Report

General Equations for P_t, P_s, and the Power of the TDT and the Affected–Sib-Pair Test

Ralph McGinnis

Biotechnology and Genetics, SmithKline Beecham Pharmaceuticals, Harlow, Essex, United Kingdom

Several equations are highlighted here, whose algebraic symmetries and generality make them very useful for understanding and comparing the properties of the transmission disequilibrium test (TDT) and affected sib-pair test. Methods using the equations are also presented that yield precise estimates of sample sizes needed for genome scans or for testing a single candidate gene, and these power methods are shown to compare favorably with alternative approaches recently described by Knapp (1999) and by Tu and Whittemore (1999). Simple relationships are also noted that summarize the relative sample sizes required for equivalent power to detect association by the TDT or case-control designs. As single-nucleotide polymorphism (SNP) maps revolutionize the search for disease-causing genes, the equations should prove useful for planning and evaluating studies of linkage and association across a broad range of possible disease models and relationships between markers and linked disease loci.

I wish to highlight several equations that are very useful for understanding and comparing the properties of the transmission disequilibrium test (TDT) (Spielman et al. 1993) and the "mean" affected sib-pair (ASP) test (Knapp et al. 1994). I previously demonstrated that these equations provide a general framework for determining the power of each test (McGinnis 1998); indeed, as I show here, the method I previously described predicts sample sizes for genomewide scans that are almost identical to those recently presented by Knapp (1999), Tu and Whittemore (1999), and Camp (1999). The equations are also among the most general presented to date, since they (*a*) cover general modes of inheritance, (*b*) describe any marker that is linked to a biallelic disease locus, and (*c*) subsume, as special cases, corresponding equations presented by Risch and Merikangas (1996) and Camp (1997, 1999). Perhaps most important, the two central equations (P_s and P_t in fig. 1) exhibit symmetries and a partitioning of the influence of basic genetic parameters and thus facilitate TDT-ASP test comparisons and help provide an intuitive understanding of the properties of each test.

The equations are applicable to understanding ASP analyses in general, but here I focus on the "mean" test (hereafter denoted "ASP test") because it is generally the most powerful form of ASP analysis (Knapp et al. 1994). The ASP test detects linkage in families considered to be single parent-ASP trios, if the two sibs share one of the parental alleles (identical-by-descent [ibd]) in significantly more than half of the trios in the data set [$\chi^2_{\text{asp}} = (n_s - n_n)^2/(n_s + n_n)$ where n_s and n_n are the total number of trios that exhibit ibd allele sharing (n_s) or nonsharing (n_n)]. The TDT also detects linkage in nuclear families but does so if a particular ("disease associated") marker allele is preferentially transmitted from heterozygous parents to individual affected offspring across an entire population or set of families [$\chi^2_{\text{tdt}} = (n_a - n_b)^2/(n_a + n_b)$ where n_a and n_b are the total number of instances that heterozygous A/B parents transmit marker allele A (n_a) or allele B (n_b) to individual affected offspring]. As explained below, the power of both χ^2's can be estimated by considering heterozygous parents in a data set to be identical, randomly selected, independent events, each having a probability (P_s) of increasing the value of n_s and χ^2_{asp}, and each having a second probability (P_t) of increasing the value of n_a and χ^2_{tdt}.

To facilitate an understanding of P_s, P_t, and related equations that I wish to highlight, I first describe their

Received June 9, 2000; accepted for publication August 10, 2000; electronically published September 29, 2000.

Address for correspondence and reprints: Dr. Ralph McGinnis, SmithKline Beecham Pharmaceuticals, New Frontiers Science Park (North), H30-2-038, Harlow, Essex CM19 5AW, United Kingdom. E-mail: Ralph_McGinnis@sbphrd.com

0002-9297/2000/6704-0033$02.00

Marker allele transmission from an informative (A/B) parent to sibs in an ASP

P_s - Probability that both sibs of the ASP received the same marker allele (A or B) identical-by-descent:

$$P_s = 0.5 + (1-2\theta)^2\left[\frac{c_1c_4 + c_2c_3}{H}\right]\left\{p^2\frac{(a-b)^2}{4} + 2p(1-p)\frac{(a-g)^2}{16} + (1-p)^2\frac{(b-g)^2}{4}\right\} = 0.5 + L_s\left[M_s\right]\left\{R_s\right\}$$

P_t - Probability that an individual affected sib received a particular marker allele (e.g. allele A):

$$P_t = 0.5 + (1-2\theta)\left[\frac{c_1c_4 - c_2c_3}{H}\right]\left\{p^2\frac{(a^2-b^2)}{4} + 2p(1-p)\frac{((a+b)^2-(b+g)^2)}{16} + (1-p)^2\frac{(b^2-g^2)}{4}\right\} = 0.5 + (L_t)\left[M_t\right]\left\{R_t\right\}$$

Proportion of parents of an ASP who are heterozygous (A/B) at a marker is given by the ratio H/F

$$H = 2c_1c_3W_{DD} + 2(c_1c_4 + c_2c_3)W_{Dd} + 2c_2c_4W_{dd} \qquad F = p^2W_{DD} + 2p(1-p)W_{Dd} + (1-p)^2W_{dd}$$

$$\text{where } W_{DD} = p^2a^2 + 2p(1-p)\left(\frac{a+b}{2}\right)^2 + (1-p)^2b^2$$

$$W_{Dd} = p^2\left(\frac{a+b}{2}\right)^2 + 2p(1-p)\left(\frac{a+2b+g}{4}\right)^2 + (1-p)^2\left(\frac{b+g}{2}\right)^2$$

$$W_{dd} = p^2b^2 + 2p(1-p)\left(\frac{b+g}{2}\right)^2 + (1-p)^2g^2$$

Figure 1 General equations for understanding and comparing the TDT and affected sib-pair (ASP) test. The equations assume (1) random ascertainment of nuclear families with an ASP, and (2) linkage between a biallelic marker (alleles A and B) and biallelic disease locus (predisposing allele D, "protective" allele d); with minor modifications, these equations also describe markers that are multiallelic or completely polymorphic (McGinnis 1998). Variables in the equations are as follows: *a, b,* and *g* are penetrances of the DD, Dd, and dd genotypes, respectively; c_1, c_2, c_3, and c_4 are population frequencies of AD, Ad, BD, and Bd haplotypes, respectively; *p* is the population frequency of disease allele D; and θ is the recombination fraction between marker and disease loci. The quantities *F* and *H* are proportional to the population frequencies of parents who have produced an ASP (*F*) and such parents who are also heterozygous (*H*). W_{DD}, W_{Dd}, and W_{dd} are "weights" that reflect the relative ability of DD, Dd, and dd parents to produce an ASP.

basic features and then discuss some results derived from them.

Basic Description of the Equations

The equations shown in figure 1 assume random ascertainment of families with an ASP and linkage between a biallelic marker (alleles A and B) and biallelic disease locus (alleles D and d); however, the same equations are easily modified to describe linked markers that are multiallelic or completely polymorphic (McGinnis 1998). P_s is the probability that a randomly ascertained, informative (A/B) parent transmitted the same marker allele (ibd) to both affected sibs; and P_t is the probability that the parent transmitted a particular marker allele (e.g., allele A) to an individual affected sib. Thus, P_s can be regarded as the probability of allele "sharing" by ASPs while P_t can be considered the probability of individual allele "transmission." The final section in figure 1 describes the numerator and denominator of *H*/*F*, which is the proportion of randomly ascertained parents of an ASP expected to be heterozygous at the marker. *H*/*F* is important in determining effective sample size since the TDT and ASP test only consider alleles transmitted from informative parents.

When effective sample size is fixed (i.e., $n_s + n_n$ and $n_a + n_b$ are constant), note that $\chi^2_{asp} = (n_s - n_n)^2/(n_s + n_n)$ will increase only if $(P_s - .5)$ increases, thereby increasing the n_s/n_n ratio; and analogously, $\chi^2_{tdt} = (n_a - n_b)^2/(n_a + n_b)$ will increase only if $|P_t - .5|$ increases (see Risch and Merikangas [1996] and McGinnis [1998]). Therefore, given their pivotal roles in driving the magnitudes of χ^2_{asp} and χ^2_{tdt}, what do the expressions in figure 1 reveal about the magnitudes and properties of P_s and P_t? In the absence of linkage, both probabilities equal 0.5; when linkage is present, each probability equals 0.5 plus the product of three factors ($P_s = 0.5 + L_sM_sR_s$, and $P_t = 0.5 + L_tM_tR_t$). Note that the leftmost factors

(L_s, L_t) depend only on the recombination fraction (θ), the middle factors (M_s, M_t) depend on haplotype frequencies (c_1,c_2,c_3,c_4) and the quantity H (see fig. 1), and the rightmost factors (R_s,R_t) depend only on the properties of the disease locus (i.e., disease-allele frequency p and penetrances a,b,g). In addition to R_s and R_t being completely independent of the marker, it was shown that $R_t > R_s$ (McGinnis 1998), and inspection also shows that $L_t \geqslant L_s$ for any marker. Thus, $|P_t - .5|$ would always exceed ($P_s - .5$) were it not for the influence of the middle factors (M_s,M_t).

Note, then, that M_t and M_s are identical except that the numerator of M_t is the disequilibrium coefficient δ (where $\delta = c_1c_4 - c_2c_3$) whereas the numerator of M_s is the two components of δ added together ($c_1c_4 + c_2c_3$). This implies that $M_s \geqslant |M_t|$ and, because $|M_t|$ reaches its minimum of 0 at equilibrium ($\delta = 0$) whereas M_s is always positive, it follows that $(P_s - .5) > |P_t - .5| \approx 0$ when disequilibrium between marker and disease locus is low. However, suppose allele frequencies at the biallelic marker and the biallelic disease locus are fixed. It was shown in McGinnis (1998) that $|M_t|$ and M_s are (a) both *maximized* and (b) *equal to each other* when the disease-causing allele exhibits maximum possible association with the marker allele nearest in frequency (denote such disequilibrium as $\delta_{\max}$). Since $R_t > R_s$ and $L_t \geqslant L_s$, it follows that $|P_t - .5| > (P_s - .5)$ as δ values move toward $\delta_{\max}$, the actual level of disequilibrium at which $|P_t - .5|$ exceeds ($P_s - .5$) being dependent on the degree of elevation of R_t above R_s.

On the basis of the algebraic symmetries in R_t and R_s, it can be shown that the R_t/R_s ratio is most extreme for disease loci conferring *modest* disease risk (McGinnis 1998). To illustrate this extreme elevation of R_t/R_s and resulting elevation in $|P_t - .5|/(P_s - .5)$ if the marker and disease locus are strongly associated, figure 2 shows P_t and P_s when the DD homozygote has only twofold-greater disease risk than the dd homozygote. Figure 2 also assumes that the tested marker *is* the disease locus—in which case, $|P_t - .5|/(P_s - .5) = R_t/R_s$. The figure's much higher values of $|P_t - .5|$ compared with ($P_s - .5$) visually illustrates why the TDT has greater power than the ASP test to detect genes of modest effect, if a marker is found that is strongly associated with the disease locus. (For fuller discussion of the relative power of the TDT and the ASP test, see Tu and Whittemore [1999] and McGinnis [1998]).

Two further points should be noted about the equations in figure 1. First, if a biallelic marker is in linkage *equilibrium* with the disease locus, the expression for P_s simplifies to

$$P_s = 0.5 + (1 - 2\theta)^2 \left[\frac{p(1-p)}{F}\right] R_s\ ,$$

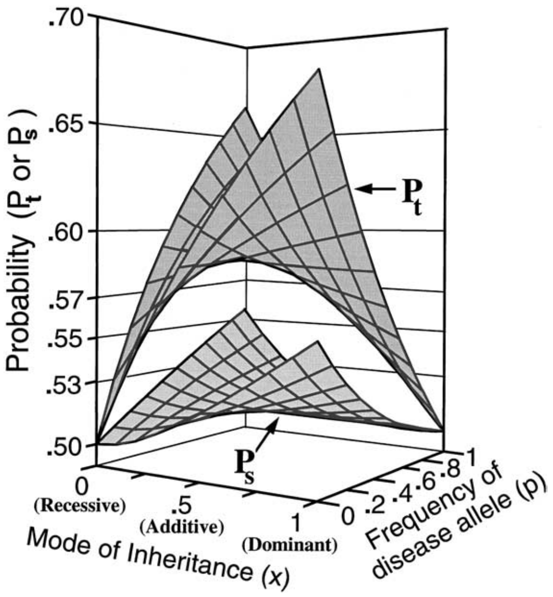

Figure 2 Values of P_t and P_s showing extreme elevation of $|P_t - .5|/(P_s - .5) = R_t/R_s$ for a disease locus that confers modest disease risk. The three-dimensional, "saddle-shaped" surfaces showing values of P_t (upper surface) and P_s (lower surface) were calculated assuming that DD penetrance (a) is only two times greater than dd penetrance (g) and assuming that the marker *is* the disease locus. Identity of the marker and disease locus implies that $L_s = L_t$ and $M_s = |M_t|$ in the expressions $P_s = .5 + L_sM_sR_s$ and $P_t = .5 + L_tM_tR_t$, and thus $|P_t - .5|/(P_s - .5) = R_t/R_s$. Values of P_t or P_s (Z-axis) are plotted as a function of the frequency (p) of disease allele D (Y-axis) and mode of inheritance (X-axis) where X denotes Dd penetrance (b) in terms of its numerical "location" ($1 \geqslant X \geqslant 0$) between dd penetrance (where $X = 0$) and DD penetrance (where $X = 1$). The much higher value of $|P_t - .5|$ compared to ($P_s - .5$) across the entire parameter space shows why the TDT has greater power than the ASP test to detect genes of modest effect, if a marker is found that is strongly associated with the disease locus.

which is identical to the expression for P_s when a marker is completely polymorphic (McGinnis 1998). This enables power to be calculated when χ^2_{asp} is applied to a completely polymorphic marker (see below). The second point is that the expressions for P_s, P_t, and H/F each simplify to corresponding expressions ("Y," "P-trA" and "h") presented by Risch and Merikangas (1996) when their assumptions are adopted (multiplicative mode of inheritance, χ^2_{asp} evaluates a completely polymorphic marker, χ^2_{tdt} evaluates a biallelic marker that *is* the disease locus) (for more details, see McGinnis [1998]). P_s, P_t, and H/F also simplify to corresponding expressions ("Y," "τ_p," and "h_p") in Camp (1997) that were derived under the same assumptions as in Risch and Merikangas (1996), except that penetrances were relaxed to cover many modes of inheritance. As noted by Camp (1999), these probabilities for allele transmission and parental heterozygosity are *correct,* even though the method of

TDT-power estimation in Camp (1997) was somewhat inaccurate. Thus, expressions in Risch and Merikangas (1996) and Camp (1997) are special cases of equations in figure 1 which are more general, in part, because they describe markers not necessarily identical with the disease locus.

Power of the ASP Test and TDT

Table 1 gives expressions for the number (N) of families required by the TDT and ASP test to achieve power of $(1-\beta)$ assuming one ASP per family is considered and that the probability of type I error is α. Using the expressions in figure 1 for P_s, P_t, and H/F, "Method 1" extends the approach of Risch and Merikangas (1996 [see note 6]) to general modes of inheritance and, following those authors, assumes that the total number of heterozygous parents randomly varies around the expected value $2NH/F$. "Method 2" was described in McGinnis (1998) and is based on a single binomial distribution that assumes the total number of heterozygous parents to be exactly $2NH/F$. Both methods can calculate χ^2_{asp} and χ^2_{tdt} samples sizes for biallelic markers of any allele frequency and in any degree of linkage disequilibrium with the disease locus. If a marker is completely polymorphic, χ^2_{asp} sample size is calculated by setting $H/F = 1$ and

$$P_s = 0.5 + (1 - 2\theta)^2\left[\frac{p(1-p)}{F}\right]R_s ,$$

for reasons noted above; in this situation, the formulae for Methods 1 and 2 become identical. Since both methods can calculate and compare the power of χ^2_{asp} and χ^2_{tdt} for general modes of inheritance and for markers distinct from the disease locus, they provide an alternative to the only other approach that, to my knowledge, is similarly general (Tu and Whittemore 1999). Below, I argue that Methods 1 and 2 are simpler and more general than the method of Tu and Whittemore, with negligible loss of accuracy.

Comparison with Alternative Methods for Calculation of TDT Power in ASPs

The expressions in figure 1 for P_s and P_t correctly account for the fact that the probability of parental allele transmission partly depends on the genotype of the parent's mate (see appendix I in McGinnis [1998]). However, Methods 1 and 2 consider parents to be randomly ascertained as a series of independent events, rather than as pairs (parent and mate), in order to simplify power calculation and derivation of a normal distribution for values of χ^2_{asp} and χ^2_{tdt} (Risch and Merikangas 1996; McGinnis 1998; see also Camp 1999 and appendix C

Table 1

Two Methods for Calculating Number (N) of ASP Families Required by the TDT and the ASP Test to Achieve Power of $(1-\beta)$

	Formulae[a]	
Test	Method 1[b]	Method 2[c]
TDT[d]	$N = \dfrac{\left[Z_{\alpha/2} - \sqrt{1 - \frac{H}{F}(2P_t - 1)^2}Z_{1-\beta}\right]^2}{4\frac{H}{F}(2P_t - 1)^2}$	$N = \dfrac{\left[Z_{\alpha/2} - 2\sqrt{P_t(1 - P_t)}Z_{1-\beta}\right]^2}{4\frac{H}{F}(2P_t - 1)^2}$
ASP test[e]	$N = \dfrac{\left[Z_{\alpha} - \sqrt{1 - \frac{H}{F}(2P_s - 1)^2}Z_{1-\beta}\right]^2}{2\frac{H}{F}(2P_s - 1)^2}$	$N = \dfrac{\left[Z_{\alpha} - 2\sqrt{P_s(1 - P_s)}Z_{1-\beta}\right]^2}{2\frac{H}{F}(2P_s - 1)^2}$

[a] α and β are probabilities of type 1 and 2 error, respectively, (i.e., $1-\beta$ = power); Z_x ($x = \alpha$, $\alpha/2$, or $1 - \beta$) is the value of the standard normal deviate (Z) such that prob($Z > Z_x$) = x; $x = \alpha$ or $\alpha/2$ for one- and two-tailed tests, respectively.

[b] Using the general expressions for P_s, P_t, and H/F in figure 1, Method 1 extends the approach of Risch and Merikangas (1996 [note 6]) to general modes of inheritance and to markers not necessarily identical with the disease locus.

[c] Method 2 was described by McGinnis (1998). Whereas Method 1 allows the total number of heterozygous parents to vary randomly around the expected value of $2NH/F$, Method 2 uses a binomial distribution that assumes exactly $2NH/F$ heterozygous parents.

[d] When multiplied by 2, these formulae calculate the number (N) of *singleton* families required by the TDT if singleton values for P_t, H, and F are substituted (denoted as P_t^*, H^*, and F^* in the appendix, which gives the appropriate expressions).

[e] If a marker is completely polymorphic, N for the ASP test is determined by setting $H/F = 1$ and $P_s = (1 - 2\theta)^2[p(1 - p)/F]R_s$ (see text).

of Tu and Whittemore 1999). Methods 1 and 2 also assume that each sib of an ASP contributes independently to the TDT χ^2 statistic, an assumption previously found to have negligible effects on χ^{2}_{tdt} power estimates (see appendix II of McGinnis [1998]). By contrast, Knapp (1999) and Tu and Whittemore (1999) recently presented more complex algebraic formulations that avoid both of these simplifying assumptions in order to achieve more precise estimates of sample size. Knapp's two formulae ("First Approximation" and "Second Approximation") both calculate TDT sample size under the assumption that a biallelic marker *is* the disease locus, with the "First Approximation" giving more precise results (judging from simulations) over all tested modes of inheritance. Tu and Whittemore's method, as noted above, is more general, since it can evaluate TDT sample size for biallelic markers distinct from the disease locus.

To quantify differences among the methods, I used each of them to calculate the number (N) of ASP families required by the TDT for 80% power to detect linkage under each of the 48 disease models previously considered by Knapp (1999). Following Knapp (1999), the linked marker was assumed to be identical to the biallelic disease locus. I calculated samples sizes needed, assuming a TDT genome scan with 500,000 single-nucleotide polymorphisms (SNPs) ($\alpha/2 = 5 \times 10^{-8}$) or assuming a single TDT test ($\alpha/2 = .025$) as would occur in evaluating one SNP in a candidate gene. For each pair of methods, table 2 shows the mean percentage difference in N for the 48 disease models ± standard deviation and, inside parenthesis, shows the largest percentage difference in N observed over all 48 models. Note that differences among the methods are generally small but are categorically larger for testing one biallelic candidate than for a genome scan. Therefore, table 3 shows the actual sample sizes for the candidate gene test, to give a more concrete sense of the degree of difference among the methods.

I wish to make several points about these results:

1. Surprisingly, Tu and Whittemore's method and the Second Approximation of Knapp gave identical numerical results over all 48 disease models. This suggests that the two approaches are essentially identical, even though Tu and Whittemore's method is more general since it can evaluate markers that are distinct from the disease locus.
2. Methods 1 and 2 showed very little divergence from the First Approximation of Knapp, in terms of both mean difference over the 48 disease models and largest single difference observed (tables 2 and 3). Since Knapp's simulations led him to conclude that the First Approximation is extremely precise, the minimal divergence with Methods 1 and 2 indicates that these methods are also very precise.
3. Although precision in sample-size estimates is important, simplicity and interpretability of the method are also important. As the only general alternative to Methods 1 and 2, actual calculation by the method of Tu and Whittemore seems much more complicated since their method requires a number of summations and step-by-step substitutions into multiple matrices and equations.
4. The magnitude of pairwise differences shown in tables 2 and 3 did *not* increase when Methods 1 and 2 and the method of Tu and Whittemore were used to calculate TDT sample sizes for biallelic markers *not* identical with the disease locus (data not shown). These pairwise comparisons were conducted for each of the 48 disease models retaining the assumption of equal frequency for the marker and disease allele, but disequilib-

Table 2

Pairwise Comparisons between the Four Methods

	Tu and Whittemore	Method 1	Method 2
Genome Scan:			
Knapp	1.4% ± 1.6% (6%)	0.7% ± 0.7% (3%)	1.0% ± 0.9% (3%)
Tu and Whittemore	...	1.2% ± 1.1% (4%)	2.3% ± 2.3% (9%)
Method 1	...	...	1.2% ± 1.4% (5%)
Candidate gene:			
Knapp	3.4% ± 3.6% (13%)	1.9% ± 1.7% (6%)	1.9% ± 1.8% (7%)
Tu and Whittemore	...	2.6% ± 2.4% (9%)	5.0% ± 5.0% (18%)
Method 1	...	...	2.6% ± 3.0% (11%)

NOTE.—Four methods are: the "first approximation" of Knapp (1999), the method of Tu and Whittemore (1999) (which gave identical numerical results to the "second approximation" of Knapp [1999]), and Methods 1 and 2 as described in the text and in table 1. Number (N) of ASP families needed by the TDT to achieve 80% power ($Z_{1-\beta} = -.84$) was calculated by each method for the 48 disease models considered by Knapp (1999). N was calculated assuming a genome scan of 500,000 SNPs ($\alpha/2 = 5 \times 10^{-8}$, $Z_{\alpha/2} = 5.33$) or a test of one candidate SNP ($\alpha/2 = .025$, $Z_{\alpha/2} = 1.96$). For each pairwise comparison between two methods, the table shows the mean percent difference in N for the 48 disease models ± SD and also shows the largest percent difference observed (in parentheses).

Table 3

Number (*N*) of ASP Families Needed for 80% Power, as Calculated by the Four Methods, Assuming TDT Test of One Candidate SNP

	Mode of Inheritance[a]															
γ and	Multiplicative				Additive				Recessive				Dominant			
p[b]	Kn	T-W	M1	M2	Kn	T-W	M1	M2	Kn	T-W	M1	M2	Kn	T-W	M1	M2
4.0:																
.01	46	53	48	43	50	57	53	47	1.5×10^5	1.5×10^5	1.5×10^5	1.5×10^5	51	62	54	49
.10	9	10	10	9	15	16	16	15	238	244	224	222	18	19	20	19
.50	12	13	12	11	26	28	27	26	19	20	19	18	99	100	102	102
.80	31	36	33	30	73	76	74	72	35	39	36	33	1473	1477	1489	1488
2.0:																
.01	399	418	405	392	399	418	405	392	2.6×10^6	2.6×10^6	2.6×10^6	2.6×10^6	418	437	425	412
.10	53	55	54	53	53	55	54	53	3109	3116	3052	3050	81	82	83	81
.50	36	37	37	36	36	37	37	36	88	89	86	86	225	226	230	230
.80	80	83	81	79	80	83	81	79	100	103	100	98	2937	2941	2965	2963
1.5:																
.01	1592	1620	1601	1582	1531	1559	1538	1519	1.3×10^7	1.3×10^7	1.3×10^7	1.3×10^7	1645	1672	1655	1636
.10	193	195	193	192	141	144	140	139	14455	14463	14336	14334	264	266	267	265
.50	99	100	99	99	47	48	47	46	295	296	291	290	520	521	528	527
.80	193	195	193	192	85	88	86	84	262	264	260	259	6171	6175	6214	6213

[a] Kn = "first approximation" of Knapp (1999); T-W = method of Tu and Whittemore (1999); M1 = Method 1; M2 = Method 2. Following Knapp (1999), the tested marker is assumed to be the disease locus. Relative to the low-risk homozygote (assigned a disease risk of 1), the higher-disease-risk homozygote and heterozygote have respective disease risks of γ^2 and γ (multiplicative inheritance), 2γ and γ (additive inheritance), γ and 1 (recessive inheritance), and γ and γ (dominant inheritance).

[b] p = frequency of the disease-predisposing allele; γ = relative disease risk of heterozygotes compared with homozygotes who lack the disease-predisposing allele

rium was allowed to vary in 25% decrements from +100% disequilibrium (highest positive value, equivalent to marker and disease locus being identical) to −100% disequilibrium (most negative value, equivalent to maximum association between the disease allele and the alternate marker allele).

5. When Methods 1 and 2 are used to calculate TDT sample size for *singleton* families (see footnote d of table 1), the results again match or surpass the degree of precision indicated by tables 2 and 3. Indeed, Methods 1 and 2 give number of singleton families required for genome scans that show little or no departure from the singleton estimates given in table 3 of Knapp (1999) and table 1 of Camp (1999) (data not shown). The expressions in the appendix for singleton families ("P_t*" and "*H**/*F**") can be shown to be related to important singleton probabilities derived by Ott (1989) and by Sham and Curtis (1995) (see McGinnis [1998]). P_t*and *H**/*F** also subsume as special cases corresponding singleton probabilities ("τ_s" and "h_s") presented by Camp (1997).

6. The expressions for P_s and H/F enable χ^2_{asp} power to be evaluated for incompletely polymorphic markers like SNPs and also enable the influence of linkage disequilibrium on χ^2_{asp} to be considered. This makes methods 1 and 2 more general than alternative approaches that only analyze χ^2_{asp} power for completely polymorphic markers or for markers in equilibrium with the disease locus. Given the imminent release of dense SNP maps with markers likely to be in linkage disequilibrium with each gene, this additional capability may prove valuable.

7. Like formulae presented by others (Risch and Merikangas 1996; Camp 1997; Knapp 1999; Tu and Whittemore 1999), the sample-size formulae in table 1 are interpreted as number of *families*—being based on the assumption that each family contributes the minimum data required for family ascertainment (i.e., one ASP per family under ASP ascertainment, one affected child per family under singleton ascertainment). This is reasonable since it can be shown that 90%–96% of ascertained families contain only the minimum data under many disease models (e.g., in all 48 models considered in table 3, if disease prevalence is ⩽1% and mean nuclear family size is <8). However, I wish to note that the equations for singleton families (P_t*, *H**/*F**) and ASP families (P_t, P_s, H/F) accurately account for the expected distribution of families with more than the minimum number of affected offspring required for ascertainment (see appendix I in McGinnis [1998]). Thus the equations accurately apply to randomly ascertained families with more than the minimum data. If the fraction of such families is substantial (due to high population prevalence of disease or large mean size of nuclear families), the formulae in table 1 would still be accurate but are better interpreted as the required number of individual affected offspring or ASPs in randomly ascertained families, rather than total number of families.

Relative Sample Sizes Required by the TDT and Case-Control Designs

Since association studies with disease cases and unrelated or related controls is an important complementary strategy to association testing by the TDT, it should be noted that two simple relationships summarize the relative sample sizes required by the TDT and two types of case-control design. Using the formulae described here (see footnote d of table 1), Darvasi and McGinnis (1998) calculated the number of trios (parents–one affected child) required by the TDT to achieve equivalent power to case-control studies having an equal number of unrelated cases and unrelated controls. With no exceptions, we found the number of required cases to be almost identical to the number of trios required by the TDT across a broad range of disease models and degrees of linkage disequilibrium between marker and disease loci. When applied to each of the 48 disease models in table 3, I found that the formulae for unrelated cases and controls in Risch and Teng (1998 [pp. 1280–1281]) also yield a required number of cases that is nearly identical to the number of trios required by the TDT, for either a genome scan or test of one candidate SNP.

A second relationship summarizes the relative sample sizes needed for equal power by the TDT and S-TDT, a design that compares cases and *related* (i.e., unaffected sib) controls (Spielman and Ewens 1998). Whittaker and Lewis (1999) discovered that if each test is applied to families with one affected child and *n* unaffected sibs, the S-TDT requires $(n + 1)/n$ times as many such families as the TDT for equal power. So, to summarize, the formulae for calculating number of singleton families required by the TDT (see footnote d of table 1) also imply approximate sample sizes required for association studies with either unrelated cases and controls, or with unrelated cases each of which is matched to *n* unaffected sibling controls.

Influence of Polygenic Background

The equations I have presented do not explicitly account for modulation of disease penetrance by polygenic background loci, a limitation shared with previous equations describing TDT and ASP power (Risch and Merikangas 1996; Camp 1997; Knapp 1999; Tu and Whittemore 1999). When such background genotypes are present, each penetrance of the "foreground" locus (*a, b,* or *g*) represents the *mean* of a distribution of penetrance values for the DD, Dd, or dd genotype, each distribution being generated by the frequencies of the background genotypes and their specific effects on DD, Dd, or dd penetrance. Preliminary investigation (author's unpublished data) indicates that the equations for singleton families (P_t*, H*, and F*) are unchanged by the model's inclusion of background genotypes, the only difference being that *a, b,* and *g* would now represent means of penetrance distributions in the general population. However, the equations for sib pairs (P_s, P_t, *H,* and *F*) appear to require each occurrence of *a, b,* or *g* to be replaced by mean penetrance in the general population *added to* a term related to the variance of the corresponding penetrance distribution. This suggests that the expressions for P_s, P_t and *H/F* may accurately account for the presence of background loci if *a, b,* and *g* are set somewhat higher than their mean values in the general population (perhaps reflecting increased DD, Dd, and dd penetrance in affected sib-pair families caused by a higher frequency in such families of disease-predisposing background alleles). The influence of polygenic background on TDT and ASP power is a topic that merits further investigation.

Concluding Remarks

In conclusion, my primary purpose is to highlight the usefulness of the equations for P_s, P_t, *H/F,* and their singleton-family counterparts (P_t*, *H*/F**) for understanding and comparing the properties of the TDT and the ASP test across a broad range of disease models and possible relationships between markers and linked disease loci. As I have demonstrated, these equations yield very accurate power estimates. Furthermore the central equations (P_s and P_t in fig. 1) exhibit symmetries and partition the contributions of basic genetic parameters, thus facilitating comparison and understanding of the TDT and the ASP test (McGinnis 1998). Methods 1 and 2 also have the additional appeal of being conceptually simple; Method 2, for example, involves a single binomial distribution with number of "trials" (2*NH*/*F* or 4*NH*/*F*) determined by the number of heterozygous parents (2*NH*/*F*) and probability of "success" equal to P_s or P_t. In view of the increasing importance of SNPs (Collins 1997) and the growing interest in linkage and disease association studies (Risch and Merikangas 1996; Schaid 1998), I hope these equations prove useful.

Acknowledgments

I thank Richard Spielman and Warren Ewens for insightful comments about figure 2. I wish to dedicate this paper to my father.

Appendix

General equations for singleton families are shown here as derived in McGinnis (1998). The equations assume (1) random ascertainment of nuclear families with at least one affected child and (2) linkage between a biallelic marker (alleles A and B) and a biallelic disease locus (predisposing allele D, "protective" allele d). With minor modifications, these equations also describe markers that are multiallelic (McGinnis 1998). Variables in the equations are as follows: *a, b,* and *g* are penetrances of the DD, Dd, and dd genotypes, respectively; c_1, c_2, c_3, and c_4 are population frequencies of AD, Ad, BD, and Bd haplotypes, respectively; *p* is the population frequency of disease allele D; and θ is the recombination fraction between marker locus and disease locus. As explained in footnote d of table 1, the expressions given below for P_{t}^{*} and $H^*\big/F^*$are used to calculate the number of singleton families required by the TDT to achieve a given power.

P_{t}^{*} is the singleton-family counterpart of P_{t}. It is the probability that marker allele A was transmitted to the affected child by a randomly ascertained parent who is informative (A/B) at the biallelic marker:

$$\mathrm{P}_{\mathrm{t}}^{*} = 0.5 + (1-2\theta)\left(\frac{c_1c_4 - c_2c_3}{H^*}\right)\left[p^2\frac{(a-b)}{2} + 2p(1-p)\frac{(a-g)}{4} + (1-p)^2\frac{(b-g)}{2}\right] = 0.5 + (L_{\mathrm{t}}^{*})(M_{\mathrm{t}}^{*})(R_{\mathrm{t}}^{*})\ .$$

$H^*\big/F^*$is the proportion of randomly ascertained parents expected to be heterozygous (A/B) at the biallelic marker. The expressions for H^* and F^* are

$$H^* = 2c_1c_3[p(a-b)+b]+2(c_1c_4+c_2c_3)\left[\frac{p(a-g)+b+g}{2}\right] + 2c_2c_4[p(b-g)+g]$$

and

$$F^* = p^2a + 2p(1-p)b + (1-p)^2g\ .$$

References

Camp NJ (1997) Genomewide transmission/disequilibrium testing—consideration of the genotypic relative risks at disease loci. Am J Hum Genet 61:1424–1430

Camp NJ (1999) Genomewide transmission/disequilibrium testing: a correction. Am J Hum Genet 64:1485–1487

Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Darvasi A, McGinnis RE (1998) Genetic dissection of complex traits in a high-density SNP map era. Am J Hum Genet Suppl 63:A229

Knapp M (1999) A note on power approximations for the transmission/disequilibrium test. Am J Hum Genet 64: 1177–1185

Knapp M, Seuchter SA, Baur MP (1994) Linkage analysis in nuclear families: Optimality criteria for affected sib-pair tests. Hum Hered 44:37–43

McGinnis RE (1998) Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT). Ann Hum Genet 62:159–179

Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6:127–130

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 8:1273–1288

Schaid DJ (1998) Transmission disequilibrium, family controls, and great expectations. Am J Hum Genet 63:935–941

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 (1998) A sibship test for linkage is the presence of association: the sib transmission/disequilibrium. 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

Tu IP, Whittemore AS (1999) Power of association and linkage tests when the disease alleles are unobserved. Am J Hum Genet 64:641–649

Whittaker JC, Lewis CM (1999) Power comparisons of the transmission/disequilibrium test and sib-transmission/disequilibrium test statistics. Am J Hum Genet 65:578–580