A theoretical treatment of interval mapping of a disease gene using transmission disequilibrium tests

P Narain

29278 Glen Oaks Blvd. W, Farmington Hills, MI 48334-2932, USA

(Email, narainprem@hotmail.com)

The genetic basis of the transmission disequilibrium test (TDT) for two-marker loci is explored from first principles. In this case, parents doubly heterozygous for a given haplotype at the pair of marker loci that are each in linkage disequilibrium with the disease gene with the further possibility of a second-order linkage disequilibrium are considered. The number of times such parents transmit the given haplotype to their affected offspring is counted and compared with the frequencies of haplotypes that are not transmitted. This is done separately for the coupling and repulsion phases of doubly heterozygous genotypes. Expectations of the counts for each of the sixteen cells possible with four-marker gametic types (transmitted vs not transmitted) are derived. Based on a test of symmetry in a square 4 x 4 contingency table, chi-square tests are proposed for the null hypothesis of no linkage between the markers and the disease gene. The power of the tests is discussed in terms of the corresponding non-centrality parameters for the alternative hypothesis that both the markers are linked with the disease locus. The results indicate that the power increases with the decrease in recombination probability and that it is higher for a lower frequency of the disease gene. Taking a pair of markers in an interval for exploring the linkage with the disease gene seems to be more informative than the single-marker case since the values of the non-centrality parameters tend to be consistently higher than their counterparts in the single-marker case. Limitations of the proposed test are also discussed.

[Narain P 2007 A theoretical treatment on interval mapping of a disease gene using transmission disequilibrium tests; *J. Biosci.* **32** 1317–1324]

1. Introduction

Studies on the genetics of complex diseases such as diabetes, coronary artery disease, schizophrenia, various types of cancer, obesity, alcoholism, Alzheimer disease, etc. are at the frontier of research activity in human genetics, which received a new impetus with the completion of the Human Genome Project at the turn of the century. Such diseases are determined by multiple genetic and environmental factors as well as the interactions between them. Association studies that involve linkage disequilibrium (LD) between markers and genes underlying such traits are being undertaken in different parts of the world. The key idea is that a disease mutation assumed to have arisen once on the ancestral haplotype of a single chromosome in the past history of the population of interest is passed on from generation to generation together with markers at tightly linked loci resulting in LD. The usual method adopted in such studies is that of case–control analysis wherein genotype or allele frequencies of candidate genes are compared in unrelated cases and controls. However, if the population is composed of a recent admixture of different ethnic groups that differ in marker allele as well as disease frequencies, spurious associations may result between the marker genotypes and the disease traits (Lander and Schork 1994). Family-based association methods such as the transmission/disequilibrium test (TDT), introduced by Spielman *et al* (1993) can circumvent such problems. This test detects linkage between the marker and the disease gene in the presence of LD between the two loci. A recent review of this test is given in Ewens and Spielman (2003).

When several adjacent marker loci are used for screening, one can examine each locus individually as in the work of Devlin and Risch (1995) and make some correction for

Keywords. Coupling and repulsion phases; disease genes; interval mapping; non-centrality parameter; transmission disequilibrium

Abbreviations used: EATDT, exhaustive allelic transmission disequilibrium tests; GWA, genomewide association; LD, linkage disequilibrium; SNP, single nucleotide polymorphism; S-TDT, sib TDT; TDT, transmission disequilibrium test

multiple testing. As this approach ignores the possible dependence among the two or more marker loci, we may lose information on linkage by conducting single-marker analysis. Several papers such as those of Clayton (1999), Clayton and Jones (1999), Zhao *et al* (2000), and Dudbridge *et al* (2000) consider multiple markers simultaneously. But their approaches have encountered one or other problem such as the discarding of families with ambiguous haplotypes, assumption of no recombination among the markers under study, method not being robust to population stratification, and related issues.

In recent times, the advent of modern genotyping technology has enabled identification of a very large number of single nucleotide polymorphisms (SNPs), providing databases of about 9 million out of the posited 10–13 million common SNPs in the human genome (International HapMap Consortium 2005). Patterns of correlations among them (LD) have been catalogued in several populations. With their help, genomewide association (GWA) studies to identify genetic variants for complex disease traits are now being undertaken using different methods. A family-based association method in the form of exhaustive allelic transmission disequilibrium tests (EATDT) has been advocated by Lin *et al* (2004). This method uses haplotype information after phase reconstruction by searching all alleles – individual SNPs as well as continuous haplotypes of all lengths – from the input sequence data of trios to find the set yielding the lowest TDT *P*-values. It utilizes heterozygous transmissions and non-transmissions for a specific allele in a given window from parent to an affected offspring via a computer algorithm. However, for a pair of markers, it does not distinguish between single and double heterozygotes and uses the usual 2 x 2 McNemar table.

We propose here an approach by which we can study the putative disease gene at any given location on the chromosome by considering only a pair of flanking markers around it rather than the whole set of markers – a sort of *interval mapping* introduced in the literature by Lander and Botstein (1989) for quantitative characters. By choosing different gene locations throughout the length of the chromosome, the behaviour of the concerned statistics can pinpoint the optimum location of the disease gene. The TDT, with two loci data on parents and offspring, then needs to be carried out for the association tests involving only the first- and second-order association parameters for which the necessary theory does not seem to be available in the literature. Once linkage between the disease gene and the two-marker haplotypes is established, usual likelihood-based methods could be employed to develop statistics for estimating the possible location of the disease gene.

We therefore develop, in this paper, a theory of TDT with two-marker loci from the first principles and derive the necessary tests and their powers for the case when the linkage phase is known without error and under the assumption of known haplotype information on parents and affected offspring. Before doing so, however, we recapitulate the known results for the TDT with a single-marker case in what follows.

2. TDT with a single marker

Let *A-a* denote a marker locus that is to be evaluated in relation to a disease trait locus *D-d* with a recombination probability between them of r_1. We assume that the random mating population under consideration is in a steady state with a constant population size, i.e. in equilibrium between the effects of genetic drift and recombination. This means that the time that has passed since the disease mutant was introduced is of the same order as the effective population size.

We consider recessive disease genes so that allele *D* is dominant over allele *d*. We further assume that only individuals with genotype *dd* are affected by the disease whereas the homozygous genotype *DD* and the heterozygous genotype *Dd* are unaffected by the disease and therefore categorized as normal individuals. The population then consists of two types of individuals, affected and normal. We take an affected individual whose genotype at the disease locus is thus known as *dd* as well as the two parents of this individual whose genotypes at the disease locus are not known but both must have contributed the allele *d* to their child. We treat this as one nuclear family, a trio, and suppose we have *N* trios in our sample. Now we genotype these trios for the marker *A-a* and examine whether a parent heterozygous at the marker locus, i.e. having the genotype *Aa* has transmitted to the child a marker allele along with the allele *d* or not. When the disease and marker loci are neither linked nor show LD, i.e. there is no association between them, the number of times a marker allele is transmitted or not transmitted to the child along with the disease allele is expected to be the same. This is the rationale behind the TDT.

The TDT thus compares the frequencies of the marker alleles *A* and *a*, transmitted from the parents *Aa* to offspring *dd* with those of alleles that are not transmitted and hence is based on a 2 x 2 table containing frequencies for the marker alleles transmitted (*T*) or not transmitted (*NT*) from parents to affected offspring in a sample of 2*N* parents of *N* affected offspring as given in table 1.

The expected values of the counts in table 1 depend on the conditional probabilities with which a parent transmits one marker allele and not the other, given that it transmits a *d* allele. In order to determine them we need to consider the population genetics model of a two-loci system as discussed below.

Let the allelic frequencies of the marker and disease trait loci be denoted by p_1, $q_1=1-p_1$ and p_d, $q_d=1-p_d$, respectively.

Table 1. Observed counts for transmitted and non-transmitted marker alleles *A* and *a* among 2*N* parents of *N* affected offspring

Non-transmitted (*NT*) allele	Transmitted (*T*) allele		
	A	*a*	Total
A	*a*	*b*	*(a+b)*
a	*c*	*d*	*(c+d)*
Total	*(a+c)*	*(b+d)*	*2N*

There are ten genotypes, taking into account the two phases of linkage with respect to the two loci *A-a* and *D-d.* There are four possible two-locus haplotypes *AD, Ad, aD* and *ad* with frequencies, say, p_{AD}, p_{Ad}, p_{aD} and p_{ad}, respectively, when the genotypes mate at random. Then the linkage disequilibrium coefficient between the two loci, denoted by D_{1d}, is defined as the deviation of the haplotype frequency from its expected frequency under equilibrium, which is simply the product of the corresponding gene frequencies. For example, if we take the haplotype *AD* we have

$$D_{1d} = p_{AD} - p_1 p_d \tag{1}$$

The disequilibrium coefficient can also be expressed entirely in terms of the four haplotype frequencies (Narain 1990), as

$$D_{1d} = p_{AD} p_{ad} - p_{Ad} p_{aD} \tag{2}$$

This coefficient measures *allelic association* that could be either due to linkage for loci on the same chromosome or just association without any linkage for loci on non-homologous chromosomes showing independent segregation at meiosis.

Due to conditioning for the recessive genotype *dd*, we have to consider the probability of only those mating types that result in the formation of gametes *Ad* and *ad*. The total frequency of these gametes being p_d, the relevant probabilities need to be divided by p_d. From a table of such probabilities, one can determine the required conditional probabilities of transmission of gametes. For instance, for the expected value of the count *b*, we determine the probability that, given that the disease trait allele *d* is transmitted, the heterozygous parent *Aa* transmits the marker allele *a* and not the other allele *A*. This is written symbolically as *Pr. [T: a, NT: A / Aa, T: d]* and is given by

$$\begin{aligned} E(b) &= 2N \Pr.\,[T: a, NT: A / Aa, T: d] \\ &= 2N\, p_d^{-1}\, [p_{Ad} p_{ad} + r_1 p_{Ad} p_{aD} + (1-r_1)\, p_{AD} p_{ad}] \\ &= 2N\, [p_1 q_1 + (r_1 - p_1)\, D_{1d} / p_d] \end{aligned}$$

In a similar manner, we get the expectations of *a*, *c* and *d*. All these expectations are given in Appendix I. From these expectations, we get

$$E(c\text{-}b) = 2N\, [(1\text{-}2r_1) D_{1d} / p_d] \tag{3}$$

$$E(c+b) = 2N\, [2p_1 q_1 + (q_1\text{-}p_1) D_{1d} / p_d] \tag{4}$$

This shows that the expectation of the difference (*c–b*) would be zero if either $r_1 = 1/2$ or $D_{1d} = 0$, which indicates either no linkage or no disequilibrium. In that case the expectations of both *c* and *b* will be the same and equal to half. The statistic for TDT is therefore

$$\chi^2 = (c-b)^2/(c+b) \tag{5}$$

which follows a chi-square distribution with one degree of freedom (df) and therefore can be used to test whether there is an association between marker *A* and the trait gene *d*. It may be noted that (*c*+*b*) provides an estimate of the variance of (*c–b*). Alternatively, the first and second marker allele of each parent can be matched in four possible types as transmitted–transmitted, not transmitted–not transmitted, transmitted–not transmitted and not transmitted–transmitted to correspond to entries *a*, *d*, *c*, and *b* respectively in table 1, each parent being counted twice. Because of the matching, the observations tend to be dependent and one has to use the test for comparing correlated proportions. This leads to McNemar test, which is the same as that given by (5). In fact, in this we test the hypothesis of *marginal homogeneity.* It implies symmetry across the main diagonal so that hypotheses of marginal homogeneity and symmetry are equivalent.

Under the alternative hypothesis that there is linkage between the marker and the disease gene, given that there is linkage disequilibrium, the chi-square statistic given by (5) follows a non-central $\chi^2(1, \lambda)$ distribution with 1 df and non-centrality parameter λ given by

$$\begin{aligned} \lambda &= [E(c) - E(b)]^2 / [E(c) + E(b)] \\ &= 2N\, [(1 - 2r_1)^2 D_{1d}^{\,2}] / p_d\, [2p_1 q_1 p_d \\ &\quad + D_{1d}\, (1\text{-}\, 2p_1)] \end{aligned} \tag{6}$$

The power of the test is then the probability that the deviate from $\chi^2(1, \lambda)$ is greater than or equal to $\chi^2(\alpha)$, the critical value of χ^2 to reject the null hypothesis at significance level α. Liu (1997) gives the power of this TDT for several values of *N*, p_1, p_d, r_1 and D_{1d}. The power increases with increase in D_{1d} but with decrease in r_1. It is high when p_d is lower. The frequency p_1 has, however, a small effect on the power. It also increases with an increase in *N*.

It may be seen that λ will be strictly zero when either $r_1 = 1/2$ or $D_{1d} = 0$, the values under the null hypothesis, in which case the chi-square follows a central χ^2 distribution with 1 df. Values of λ, therefore, under different values of the five parameters, *N*, r_1, p_1, p_d and D_{1d} reflect the power of the TDT with a single marker. We give these values in table 2 below for N=200 and D_{1d}=0.1 when r_1 varies between 0.45 and 0.01 for each of the two values of p_d and p_1.

From table A.3 in Weir (1996), we find that for the significance value of 0.05 and 1 df, the power is 0.90

Table 2. Values of non-centrality parameters for different recombination probabilities between the marker *A-a* and the disease locus for two combinations of gene frequencies when *N=20*

r_1	$p(d)=0.5, p(1)=0.2$	$p(d)=p(1)=0.2$
0.45	0.36	1.61
0.30	5.82	25.01
0.20	13.09	58.06
0.10	23.27	103.22
0.04	30.78	136.52
0.01	34.92	154.90

and 0.99 for non-centrality parameters of 10.5 and 18.4, respectively. As such, when r_1 is 0.2, the power of the test for $p_d = 0.5$ and $p_1 = 0.2$ would be greater than 0.90 but it would be even greater than 0.99 for $p_d = p_1 = 0.2$, indicating thereby that the power is high when the value of p_d is small. Thus, we can compare the power of the tests in terms of non-centrality parameters.

When the marker is at the disease gene locus itself, $r_1 = 0$, $p_1 = p_d$, and $D_{1d} = p_d q_d$, giving

$$\lambda = 2N q_d \tag{7}$$

3. TDT with two-marker loci

Now we consider another marker locus *B-b* tightly linked with *A-a* with a small probability of recombination between them as *r* and the trait locus *D-d* is located between them with an r_1 recombination probability between *A* and *D*, and r_2 recombination probability between *B* and *D*. Then, given *r* and assuming no interference, r_2 can be expressed in terms of r_1 using the relation $r = r_1 + r_2 - 2 r_1 r_2$ so that there is only one unknown to handle. We now have to consider the frequencies of the marker gametes *AB, Ab*, *aB* and *ab* transmitted from the doubly heterozygous parents (*AaBb*), which could be of two types (*AB/ab or Ab/aB)* depending upon the phase of the linkage, to affected offspring having genotypes *dd* with those of the gametes not transmitted. It is based on a 4 x 4 table containing frequencies of the marker gametes transmitted (*T*) or not transmitted (*NT*) from parents to affected offspring in a sample of 2*N* parents of *N* affected offspring as given in table 3.

The expected values of the counts in table 3 depend upon the conditional probabilities with which a parent transmits one marker haplotype and not the other, given that a *d* allele is transmitted. In order to determine them we need to consider the population genetics model of a three-loci system as discussed below.

Let the gene frequencies of the two markers *A-a* and *B-b* be denoted, respectively, by p_1, $q_1=1-p_1$ and p_2, $q_2=1-p_2$ and

Table 3. Observed counts for transmitted (T) and non-transmitted (NT) marker gametes *AB*, *Ab*, *aB*, and *ab* among 2*N* parents of *N* affected offspring

Non-transmitted gamete (NT)	Transmitted gamete (T)				
	AB	Ab	aB	ab	Total
AB	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
Ab	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
aB	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
ab	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	*2N*

that of the disease locus, as before, by p_d, $q_d=1-p_d$. There are now 36 genotypes with respect to the three loci, *A-a, B-b* and *D-d*, there being 12 double heterozygotes and 4 triple heterozygotes with possible linkage phases. We get eight three-locus haplotypes *ABD, ABd, AbD, Abd, aBD, aBd, abD* and *abd* with respective frequencies, say p_{ABD}, p_{ABd}, p_{AbD}, p_{Abd}, p_{aBD}, p_{aBd}, p_{abD} and p_{abd} when the genotypes mate at random. Let the pair-wise (first-order) LD parameters between *A-a* and *B-b* be denoted by D_{12}, between *A-a* and *D-d*, as before, by D_{1d}, and between *B-b* and *D-d* by D_{2d}. The three-locus (second-order) LD parameter between (*A*, *B)* and *D*, denoted by D_{12d} is defined by

$$D_{12d} = p_{ABD} - p_1 D_{2d} - p_2 D_{1d} - p_d D_{12} - p_1 p_2 p_d. \tag{8}$$

It may be noted that to describe the association among alleles at three loci, the effects of two-locus disequilibrium are removed by subtraction (Bennett 1954). This disequilibrium coefficient can also be expressed entirely in terms of the eight three-locus haplotype frequencies as

$$D_{12d} = [\, p_{ABD} p_{abd} - p_{ABd} p_{abD}] - [\, p_{AbD} p_{aBd} - p_{Abd} p_{aBD}]. \tag{9}$$

Due to conditioning for the recessive genotype *dd*, we have to consider the probability of only those mating types that result in the formation of gametes *AdB*, *Adb*, *adB* and *adb*. The total frequency of these gametes being p_d, the relevant probabilities need to be divided by p_d. From a table of such probabilities, one can determine the required conditional probabilities for transmission of gametes. For instance, for the expected value of the count n_{41}, we determine the probability that, given that the disease trait allele *d* is transmitted, the doubly heterozygous parent *AaBb*, in the coupling phase, transmits the marker gamete *AB* and not the other gamete *ab*. This is written symbolically as

Pr. [T: AB, NT: ab / AB / ab, T: d] and is given by

$$E(n_{41}) = Pr.\,[T : AB, NT: ab / AB / ab, T: d]$$
$$= 2N p_d^{-1} [\{(1-r)\, p_{AdB} p_{adb} + r\, p_{Adb} p_{adB}\} + (1-r_1)(1-r_2)\, p_{AdB} p_{aDb}$$
$$+ r_2\,(1-r_1)\, p_{Adb}\, p_{aDB} + r_1\, r_2\, p_{ADB}\, p_{adb} + r_1\,(1-r_2)\, p_{ADb}\, p_{adB}]$$
$$= 2N [\{p_1 p_2 q_1 q_2 + (p_1 p_2 + q_1 q_2 - r)\, D_{12} + D_{12}^2\} + \{-p_2 q_2 (p_1 - r_1) D_{1d} - p_1 q_1 (p_2 - r_2) D_{2d} + (p_1 - r_1)(p_2 - r_2) D_{12d} - (q_2 - r_2) D_{12} D_{1d} - (q_1 - r_1) D_{12} D_{2d} + D_{12} D_{12d}\}/p_d].$$

In a similar manner we get the expectations of the other counts of table 3. However, to conserve space, we give, in Appendix II, only the expectations of the 4 cell counts (n_{14}, n_{23}, n_{32}, n_{41}) that are relevant to our study.

It may be verified that by pooling appropriate cell counts in table 2, we can get the expectations of cell counts in table 1. For instance, if we pool over the different levels of the locus *B-b,* in the third and fourth rows and the first and second columns, we get *E(c)*. Similarly, we can get the expectations of cell counts in the table (not shown) for the TDT applied to the locus *B-b* by pooling over the different levels of the locus *A-a*.

3.1 *Various tests*

In TDT the 4 entries in the diagonal do not contribute to the test since these pertain to doubly homozygous parents. Of the 12 remaining entries, the 6 above the diagonal are matched with the 6 below the diagonal. Of the 6 pairs so formed, 4 pertain to the singly heterozygous parents at each of the two markers (there being two possible homozygotes at the other marker locus) and 2 to the doubly heterozygous parents (one in the coupling phase and the other in the repulsion phase). When there is no association between the markers and the disease gene making all the Ds zero or when the markers and the disease gene are not linked, i.e. $r_1 = r_2 = ½$ so that $r = ½$ also, the expectation of the matched entries below and above the diagonal are same. Symbolically, $E(n_{ij}) = E(n_{ji})$ for $i<j$, $i,j=1,2,3,4$. The 4 x 4 table therefore satisfies the condition of *symmetry*. In this case, marginal homogeneity occurs since the expectation of marginal totals $E(n_{i.})$ and $E(n_{.j})$ become the same. But here the symmetry is not equivalent to marginal homogeneity as is the case in the 2 x 2 table for the TDT with a single marker. Marginal homogeneity can occur *without* symmetry. Therefore, we need here a test for symmetry.

Following Bowker (1948), the test of symmetry in the square 4 x 4 contingency can be performed with the help of the statistic

$$\chi^2 = \Sigma\,\Sigma\, (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji}) \text{ for } 2 \le i \le 4,\ 1 \le j \le (i-1).$$

This statistic follows a chi-square distribution with 6 df. This is a composite statistic testing for the linkages between the disease gene and either of the two markers, either singly or jointly, on the condition that all the pair-wise disequilibria as well as second-order disequilibrium exist. It can be partitioned into six components corresponding to the six 2 x 2 contingency tables formed by conditioning the data *only* for the given table. For the table with entries, n_{ii}, n_{ij}, n_{ji}, n_{jj}, the chi-square with 1 df would be

$$\chi^2 = (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji}) \text{ for } 2 \le i \le 4,\ 1 \le j \le (i-1). \quad (10)$$

The expectations of the difference $(n_{ij} - n_{ji})$ and the sum $(n_{ij} + n_{ji})$ for different linkage tests involving double heterozygotes in the coupling and repulsion phases would be as given in Appendix III (the expectations for other linkage tests are not given for the sake of brevity). On the null hypothesis of no linkage between the disease gene and the markers, each of the expectations of the *difference*, given in Appendix III, would be zero. For data conditioned in a given 2 x 2 table, the null hypothesis for binary matched pairs is

$$H_0: E(n_{ij}) = E(n_{ji}) \text{ or } E(n_{ij})/E(n_{ij} + n_{ji}) = 0.5.$$

Under H_0, n_{ij} has a binomial distribution $(n_{ij}^*, 1/2)$ with $n_{ij}^* = n_{ij} + n_{ji}$ that, for large samples, is approximately normal with mean $(1/2)n_{ij}^*$ and variance $n_{ij}^*(1/2)(1/2)$. The standardized normal test statistic is then

$$[n_{ij} - (1/2)n_{ij}^*]/[n_{ij}^*(1/2)(1/2)]^{1/2} = (n_{ij} - n_{ji})/(n_{ij} + n_{ji})^{1/2}$$

leading to a χ^2 test statistic with 1 df as already given above.

For interval mapping of the disease gene, we have to consider a situation when the parental genotype is doubly heterozygous. However, double heterozygotes of the type *AaBb* can have two different kinds of allelic arrangements on the homologous pairs of chromosomes, namely *AB/ab* and *Ab/aB*. The former, where both the dominant genes are located on the same chromosome, is said to have the linkage in the *coupling phase* while the latter, where one dominant gene is on the first member and the other dominant gene is on the second member of the pair of chromosomes, is said to have the linkage in the *repulsion phase* (Narain 1990). Segregation of the heterozygous parents in the two phases is normally done with the help of progeny tests or pedigree analysis. In the latter case, pedigrees are often ascertained through a child affected by the disease. For rare diseases, this means we have families with heterozygous parents. If we have marker information on both the parents and the child for each family ascertained, we may be able to determine the phase of the double heterozygotes from such data as shown in a study on cystic fibrosis by Weir (1989).

There would, thus, be two chi-square tests, each on 1 df, for testing the relevant null hypothesis of no linkage, given by:

$$\chi_1^2 (AB/ab) = (n_{14} - n_{41})^2 / (n_{14} + n_{41}) \quad (11)$$

$$\chi_1^2 (Ab/aB) = (n_{23} - n_{32})^2 / (n_{23} + n_{32}). \quad (12)$$

In the first case, the statistic tests whether the marker gamete *AB* is linked with the disease gene when the parent is in the coupling phase, whereas the second statistic tests whether the marker gamete *Ab* is linked with the disease gene when the parent is in the repulsion phase, on the assumption that both first- as well as second-order disequilibrium exist. In either of the cases, it tests whether the disease gene is in the given interval of the two markers, the null hypothesis being that the disease gene is *not* in the interval.

3.2 *Power of the tests*

Under the alternative hypothesis that the gene does lie in the interval, namely, that the disease gene is linked with *both* the markers, given that the disequilibrium coefficients are non-zero, the chi-squares, given by (11) and (12), follow approximately a non-central chi-square distribution with 1 df and with non-centrality parameters λ_1 and λ_2 given respectively by

$$\begin{aligned}\lambda_1 &= [E(n_{14}) - E(n_{41})]^2 / [E(n_{14}) + E(n_{41})] \\ &= 4N^2 [(1 - 2r_1) \{p_2 q_2 D_{1d} + D_{12} D_{2d} \\ &\quad + (1/2)(q_2 - p_2) D_{12d}\} + (1 - 2r_2)\{p_1 q_1 D_{2d} + D_{12} D_{1d} \\ &\quad + (1/2)(q_1 - p_1) D_{12d}\}]^2 / S_1 \end{aligned} \quad (13)$$

$$\begin{aligned}\lambda_2 &= [E(n_{23}) - E(n_{32})]^2 / [E(n_{23}) + E(n_{32})] \\ &= 4N^2 [(1 - 2r_1)\{p_2 q_2 D_{1d} + D_{12} D_{2d} \\ &\quad + (1/2)(q_2 - p_2) D_{12d}\} - (1 - 2r_2)\{p_1 q_1 D_{2d} \\ &\quad + D_{12} D_{1d} + (1/2)(q_1 - p_1) D_{12d}\}]^2 / S_2 \end{aligned} \quad (14)$$

where S_1 and S_2 are given by (AIII.2) and (AIII.4), respectively.

The alternative hypothesis here is that *both* the markers are linked with the disease locus, i.e. $r_1 \neq 1/2$ and $r_2 \neq 1/2$. The situation when only one of the markers is linked with the disease locus, i.e. say $r_1 \neq 1/2$ but $r_2 = 1/2$, is not tenable since in that case $r = r_1 + r_2 - 2r_1 r_2 = 1/2$, which violates the assumption of tightly linked markers.

It may, however, be noted that, in the above discussion, the non-centrality parameters are determined *approximately*. Their computation therefore does not seem to give an exact answer. Also, when *r* is small, rejection of the symmetry test always means that both r_1 and r_2 are not equal to *1/2* and one of the LD coefficients is not zero. This shows that the information obtained by the proposed interval mapping may, under some scenarios, be compromised.

If *a priori* information indicates that the two markers are likely to have the *same* recombination probability with the disease locus, i.e. $r_1 = r_2$, the non-centrality parameter is simplified in the two cases to

$$\lambda_1^* = 4N^2 p_d^{-2} (1-2r_1)^2 (C_1 + C_2)^2 / S_1^* \quad (15)$$

$$\lambda_2^* = 4N^2 p_d^{-2} (1-2r_1)^2 (C_1 - C_2)^2 / S_2^* \quad (16)$$

where

$$C_1 = p_2 q_2 D_{1d} + D_{12} D_{2d} + (1/2)(q_2 - p_2) D_{12d}$$

$$C_2 = p_1 q_1 D_{2d} + D_{12} D_{1d} + (1/2)(q_1 - p_1) D_{12d}$$

S_1^* and S_2^* being given by (AIII.2) and (AIII.4), respectively, with $r = 2r_1 (1-r_1)$. It is seen that in this case the λs will become zero when either $r_1 = 1/2$, i.e. no linkage between the markers and the disease locus or the *D*s, the disequilibrium coefficients in *C*s become zero.

It may be seen that, in general, λs will be strictly zero when either $r_1 = 1/2$, and $r_2 = 1/2$ or else $D_{1d} = D_{2d} = D_{12d} = 0$, the values under the null hypothesis, in which case the chi-squares follow a central χ^2 distribution with 1 df. Values of λs, therefore, under different values of the ten parameters N, p_1, p_2, D_{12}, p_d, D_{1d}, D_{2d}, D_{12d}, r_1 and r_2 reflect the power of the TDT with two markers. We give in table 4 below the values of λ_1 for $N = 200$, $D_{1d} = D_{2d} = D_{12d} = 0.1$, $D_{12} = 0.1$, and $r_2 = 0.05$ for different values of r_1 for each of the two combinations of gene frequencies.

The results indicate that the power increases with a decrease in the values of r_1 and that power is high with smaller values of p_d. Most of the results, if not all, particularly the effect of disequilibrium coefficients true for the single-marker case are, therefore, carried over to the two-markers case. Compared with the values given in table 2, the values of the non-centrality parameter are consistently higher in table 4, indicating the benefits of having information about linkage from the second marker.

When one of the markers, say *A-a*, is at the disease gene locus itself, $r_1 = 0$, $p_1 = p_d$, $q_1 = q_d$, $D_{1d} = p_d q_d$, $D_{12} = D_{2d}$, $D_{12d} = (q_d - p_d) D_{2d}$ and we get

$$\begin{aligned}\lambda_1 &= 2N [p_2 q_2 p_d q_d + \{(1 - 2r_2) \\ &\quad + (q_2 - p_2)(q_d - p_d)\} D_{2d} / 2 + D_{2d}^2] \end{aligned} \quad (17)$$

$$\begin{aligned}\lambda_2 &= 2N [p_2 q_2 p_d q_d - \{(1 - 2r_2) \\ &\quad - (q_2 - p_2)(q_d - p_d)\} D_{2d} / 2 + D_{2d}^2]. \end{aligned} \quad (18)$$

Table 4. Values of the non-centrality parameter λ_1 for different recombination probabilities between the marker *A-a* and the disease locus for each of the two combinations of the gene frequencies when $N = 200$

r_1	$p(d)=0.5, p(1)=p(2)=0.2$	$p(d)=p(1)=p(2)=0.2$
0.45	23.27	85.99
0.30	31.45	115.41
0.20	36.94	135.12
0.10	42.45	154.87
0.04	45.76	166.74
0.01	47.42	172.67

When the other marker, *B-b*, is also at the disease gene locus, i.e. at the other marker *A-a* itself, $r_2 = 0$, $p_2 = p_d$, $q_2 = q_d$, $D_{2d} = p_d q_d$, and we get, for λ_1 only as coupling phase is only possible in such a case,

$$\lambda_1 = 2N q_d,$$

the same as (7), as it should since the whole system now reduces to a single locus case.

4. Discussion

The theory of TDT with a single-marker locus has been successfully extended to two-linked marker loci with first- and second-order disequilibria. A new test statistic for testing linkage with 1 df based on the test for symmetry has been proposed. It uses data only on doubly heterozygous parents who transmit the given haplotype to their affected offspring. The power of the test has also been discussed in terms of non-centrality parameters. Further extension of TDT to three or more markers is quite involved due to a commensurate increase in the parameters of disequilibrium coefficients of various orders, besides the increase in parameters pertaining to gene frequency and linkage. However, for the purpose of interval mapping of the disease gene, this is not required. We need consider only a *pair* of flanking markers around the putative disease gene.

The major assumption in this study is that two-loci haplotypes are known in parents. The traditional method to determine haplotypes is either pedigree analysis or molecular haplotyping. Both these methods require a lot of work, of either collecting a large number of pedigree members or in performing costly laboratory tests. Due to these limitations, the current trend is to use appropriate statistical methods and develop computer algorithms to infer the phase of the linkage from the genotypes and thus reconstruct the haplotypes. The methods include a parsimony approach given by Clark (1990), a maximum-likelihood method via an expectation–maximization (EM) algorithm (Excoffier and Slatkin 1995), and a Bayesian approach based on priors from population genetics (Stephens *et al* 2001). The inferences in these cases are, however, drawn from unrelated individuals and are therefore not applicable to the TDT as presented in this paper. Marchini *et al* (2006) extended their algorithms for phase inference, which can handle data on related individuals such as father–mother–child trios. This could be useful for data collected on nuclear families such as the TDT with two-linked marker loci considered in this paper. In fact, EATDT – a TDT type test used in the study by Lin *et al* (2004) – makes use of this approach for phase determination before using haplotype information for the test.

Another limitation of this study is that when the disease under study has a late age of onset, the parental marker genotypes may not be available at all. In this situation, the missing parental genotypes could be reconstructed from the genotypes of their offspring and treated as if they have been typed (Spielman and Ewens 1996). However, a better way would be to generalize the test proposed in this paper to the 'sib TDT' or S-TDT type procedure discussed in Spielman and Ewens (1998), where data consist of marker genotypes of the offspring only, both affected and unaffected, for each family.

Acknowledgements

This work was supported by the Indian National Science Academy, New Delhi under their programme 'INSA Honorary Scientist'. The author is grateful to the referees for valuable suggestions, which improved the quality of the paper.

Appendix I

Expectations of cell counts in the single-marker case

$$E(a) = 2N[p_1^2 + p_1 D_{1d}/p_d] \quad \text{(AI.1)}$$

$$E(b) = 2N[p_1 q_1 + (r_1 - p_1) D_{1d}/p_d] \quad \text{(AI.2)}$$

$$E(c) = 2N[p_1 q_1 + (q_1 - r_1) D_{1d}/p_d] \quad \text{(AI.3)}$$

$$E(d) = 2N[q_1^2 - q_1 D_{1d}/p_d] \quad \text{(AI.4)}$$

Appendix II

Expectations of the four relevant cell counts in the two-markers case.

$$E(n_{14}) = 2N \Pr[\text{T: ab, NT: AB, / AB/ab, T: d}]$$
$$= 2N[\{p_1 p_2 q_1 q_2 + (p_1 p_2 + q_1 q_2 - r) D_{12} + D_{12}^2\} + \{p_2 q_2 (q_1 - r_1) D_{1d} + p_1 q_1 (q_2 - r_2) D_{2d} + (q_1 - r_1)(q_2 - r_2) D_{12d} + (p_2 - r_2) D_{12} D_{1d} + (p_1 - r_1) D_{12} D_{2d} + D_{12} D_{12d}\}/p_d] \quad \text{(AII.1)}$$

$$E(n_{23}) = 2N \Pr[\text{T: aB, NT: Ab / Ab/aB, T: d}]$$
$$= 2N[\{p_1 p_2 q_1 q_2 - (p_1 q_2 + q_1 p_2 - r) D_{12} + D_{12}^2\} + \{p_2 q_2 (q_1 - r_1) D_{1d} - p_1 q_1 (p_2 - r_2) D_{2d} - (q_1 - r_1)(p_2 - r_2) D_{12d} - (q_2 - r_2) D_{12} D_{1d} + (p_1 - r_1) D_{12} D_{2d} + D_{12} D_{12d}\}/p_d] \quad \text{(AII.2)}$$

$$E(n_{32}) = 2N \Pr[\text{T: Ab, NT: aB / Ab/aB, T: d}]$$
$$= 2N[\{p_1 p_2 q_1 q_2 - (p_1 q_2 + q_1 p_2 - r) D_{12} + D_{12}^2\} + \{-p_2 q_2 (p_1 - r_1) D_{1d} + p_1 q_1 (q_2 - r_2) D_{2d} - (p_1 - r_1)(q_2 - r_2) D_{12d} + (p_2 - r_2) D_{12} D_{1d} - (q_1 - r_1) D_{12} D_{2d} + D_{12} D_{12d}\}/p_d] \quad \text{(AII.3)}$$

$$E(n_{41}) = 2N \Pr[\text{T: AB, NT: ab / AB/ab, T: d}]$$
$$= 2N[\{p_1 p_2 q_1 q_2 + (p_1 p_2 + q_1 q_2 - r) D_{12} + D_{12}^2\} + \{-p_2 q_2 (p_1 - r_1) D_{1d} - p_1 q_1 (p_2 - r_2) D_{2d} + (p_1 - r_1)(p_2 - r_2) D_{12d} - (q_2 - r_2) D_{12} D_{1d} - (q_1 - r_1) D_{12} D_{2d} + D_{12} D_{12d}\}/p_d] \quad \text{(AII.4)}$$

Appendix III

Expectations of the difference ($n_{ij} - n_{ji}$) and the sum ($n_{ij} + n_{ji}$) for the linkage tests in the two-marker case (linkage between *D-d* and *A-a* as well as *B-b*)

(a) Genotype for the markers in the coupling phase

$$E(n_{14} - n_{41}) = 2N\, p_d^{-1}[(1 - 2r_1)\{p_2q_2\, D_{1d} + D_{12}D_{2d} + (1/2)(q_2 - p_2)\, D_{12d}\} + (1 - 2r_2)\{p_1q_1D_{2d} + D_{12}D_{1d} + (1/2)(q_1 - p_1)\, D_{12d}\}] \quad \text{(AIII.1)}$$

$$E(n_{14} + n_{41}) = 2N\, p_d^{-1}[\, p_d\, \{2p_1\, q_1p_2q_2 + 2(p_1p_2 + q_1q_2 - r)D_{12} + 2D_{12}^2\} + \{p_2q_2(q_1 - p_1)D_{1d} + p_1q_1(q_2 - p_2)D_{2d} + (p_1p_2 + q_1q_2 - r)D_{12d} - (q_2 - p_2)D_{12}D_{1d} - (q_1 - p_1)D_{12}D_{2d} + 2D_{12}D_{12d}\}] \quad \text{(AIII.2)}$$

(b) Genotype for the markers in the repulsion phase

$$E(n_{23} - n_{32}) = 2N\, p_d^{-1}\, [(1 - 2r_1)\{p_2q_2\, D_{1d} + D_{12}D_{2d} + (1/2)(q_2 - p_2)\, D_{12d}\} - (1 - 2r_2)\{p_1q_1\, D_{2d} + D_{12}D_{1d} + (1/2)(q_1 - p_1)\, D_{12d}\}] \quad \text{(AIII.3)}$$

$$E(n_{23} + n_{32}) = 2N\, p_d^{-1}[\, p_d\, \{2p_1\, q_1p_2q_2 - 2(p_1q_2 + q_1p_2 - r)D_{12} + 2D_{12}^2\} + \{p_2q_2(q_1 - p_1)D_{1d} + p_1q_1(q_2 - p_2)D_{2d} - (p_1q_2 + q_1p_2 - r)D_{12d} - (q_2 - p_2)D_{12}D_{1d} - (q_1 - p_1)D_{12}D_{2d} + 2D_{12}D_{12d}\}] \quad \text{(AIII.4)}$$

References

Bennett H J 1954 On the theory of random mating; *Ann. Eugenics* **18** 311–317

Bowker A H 1948 A test for symmetry in contingency tables; *J. Am. Statist. Assoc.* **43** 572–574

Clark AG 1990 Inferences of haplotypes from PCR-amplified samples of diploid populations; *Mol. Biol. Evol.* **7** 111–122

Clayton D 1999 A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission; *Am. J. Hum. Genet.* **65** 1170–1177

Clayton D and Jones H 1999 Transmission/disequilibrium tests for extended marker haplotypes; *Am. J. Hum. Genet*. **65** 1161–1169

Devlin B and Risch N 1995 A comparison of linkage disequilibrium measures for fine-scale mapping; *Genomics* **29** 311–322

Dudbridge F, Koeleman B P C, Todd J A and Clayton D G 2000 Unbiased application of the transmission/disequilibrium test to multilocus haplotypes; *Am. J. Hum. Genet*. **66** 2009–2012

Ewens W J and Spielman R S 2003 The transmission/disequilibrium test; in *Handbook of statistical genetics*, 2nd edition, (eds) D J Balding, M Bishop and C Cannings (New York:John Wiley) pp 961–972

Excoffier L and Slatkin M 1995 Maximum likelihood estimation of molecular haplotype frequencies in a diploid population; *Mol. Biol. Evol*. **12** 921–927

International HapMap Consortium 2005 A haplotype map of the human genome; *Nature* **437** 1299–1320

Lander E S and Botstein D 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps; *Genetics* **121** 185–199

Lander E S and Schork N J 1994 Genetic dissection of complex traits; *Science* **265** 2037–2048

Lin S, Chakravarti A and Cutler D 2004 Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies; *Nat. Genet*. **36** 1181–1188

Liu Ben-Hui 1997 *Statistical genomics – linkage, mapping, and QTL analysis* (New York: CRC Press)

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin Z S, Munro H M, Abecasis G and Donnelly P 2006 A comparison of phasing algorithms for trios and unrelated individuals; *Am. J. Hum. Genet*. **78** 437–450

Narain P 1990 *Statistical genetics* (New York: John Wiley; reprinted 1993 New Delhi: Wiley Eastern Ltd.; Published 1999 New Delhi: New Age International Pvt. Ltd.)

Spielman R S and Ewens W J 1996 The TDT and other family based tests for linkage disequilibrium and association; *Am. J. Hum. Genet*. **59** 983–989

Spielman R S and Ewens W J 1998 A sibship test for linkage in the presence of association: the sib-transmission/disequilibrium test; *Am. J. Hum. Genet*. **62** 450–458

Spielman R S, Mcginnis R E and Ewens W J 1993 Transmission tests for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM); *Am. J. Hum. Genet*. **52** 506–516

Stephens M, Smith N J and Donnelly P 2001 A new statistical method for haplotype reconstruction from population data; *Am. J. Hum. Genet*. **68** 978–989

Weir B S 1989 Locating the cystic fibrosis gene on the basis of linkage disequilibrium with markers?; *in Multipoint mapping and linkage based upon affected pedigree members: genetic analysis workshop 6* (eds) R C Elston, M A Spence, S E Hodge and J W MacCluer (New York: Liss) pp 81–86

Weir B S 1996 *Genetic data analysis II. Methods for discrete population genetic data* (Sunderland, Massachusetts: Sinauer Associates)

Zhao H, Zhang S, Merikangas K R, Trixler M, Wildenauer D B, Sun F and Kidd K K 2000 Transmission/disequilibrium tests using multiple tightly linked markers; *Am. J. Hum. Genet*. **67** 936–946

MS received 23 February 2007; accepted 5 October 2007

*e*Publication: 18 October 2007

Corresponding editor: Partha P Majumder