783	.880	.931	.967
.3	.1	.8	.571	.770	.868	.924	.957
.3	.5	.5	.573	.762	.844	.889	.957
.3	.5	.8	.578	.777	.870	.926	.971
.5	.1	.5	.636	.858	.956	1.000	1.040
.5	.1	.8	.630	.850	.951	1.005	1.035
.5	.5	.5	.604	.793	.870	.900	.944
.5	.5	.8	.636	.847	.938	.981	1.021

[a] Ratio of the sample size required for the T_{QP} test, with samples of t children, to achieve 80% power to the sample size required for the T_{QS} test. All models are for a QTL with additive mode of inheritance.

After further investigation, we have found that, for higher values of heritability, it is possible to have the ratio >1 (results not shown). We also computed the F_P/F_S ratio for the same 12 marker/QTL models with dominant and recessive modes of inheritance (data not shown). Identical conclusions were reached.

Permutation Procedure for T_{QP} *and* T_{QS}

We provide evidence that our permutation procedures for T_{QP} and T_{QS} are valid, by estimation of the significance levels for the T_{QP} and T_{QS} tests for the 36 marker/QTL models. Estimates were based on data from 200 families with a constant number of children t. The range of the t value was 1–5, with the exception that T_{QS} was not applicable for $t = 1$. The 36 × 5 marker/QTL/sampling models were numbered from 1–180. Figure 3 contains a plot of the estimates of significance levels. Two SD lower and upper bounds are indicated. The significance-level estimates fall satisfactorily within two SDs of .01. This is the case for both the T_{QP} test (fig. 3*A*) and the T_{QS} test (fig. 3*B*).

Validity of the T_{QP} *and* T_{QS} *Tests under Stratification*

To demonstrate that the T_{QP} and T_{QS} tests are valid when there is population stratification, we simulated a population that is a mixture of two homogenous subpopulations. We considered the scenarios where .5 and .75 of our sampled families are from subpopulation 1. For the 12 QTL/marker models with heritability of .1, we simulated samples of 500 families with five children, for all possible assignments of one of these models to subpopulation 1 and of another of the models to subpopulation 2 (132 possibilities). Table 5 contains the mean estimate of the significance level, across these 132 models, for both the T_{QP} test and the T_{QS} test, when subpopulation 1 comprises .5 and .75 of the population. We have also given 95% confidence intervals for each mean. Two of the confidence intervals do not contain the expected significance level of .01. We attribute this to our χ^2 approximation. The estimates deviate from .01 by very little. Furthermore, the deviation is in the opposite direction of that which would be expected as a result of problems with stratification.

Discussion

Family-based methods have previously been introduced for testing association of markers or candidate genes with a QTL. These tests avoid the increase in the false-positive rate that occurs in the typical case-control test if there is population stratification. However, the current family-based association tests have sampling restrictions that result in a loss of information. However, if these sampling restrictions are not followed, then the false-positive rate of the tests will increase. Furthermore, we have demonstrated, through simulation, that the amount of increase in the false-positive rate will be

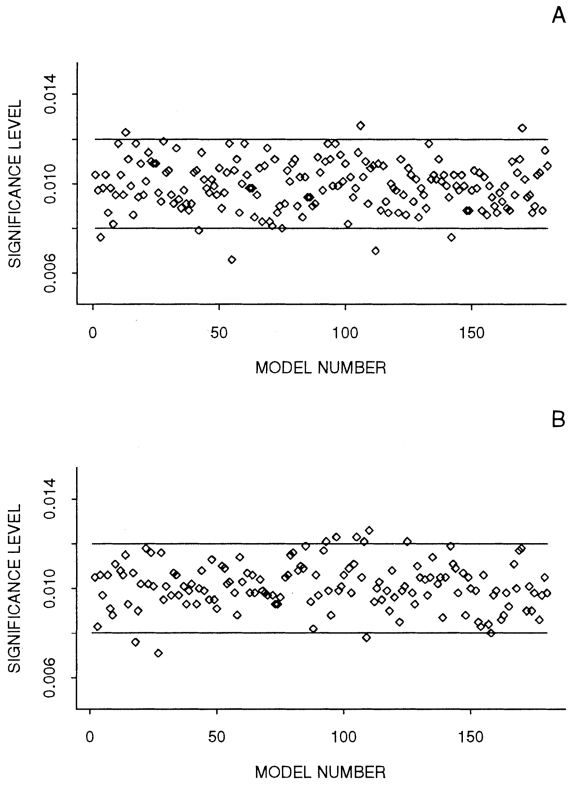

Figure 3 Estimates of the significance level for the T_{QP} and T_{QS} tests for the 36 marker/QTL models. Estimates are given for samples of 200 families of constant size t, where $t = 1$–5. The 36 × 5 models were numbered 1–180. Note that T_{QS} is not applicable for $t = 1$.

model dependent. Since the true underlying model is rarely known, a researcher will be unaware of the true effect of nonvalidity. This problem will be compounded as more and more markers or candidate genes are tested, since a researcher will have no sense of how many of their positive results may be in error.

We have developed three tests that are valid tests of association (and linkage) without the sampling restrictions of current tests. In addition, we have shown that a great deal of power is acquired by use of all children. In fact, for all three of our tests, the power increases when more children are used. If association tests are to be performed with the use of a previously obtained data set, then this would imply that all children should be used in the test of association. If a study is being designed, then other considerations, such as ascertainment and genotyping costs, will determine whether fewer large families or more small families should be sampled. For the majority of the models we have considered, the use of families with parental-genotype information will allow for a more-powerful test than will tests in which only sibship information is used. The size of this increase in power is largely dependent on how many children have been sampled. As the number of children in the family increases, the power gained by having parental information diminishes. In practice, a data set will contain families with and without parental genetic information. While one strategy would be to ignore the parental information and to use the T_{QS} test, we would recommend the use of our T_{QPS} test. Scenarios do exist where use of sibship information only can result in a more-powerful test than will use of the available parental-transmission information; however, these models generally have high heritability. Although we have shown that considerable power can be obtained by sampling more than the minimal number of children per family, it may be that current tests could be used, in conjunction with special reduction strategies, to reduce families or sibships to minimal size. One strategy would be to randomly sample the minimal number of children from each family and to compute the significance level on the basis of this reduced data set. Other strategies might sample the minimal number of children but might specify that they have extreme trait values (e.g., largest in family, a discordant pair, etc.). These methods should account for some of the power difference seen with use of all children compared with use of the minimal number of children needed (one for T_R and two for T_A). We are currently investigating these strategies.

For simplicity, we have assumed that all families have an equal number of children. If samples of families with an arbitrary number of children are available, then our statistics have the same algebraic form, and only the interpretation of the statistic changes. The mean of the random variable studied (U_i when parental information is available and V_i when sibship information is used) is now a weighted mean across family size, and the variance of the random variable is also a weighted variance across family size. The asymptotics rely on the number of families within each class, where class is defined by family size. If any of the classes have few families, then the normal approximation will be poor, and we would recommend the use of our permutation procedures.

The tests that we have proposed are for a diallelic marker; however, they are easily extended to multiallelic markers. There are two straightforward extensions. A statistic can be computed for each marker allele, and, as an overall statistic, the maximum of their absolute value or the sum of their squares can be used. The permutation procedures can then be used to measure significance. In addition to extensions to multiple alleles, it is possible to extend these methods to the use of multiple tightly linked markers. One could compute a statistic for each marker and then could define an overall statistic—perhaps the maximum or sum across markers. The permutation procedure can be used to measure significance, by simultaneously shuffling across the markers (Lazzeroni and Lange 1998; McIntyre et al., in press).

We have not included a technical discussion of costs. A researcher will have to weigh the genotyping and ascertainment costs, to determine whether resources should be spent on sampling additional children or on

Table 5

Estimates of Significance Level for the T_{QP} and T_{QS} Tests in a Stratified Population

Proportion of Sample from		Mean[a] Estimated Significance Level (95% CI) for	
Subpopulation 1	Subpopulation 2	T_{QP}	T_{QS}
.50	.50	.0097 (.00957, .00991)	.0097 (.00955, .00989)
.75	.25	.0099 (.00975, .01009)	.0100 (.00981, .01015)

[a] The mean estimated significance level for the 132 marker/QTL models (see text for details). Estimates are based on 10,000 simulated samples of 500 families with five children, for a QTL with heritability of .1.

sampling parental-genotype information. In terms of power, for our models, larger families with parental-genotype information provide the greatest power; however, costs will determine the "optimal" sampling scheme. Allison (1997) has provided a discussion of costs associated with ascertainment, genotyping, and phenotyping.

Acknowledgments

The authors would like to thank Dr. Bruce S. Weir, for numerous helpful discussions. S.A.M. would particularly like to thank the National Institute of Environmental Health Sciences, for training through an Intramural Research Training Award. This material is also based on work supported under a National Science Foundation graduate fellowship (to S.A.M.).

Appendix A

Effect of Population Stratification on the Expectation of U_i

Consider a population that is composed of B subpopulations, where the probability of a random family from subpopulation b is ϕ_b. Suppose that there is random mating within each of the subpopulations, so that each subpopulation will have an expectation and variance of U_i specific to that subpopulation's allele frequencies and linkage-disequilibrium coefficient. Denote the expectation and variance as μ_b and σ_b^2, respectively. The expectation and variance of U_i, for a random family, are therefore $E(U_i) = \Sigma_{b=1}^{B} \phi_b \mu_b$ and

$$\begin{aligned}\mathrm{Var}(U_i) &= E[\mathrm{Var}(U_i|\text{subpopulation } b)] + \mathrm{Var}[E(U_i|\text{subpopulation } b)] \\ &= \sum_{b=1}^{B} \phi_b \sigma_b^2 + \mathrm{Var}(\mu_b) \\ &= \sum_{b=1}^{B} \phi_b \sigma_b^2 + \sum_{b=1}^{B} \phi_b \mu_b^2 - \left(\sum_{b=1}^{B} \phi_b \mu_b\right)^2 .\end{aligned}$$

Under the null hypothesis, there is no association, within each subpopulation, between alleles at the marker locus and the QTL. It follows, from equation (1) in the text, that the expectation of U_i, for each of the subpopulations, is 0—that is, $\mu_b = 0$ for $b = 1,...,B$. Thus, the expectation and variance of U_i, for a random family, are 0 and $\sum_{b=1}^{B} \phi_b \sigma_b^2$, respectively. The estimate of variance used in the construction of T_{QP} is an estimate of this variance, so that T_{QP} will be asymptotically standard normal under the null hypothesis. A discussion of the effects of population stratification and admixture is provided elsewhere (Ewens and Spielman 1995).

Appendix B

Derivation of the Expectation of U_i

Families with one or two heterozygous parents provide within-family information about association between alleles at the marker locus and QTL. We begin by considering families with one heterozygous parent. Without loss of generality, assume that the mother is heterozygous for the marker. Then the random variable U_i reduces to $U_i = \frac{1}{t_i}\Sigma_{j=1}^{t_i}(Y_{ij} - \bar{Y})(X_{ijM} - 0.5)$. Let $p_{i|r}$ be the conditional marker-allele probability $Pr(A_i|Q_r)$. Denote the mother's marker/QTL haplotypes as H_{iM1} and H_{iM2}. Let $tr{\rightarrow}Q_r$ represent the event that allele Q_r has been transmitted to the given individual. Denote the probability that a marker-homozygous parent transmits the QTL allele, Q_1, to a child, by use of $\tilde{q}_1$, where $\tilde{q}_1 = q_1 + D(p_1 - p_2)/(p_1^2 + p_2^2)$, and denote the probability that it transmits a Q_2, by use of $\tilde{q}_2 = 1 - \tilde{q}_1$. Conditional on $X_{iM}^* = 1$, $X_{iF}^* = 0$ and with the assumption that $\bar{Y} \approx \mu$ and $t_i = t$,

$$
\begin{aligned}
E(U_i) &= \frac{1}{t}\sum_{j=1}^{t} E[(Y_{ij}-\mu)(X_{ijM}-.5)] \\
&= E[(Y_{i1}-\mu)(X_{i1M}-.5)] \\
&= \sum_{r=1}^{2}\sum_{s=1}^{2} E[(Y_{i1}-\mu)(X_{i1M}-.5)|H_{iM1}=A_1Q_r,\ H_{iM2}=A_2Q_s] \\
&\quad \times Pr(H_{iM1}=A_1Q_r,\ H_{iM2}=A_2Q_s) \\
&= \sum_{r=1}^{2}\sum_{s=1}^{2} E[(Y_{i1}-\mu)(X_{i1M}-.5)|H_{iM1}=A_1Q_r,\ H_{iM2}=A_2Q_s]\left(\frac{p_{1|r}p_{2|s}q_rq_s}{p_1p_2}\right) \\
&= \left(\frac{p_{1|1}p_{2|2}q_1q_2}{p_1p_2}\right) E[(Y_{i1}-\mu)(X_{i1M}-.5)|H_{iM1}=A_1Q_1,\ H_{iM2}=A_2Q_2] \\
&\quad + \left(\frac{p_{1|2}p_{2|1}q_1q_2}{p_1p_2}\right) E[(Y_{i1}-\mu)(X_{i1M}-.5)|H_{iM1}=A_1Q_2,\ H_{iM2}=A_2Q_1] \\
&= \left(\frac{p_{1|1}p_{2|2}q_1q_2}{p_1p_2}\right)\left(\frac{1}{4}(1-2\theta)\{E[(Y_{i1}-\mu)|tr\to Q_1]-E[(Y_{i1}-\mu)|tr\to Q_2]\}\right) \\
&\quad + \left(\frac{p_{1|2}p_{2|1}q_1q_2}{p_1p_2}\right)\left(\frac{1}{4}(1-2\theta)\{E[(Y_{i1}-\mu)|tr\to Q_2]-E[(Y_{i1}-\mu)|tr\to Q_1\}\right) \\
&= \frac{1-2\theta}{4}\times\frac{q_1q_2}{p_1p_2}\times(p_{1|1}p_{2|2}-p_{1|2}p_{2|1}) \\
&\quad \times\{E[(Y_{i1}-\mu)|tr\to Q_1]-E[(Y_{i1}-\mu)|tr\to Q_2]\} \\
&= \frac{1-2\theta}{4}\times\frac{q_1q_2}{p_1p_2}\times(p_{1|1}p_{2|2}-p_{1|2}p_{2|1})\times\{\tilde{q}_1E[(Y_{i1}-\mu)|tr\to Q_1,\ tr\to Q_1]\} \\
&\quad +(\tilde{q}_2-\tilde{q}_1)E[(Y_{i1}-\mu)|tr\to Q_1,tr\to Q_2]-\tilde{q}_2E[(Y_{i1}-\mu)|tr\to Q_2,\ tr\to Q_2] \\
&= \frac{1-2\theta}{4}\times\frac{q_1q_2}{p_1p_2}\times(p_{1|1}p_{2|2}-p_{1|2}p_{2|1}) \\
&\quad \times[(\mu_{11}-\mu)\tilde{q}_1+(\mu_{12}-\mu)(\tilde{q}_2-\tilde{q}_1)-(\mu_{22}-\mu)\tilde{q}_2)] \\
&= \frac{1-2\theta}{4}\times\frac{q_1q_2}{p_1p_2}\times(p_{1|1}p_{2|2}-p_{1|2}p_{2|1})\times(\mu_{11}\tilde{q}_1+\mu_{12}(\tilde{q}_2-\tilde{q}_1)-\mu_{22}\tilde{q}_2) \\
&= \frac{D}{4p_1p_2}(1-2\theta)\left[a+(q_2-q_1)d+2dD\left(\frac{p_2-p_1}{p_1^2+p_2^2}\right)\right] .
\end{aligned}
$$

We obtain the same derivation for families in which only the father is heterozygous. Thus, for families with one heterozygous parent:

$$
E(U_i|h_i=1)=\frac{D}{4p_1p_2}(1-2\theta)\left[a+(q_2-q_1)d+2dD\left(\frac{p_2-p_1}{p_1^2+p_2^2}\right)\right] . \tag{B1}
$$

For families with two heterozygous parents, we need the probability that a marker-heterozygous parent transmits the QTL allele, Q_1, to a child. Denote this probability as $\breve{q}_1$, where $\breve{q}_1 = q_1 - D(p_1-p_2)/(2p_1p_2)$, and denote the probability that a Q_2 is transmitted as $\breve{q}_2 = 1-\breve{q}_1$. We can write U_i as the sum of two components:

$$U_i = \frac{1}{t}\sum_{j=1}^{t}(Y_{ij} - \bar{Y})X^*_{iM}(X_{ijM} - .5) + \frac{1}{t}\sum_{j=1}^{t}(Y_{ij} - \bar{Y})X^*_{iF}(X_{ijF} - .5)$$
$$= U_{iM} + U_{iF} .$$

Thus, conditional on $X^*_{iM} = X^*_{iF} = 1$ and assuming that the parents are of the same genetic background, we have $E(U_i) = 2E(U_{iM})$. The above derivation can be used to compute the expectation for U_{iM}, with one alteration. Transmissions from the other parent are now from a heterozygous parent, and, so, $\tilde{q}_1$ ($\tilde{q}_2$) must be replaced by $\breve{q}_1$ ($\breve{q}_2$). From this, we get

$$E(U_{iM}|h_i = 2) = \frac{D}{4p_1p_2}(1 - 2\theta)\left[a + (q_2 - q_1)d - dD\left(\frac{p_2 - p_1}{p_1p_2}\right)\right] . \quad \text{(B2)}$$

Using equations (B1) and (B2), we can derive the expectation of U_i for a family with at least one heterozygous parent:

$$\begin{aligned}
E(U_i) &= Pr(h_i = 1|h_i = 1 \text{ or } h_i = 2)E(U_i|h_i = 1) \\
&\quad + Pr(h_i = 2|h_i = 1 \text{ or } h_i = 2)E(U_i|h_i = 2) \\
&= \frac{4p_1p_2(1 - 2p_1p_2)}{4p_1p_2(1 - 2p_1p_2) + 4p_1^2p_2^2} \times \frac{D}{4p_1p_2}(1 - 2\theta) \\
&\quad \times \left[a + (q_2 - q_1)d + 2dD\left(\frac{p_2 - p_1}{p_1^2 + p_2^2}\right)\right] \\
&\quad + \frac{4p_1^2p_2^2}{4p_1p_2(1 - 2p_1p_2) + 4p_1^2p_2^2} \times \frac{D}{2p_1p_2}(1 - 2\theta) \\
&\quad \times \left[a + (q_2 - q_1)d - dD\left(\frac{p_2 - p_1}{p_1p_2}\right)\right] \\
&= \frac{D(1 - 2\theta)[a + d(q_2 - q_1)]}{4p_1p_2(1 - p_1p_2)} .
\end{aligned}$$

Thus, we have

$$E(U_i|h_i = 1 \text{ or } h_i = 2) = \frac{D(1 - 2\theta)[a + d(q_2 - q_1)]}{4p_1p_2(1 - p_1p_2)} .$$

Appendix C

Derivation of the Expectation of V_i for an Informative Family

As was the case for U_i, the random variable V_i is the sum of two components, one from the mother (V_{iM}) and one from the father (V_{iF}):

$$
\begin{aligned}
V_i &= \frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij}-\bar{Y})(X_{ijM}+X_{ijF}-\bar{X}_i)\\
&= \frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij}-\bar{Y})(X_{ijM}-\bar{X}_{iM})+\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij}-\bar{Y})(X_{ijF}-\bar{X}_{iF})\\
&= V_{iM}+V_{iF}\ ,
\end{aligned}
$$

where $\bar{X}_{iM}$ ($\bar{X}_{iM}$) is the part of $\bar{X}_i$ corresponding to the mother (father). There are three types of family defined by the number of heterozygous parents. Obviously, if family i has no heterozygous parents, then $E(V_i) = 0$. Next, consider families with exactly one heterozygous parent. Without loss of generality, suppose that the mother is heterozygous. Then, conditional on $X^*_{iM} = 1$, $X^*_{iF} = 0$ and letting $\bar{Y} \approx \mu$ and $t_i = t$, we have

$$
\begin{aligned}
E(V_i) &= E(V_{iM})\\
&= E\left[\frac{1}{t}\sum_{j=1}^{t}(Y_{ij}-\mu)(X_{ijM}-\bar{X}_{iM})\right]\\
&= E\left[\frac{1}{t}\sum_{j=1}^{t}(Y_{ij}-\mu)(X_{ijM}-.5+.5-\bar{X}_{iM})\right]\\
&= E(U_{iM})+E[(Y_{i1}-\mu)(.5-\bar{X}_{iM})]\\
&= E(U_{iM})+.5E[Y_{i1}-\mu]\\
&\quad -\frac{1}{t}E[(Y_{i1}-\mu)X_{i1M}]-\frac{t-1}{t}E[(Y_{i1}-\mu)X_{i2M}]\\
&= E(U_{iM})+\frac{1}{2t}E(Y_{i1}-\mu)-\frac{1}{t}E[(Y_{i1}-\mu)X_{i1M}]\\
&= E(U_{iM})\\
&\quad +\frac{1}{2t}\sum_{r=1}^{2}\sum_{s=1}^{2}E[Y_{i1}-\mu|H_{iM1}=A_1Q_r,\ H_{iM2}=A_2Q_s]\\
&\qquad\qquad \times Pr(H_{iM1}=A_1Q_r,\ H_{iM2}=A_2Q_s)\\
&\quad -\frac{1}{t}\sum_{r=1}^{2}\sum_{s=1}^{2}E[(Y_{i1}-\mu)X_{i1M}|H_{iM1}=A_1Q_r,\ H_{iM2}=A_2Q_s]\\
&\qquad\qquad \times Pr(H_{iM1}=A_1Q_r,\ H_{iM2}=A_2Q_s)\\
&= \frac{D}{4p_1p_2}(1-2\theta)\left[a+d(q_2-q_1)+2dD\left(\frac{p_2-p_1}{p_1^2+p_2^2}\right)\right]\\
&\quad +\frac{1}{2t}\times\frac{D(p_2-p_1)^2}{2p_1p_2(p_1^2+p_2^2)}\times\{[a+d(q_2-q_1)](p_2-p_1)+2dD\}\\
&\quad -\frac{1}{t}\left(\frac{D}{2p_1}\left\{a+d(q_2-q_1)+[2dD-p_1a-p_1d(q_2-q_1)]\left(\frac{p_2-p_1}{p_1^2+p_2^2}\right)\right\}\right)
\end{aligned}
$$

$$-\frac{D\theta}{2p_1p_2}\left[a + d(1-2q_1) + 2dD\left(\frac{p_2-p_1}{p_1^2+p_2^2}\right)\right]$$
$$=\frac{t-1}{t}\times\frac{D}{4p_1p_2}\times(1-2\theta)\left[a + d(q_2-q_1) + 2dD\left(\frac{p_2-p_1}{p_1^2+p_2^2}\right)\right] .$$

Thus, we have

$$E(V_i|h_i=1)=\frac{t-1}{t}\times\frac{D}{4p_1p_2}\times(1-2\theta)\left[a + d(q_2-q_1) + 2dD\left(\frac{p_2-p_1}{p_1^2+p_2^2}\right)\right] . \quad \text{(C1)}$$

For families with two heterozygous parents, we have $E(V_i) = 2E(V_{iM})$ (with the assumption that the parents are from the same population). However, $E(V_{iM})$ is not that of equation (C1), since we are conditioning on two heterozygous parents. It can be shown that, conditional on there being two heterozygous parents,

$$E(V_i|h_i=2)=\frac{t-1}{t}\times\frac{D}{2p_1p_2}\times(1-2\theta)\left[a + d(q_2-q_1) - dD\left(\frac{p_2-p_1}{p_1p_2}\right)\right] . \quad \text{(C2)}$$

The T_{QS} test is recommended when parental information is not available, and, so, there will be no knowledge of how many of the family's parents are heterozygous. All that can be determined is whether a sibship is informative for the marker. Thus, we need the expectation of V_i, conditional on a family being informative. Using equations (C1) and (C2), we get

$$E[V_i|I(\text{info})]=\frac{E(V_i)}{Pr(\text{info})}$$
$$=\frac{1}{Pr(\text{info})}[Pr(h_i=1)E(V_i|h_i=1) + Pr(h_i=2)E(V_i|h_i=2)]$$
$$=\frac{1}{Pr(\text{info})}\left(\frac{t-1}{t}\right)D(1-2\theta)[a + d(q_2-q_1)] ,$$

where

$$Pr(\text{info}) = 4p_1p_2(1-2p_1p_2)\left(1-\frac{1}{2^{t-1}}\right) + 4p_1^2p_2^2\left(1-\frac{1}{2^{2t-1}}-\frac{1}{2^t}\right) .$$

Electronic-Database Information

The URL for data in this article is as follows:

University of Washington School of Public Health and Community Medicine Biostatistics, http://www.biostat.washington.edu/ steph/PROGRAMS/qtlassoc.html

References

Allison DB (1997) Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60:676–690

Allison DB, Heo M, Kaplan N, Martin ER (1999) Sibling-based tests of linkage and association for quantitative traits. Am J Hum Genet 64:1754–1763

Bickeboller H, Clerget-Darpoux F (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Genet Epidemiol 12:865–870

Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th ed. Longman Group, Harlow, Essex, England

Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–1897

Lazzeroni LC, Lange K (1998) A conditional inference frame-

work for extending the transmission/disequilibrium test. Hum Hered 48:67–81

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448

McIntyre LM, Martin ER, Simonson KL, Kaplan NL. Circumventing multiple testing: a multi-locus Monte Carlo approach to testing for association. Genet Epidemiol (in press)

Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. Am J Hum Genet 63:1507–1516

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342–350

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Schaid DJ, Rowland CR (1998) The use of parents, sibs, and unrelated controls to detection of associations between genetic markers and disease. Am J Hum Genet 63:1492–1506

Schaid DJ, Rowland CM (1999) Quantitative trait transmission disequilibrium test: allowance for missing parents. Genet Epidemiol 17:S307–S312

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. Ann Hum Genet 59:323–336

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

——— (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland, MA