Proc Indian Natn Sci Acad **80** No. 3 September 2014 pp. 621-632
© *Printed in India*.
DOI: 10.16943/ptinsa/2014/v80i3/55137

Research Paper

Transmission/Disequilibrium Tests

P NARAIN, FNA
INSA Honorary Scientist, B-3/27A, Lawrence Road, Keshav Puram, New Delhi 110 035 and 29278 Glen Oaks Blvd. W., Farmington Hills, MI 48334-2932, USA

(Received on 10 June 2013; Revised on 29 December 2013; Accepted on 23 April 2014)

A general theory of transmission/disequilibrium test (TDT), given the information on an arbitrary number of markers on the two parents and their child, affected by the disease is developed. With k biallelic SNP markers and $N=2^k$ types of possible haplotypes we consider a N x N table of transmission events in which N diagonal elements correspond to homozygous transmissions, not relevant to the problem of testing linkage, and $N(N-1)$ off-diagonal elements resulting in $N(N-1)/2$ matched pairs of transmissions. This gives rise to $N/2$ chi-square tests, corresponding to different phases of the k-tuple heterozygote, each based on 1 d.f. on the null hypothesis of no linkage between the disease-susceptibility gene and all the k markers jointly. The power of the test, in terms of non-centrality parameter, has been derived algebraically for $k=1$ and 2 and studied numerically to show that tighter linkage results in higher power. Two pairs of markers give higher power over single marker case. The theory has been validated with the help of simulated data on 4 markers with 500 trios. Possible future investigation on TDT with two adjacent linked QTLs following risk of duplicate dominant epistasis type and involving a single marker has been indicated.

Key Words: Transmission/Disequilibrium Test; Disease Genes; Interval Mapping; Non-centrality parameters; Risk parameters; Linkage disequilibrium; Single nucleotide polymorphism

Introduction

Genetic disorders in humans are broadly grouped into two categories. In the first category, falls those diseases whose genetic control follows simple Mendelian principles such as cystic fibrosis and Huntington disease. In these diseases, a single-nucleotide polymorphism in the CFTR gene profoundly affects the bearer's digestive, reproductive, and respiratory systems and causes excessive loss of salt through sweating. These symptoms collectively are known as cystic fibrosis. In the second category, we have complex diseases such as diabetes, schizophrenia, cancer etc. where the mode of inheritance does not follow a Mendelian pattern. Many genes, each with small and supplementary effects are involved in such cases that cannot be identified individually and followed through generations as we do in the case of Mendelian genes. Recent advances in DNA chip technology, microarrays etc. have opened up new possibilities for the identification of such disease genes by the correlation between the disease genes and the specific DNA markers (Narain, 2000). This involves methods of genetic analysis such as quantitative trait locus (QTL) (Narain, 2003, 2005) and linkage disequilibrium (LD) mapping. The method of LD mapping to ascertain linkage between the disease genes and molecular markers does not require assumptions about the mode of inheritance of the disease genes. Such methods use either case-control studies or family-based controls. In the latter category, a powerful test known as *transmission/disequilibrium test* (TDT) was developed by Spielman *et al.* (1993) for studying insulin-dependent diabetes mellitus (IDDM) considering only a given marker. But when

*Author for Correspondence: E-mail: narainprem@hotmail.com several adjacent marker loci are used for screening, one can examine each locus individually and make some correction for 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 have appeared in the literature that consider multiple markers simultaneously that yield more genetic information than the study of single markers. But their approach has encountered the problem of the occurrence of families with ambiguous haplotypes. In the INSA project we adopted a different approach in which we 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*. By choosing different gene locations throughout the length of the chromosome, the behavior of the concerned statistics can pinpoint the optimum location of the disease gene. This new approach has been developed for the first time in this project although Nielsen *et al*. (2004) did study the effect of two- and three-locus linkage disequilibrium on the power of case-control tests. The interesting results obtained in the INSA project prompted extension to more than a pair of markers. Further, for validation of the theory developed, simulated data from Lin *et al.* (2004) were used.

These results including theory of such tests along with their powers, wherever possible, have been reviewed and discussed primarily from a methodological point of view.

Case-Control analysis vs. TDT

The association between the markers and the disease phenotypes depends on the linkage disequilibrium (LD) between the disease susceptibility gene and the marker that got generated when the disease gene arose by mutation from the normal type in the past. The conservation of LD over the generations, either because of linkage with the gene and the marker on the same chromosome or because of association when they lie on different chromosomes, produces the correlation between the marker and the phenotype.

The most common method for discovering such an association is Case-Control analysis in which marker allele frequencies in unrelated cases and controls are compared. Any significant difference indicates the association. The test statistic is a ‘relative risk’ (RR) statistic or contingency statistic for association between disease and marker allele status in terms of a chi-square with 1 d.f.. However, if the population is composed of a recent admixture of different ethnic groups that differ in marker and disease gene frequencies, the association inferred would be spurious. The population admixture/ stratification tends to mask or even reverse true genetic effects as shown by Ewens and Spielman (1995). Based on multi-generation models of population history and subdivision they showed that the expected value of the non-squared numerator of the RR statistic for testing population association contains, besides the term containing disequilibrium and linkage parameters, an extra term, independent of these parameters, that represents a ‘spurious’ association between marker and disease gene frequencies due solely to the admixture process. They concluded that because of this term the RR statistic does not provide an appropriate test of the null hypothesis of no association in subdivided populations.

TDT method, on the other hand, evaluates whether the frequency of transmission of alleles from heterozygous parents to their affected children deviates from 50%, the expected Mendelian frequency when there is no linkage. The parents of the disease cases serve here as a within family control. The expectation of the non-squared numerator contains no spurious term due to the admixture process and is zero when disease and marker locus are unlinked. TDT is, therefore, a valid test for linkage and/or association between marker loci and the disease susceptibility gene under population subdivision and admixture. Not only that TDT statistic is made even more powerful by the admixture process due to the presence of disequilibrium parameters of different subdivided populations. A recent review by Laired and Lange (2006), giving a unified account of ‘family-based association tests’ (FBATs), indicates how TDT can be extended in several directions including parametric likelihoodbased approach. The review describes the TDT, discussed in this paper, as the simplest family-based design using data from trios-the affected offspring and his or her two parents-and stresses that the only testable null hypothesis is that the marker is both linked and associated with a DSL affecting the trait. It also emphasizes that family-based designs are robust against population admixture and stratification, allow both linkage and association to be tested and offer a solution to the problem of model building.

Risk Parameters

Let the disease susceptibility locus (DSL) be denoted by *M*-*m* where *M*, with frequency p_M, is associated with the disease status i.e. increasing disease risk with more copies of this gene and *m*, with frequency $q_M = (1-p_M)$ is the other allele. We consider an infinitely large random mating population under Hardy – Weinberg equilibrium with no segregation distortion. Let the penetrance, defined as Prob.(disease/genotype at the DSL), of the three possible genotypes, *MM*, *Mm* and *mm* at the DSL be denoted by ϕ_{MM} , ϕ_{Mm} and ϕ_{mm} respectively with $\phi_{mm} \leq \phi_{Mm} \leq \phi_{MM}$. The population disease prevalence is then

$$\phi = p_M^2\phi_{MM} + 2p_M q_M \phi_{Mm} + q_M^2\phi_{mm}. \qquad (1)$$

We also define

$$\begin{aligned} C &= p_M \phi_{MM} + (q_M - p_M)\, \phi_{Mm} - q_M\phi_{mm} \\ &= (p_M \phi_{MM} + q_M \phi_{Mm} - \phi) - (p_M \phi_{Mm} + q_M \phi_{mm} - \phi) \qquad (2) \end{aligned}$$

which is the average effect of gene substitution with respect to the penetrance as a trait in random mating populations (Edwards, 2000).

The genotypic relative risk (GRR) of a genotype is defined as the ratio of its prevalence to that of a reference genotype, with no copy of the risk allele i.e. *mm* here. Denoting them by g_2, g_1 and g_0 respectively for the genotypes *MM, Mm* and *mm*, we have $g_2 = \phi_{MM} / \phi_{mm}$, $g_1 = \phi_{Mm} / \phi_{mm}$ and $g_0 = 1$. However, we need only one of g_2 or g_1 say, $g_2 \geq 1$, to express the composite parameter (ϕ /C) in terms of its value and p_M, the disease gene frequency, depending on the mode of inheritance (MOI) i.e. additive, recessive, dominant and multiplicative with respect to disease allele *M* as given below:

Additive: $\phi_{Mm} = (\phi_{MM} + \phi_{mm})/2$, $(\phi /C) = 2\,[p_M + (g_2 - 1)^{-1}]$.

Recessive: $\phi_{mm} = \phi_{Mm}$, $(\phi /C) = [p_M^2 + (g_2 - 1)^{-1}]/p_M$.

Dominant: $\phi_{MM} = \phi_{Mm}$, $(\phi /C) = [g_2\,(g_2 - 1)^{-1} - q_M^2]/q_M$.

Multiplicative: $\phi_{Mm}^2 = \phi_{MM}\, \phi_{mm}$, (ϕ /C)

$$= [p_M + (g_2^{1/2} - 1)^{-1}].$$

Single Marker Case

Let *A*-*a* denote a marker locus with allelic frequencies *p* for *A* and *q*=*1*-*p* for *a* that is to be evaluated in relation to a disease trait locus *M*-*m* with a recombination probability between them as *r*. The TDT compares the frequencies of marker alleles, *A* and *a*, transmitted from the parents to affected offspring with those of the alleles that are not transmitted and so is based on a 2x2 table containing frequencies for the marker alleles transmitted (*T*) or not transmitted (*NT*) from parents to affected offspring in a random sample of 2*N* parents ascertained through their *N* affected offspring from a population in Hardy-Weinberg equilibrium as given in Table 1.

Table 1: Observed counts for transmitted and non-transmitted marker alleles *A* and *a* among 2N 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*)	2*N*

The expected values of the observed counts in Table 1 are conditional probabilities with which a parent transmits one marker allele and not the other, given that the offspring is affected. In order to determine them we need to consider the population genetics model of a two loci system as discussed by Narain (2007a).

(a) Test Statistic

From the expectations given in Narain (2007a), we get

$$E(c\text{-}b) = 2N\,[(1–2r)\,D\,(C\,/\,\phi)] \qquad (3)$$

$$E(c+b) = 2N\,[2pq + (q\text{-}p)\,D\,(C\,/\,\phi)] \qquad (4)$$

where $D=(p_{AM} - pp_M)$, p_{AM} being the frequency of two-locus haplotype *AM*, is the coefficient of linkage disequilibrium between the disease gene and the marker locus.

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

$$\gamma^2 = (c\text{-}b)^2/(c+b) \qquad (5)$$

that follows a chi-square distribution with one degree of freedom and therefore can be used to test whether there is an association between marker *A* and the trait gene *M*. It may be noted that (*c*+*b*) provides with an estimate of the variance of (*c*-*b*).

(b) The Insulin Gene Region and insulin-dependent diabetes mellitus (IDDM)

In the study on insulin dependent diabetes mellitus (IDDM) reported by Spielman *et al*. (1993), there were 94 families of which 57 parents were heterozygous for a marker, with “1” and “X” alleles, on chromosome 11p. The total of 62 children affected by IDDM indicated 124 alleles transmitted to the children. Among them there were c=78, “1” alleles and b = 46 “X” alleles. This gave, by (5), a chi-square value of 8.26 with 1 d.f. that is highly significant with p=0.004, thus demonstrating that the marker is linked to the susceptible gene for IDDM.

(c) Power of the test

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 $\gamma^2(1, \lambda)$ distribution with 1 d.f. and non-centrality parameter λ given by

$$\lambda = [E(c) – E(b)]^2\,/\,[E(c) + E(b)]$$

$$= 2N\,[(1 – 2r\,)^2\,D^2]/S \qquad (6)$$

where $S= (\phi\,/C)[2p\,q(\phi\,/C) + D\,(q – p)]$.

The power of the test is then the probability that the deviate from $\gamma^2(1, \lambda)$ is greater than or equal

Table 2. Power in the single marker case for different recombination probability (*r*) between the marker and the disease locus for each of the two combinations of the disease gene frequencies (p_M) under different MOIs when N = 200, g_2 = 2, D = 0.1, p = 0.2 and α = 0.05

	Additive		Recessive P_M		Dominant		Multiplicative	
r_1	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
0.45	0.052	0.051	0.050	0.052	0.054	0.051	0.052	0.052
0.30	0.088	0.074	0.058	0.085	0.124	0.068	0.081	0.076
0.20	0.136	0.105	0.069	0.129	0.220	0.091	0.121	0.109
0.10	0.206	0.150	0.083	0.194	0.352	0.123	0.178	0.157
0.04	0.257	0.183	0.094	0.241	0.442	0.148	0.220	0.192
0.01	0.285	0.202	0.100	0.267	0.489	0.161	0.244	0.212

to $\chi^2(\alpha)$, the critical value of χ^2 to reject the null hypothesis at significance level α.

Although not required for the validity of the TDT, for power computation we need a model for the mode of inheritance (MOI) in terms of the genotypic relative risk (GRR) as discussed in Section 3. To study the effect of linkage on the power of the test, we determine the values of power for different values of the non-centrality parameter for $N = 200$, D = 0.1, p = 0.2 and (C/ϕ) under different MOIs for different values of r for each of the two combinations of (0.2 and 0.5) of the disease gene frequencies. The results are given in Table 2. They indicate that power increases with decrease in the values of r showing that tighter linkage increases the power of the test. It is also higher with smaller values of disease gene frequencies except in the case of recessive MOI where it is lower with smaller values of p_M.

Case of a pair of markers

When two adjacent marker loci are used for screening, one can examine each locus individually using TDT. However, this approach ignores the possible dependence among the two marker loci. We may, therefore, tend to lose information on linkage by conducting single marker analysis. Narain (2007b) developed, probably for the first time, a theory of TDT with two linked flanking marker loci in the context of interval mapping of a disease gene under the assumption that the phase in the two-loci haplotypes is known in the parents. But the theory was restricted to the case when, at the disease locus, *only* the recessive homozygous genotypes are affected by the disease and the other homozygous and heterozygous genotypes are unaffected and therefore treated as normal individuals. In actual practice, however, every individual in the population has a certain risk of being affected by the disease. The three genotypes have therefore, associated with them, certain risk parameters. If we incorporate this factor in the theory of TDT with two linked marker loci with interval mapping we get a complete account of the theory as done by Narain (2009).

When we consider a disease gene flanked by two marker loci, such as SNP or micro-satellite, *A-a* and *B-b* linked with a recombination probability between them as r and the DSL is such that the r_1 is the recombination probability between A and M and r_2 is the recombination probability between B and M. 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. It may be noted that since we are considering *interval mapping* approach of the disease gene, the order of the three loci is necessarily *ADB*. In interval mapping approach, an entire chromosome say of 100 cM length is considered with a series of markers $A_1 - A_2$ - $A_3 - A_4 - A_5$ - A_6 spaced say every 20 cM and the TDT is used for each successive interval separately. The *lowest* of the significant P values over the intervals then indicates the interval harboring the disease gene. We take one affected child and the two parents to form one nuclear family-a triad- and suppose we have a sample of N triads. We genotype the triads for the two markers and assume that the phase of the linkage in the haplotypes have been either determined by pedigree analysis/molecular haplotyping or inferred by statistical methods/ computer algorithms. We consider the frequencies of marker haplotypes *AB*, *Ab*, *aB* and *ab* transmitted from a given parent of a specified genotype to affected offspring with those of the haplotypes not transmitted. It results in a *4* x *4* table containing frequencies of the marker haplotypes transmitted (T) or not transmitted (NT) from parents to affected offspring in a sample of $2N$ parents of N affected offspring as given in Table 3.

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

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

(a) Expected values of the counts

Let the gene frequencies of the two markers *A-a* and *B-b* be denoted by p_1 , q_1 = 1-p_1 and p_2 , q_2 = 1-p_2 respectively. Let the pair-wise (first order) disequilibrium parameters between *A-a* and *B-b* be denoted by D_{12} , between *A-a* and *D-d*, by D_{1d} and between *B-b* and *D-d* by D_{2d} . These can be expressed in terms of two-locus haplotypes frequencies and gene frequencies (Narain, 1990; Weir, 1996). In addition, we have also a three-locus (second order) linkage disequilibrium parameter between (*A, B)* and *D* denoted by D_{12d}. The expectations of the observed counts in Table 3 can be obtained by taking a given parental genotype and determining first the compound probability that its one of the two marker haplotypes is transmitted and not the other as well as the offspring is affected by the disease. This probability is then divided by the ϕ, the probability of an offspring being affected by the disease, to obtain the conditional probability that, given that the offspring is affected by the disease trait, the parent transmits one marker haplotype and not the other. These are derived in Narain (2009). From these we get

$$E(n_{14} – n_{41}) = 2N\,[(1–2r_1)C_1 + (1–2r_2)C_2](C/\phi) \qquad (7)$$

$$E(n_{23} – n_{32}) = 2N\,[(1–2r_1)C_1 – (1–2r_2)C_2](C/\phi) \qquad (8)$$

where

$$C_1 = p_2q_2\,D_{1d} + D_{12}D_{2d} + (1/2)(q_2 – p_2)\,D_{12d}$$

$$C_2 = p_1q_1\,D_{2d} + D_{12}D_{1d} + (1/2)(q_1 – p_1)\,D_{12d}$$

It is important to note that when $D_{1d} = D_{2d} = D_{12d} = 0$, the expectations are independent of r_1 and r_2 and therefore there can be *no power* of various tests of the hypotheses involving r_1 = 1/2 and/or r_2 = 1/2. That is, the TDT tests with two linked marker loci have no power unless there are associations of different orders between the genes at the DSL and those at the marker loci. To have some power at least one of the association parameters need be non-zero. At the same time these considerations also show that more the magnitudes of the associations, the greater would be the power of TDT. It may further be seen that by pooling appropriate cell counts in Table 3 and taking their expectations, we can get the expectations of the cell counts for a 2 x 2 table for a single marker case.

(b) Various Tests for Linkage

In Table 3, the four entries in the diagonal pertaining to doubly homozygous parents are uninformative about linkage and therefore do not contribute to the test. Of the twelve remaining entries, six above the diagonal are matched with six below the diagonal. Of the six pairs so formed, four pertain to the singly heterozygous parents at each of the two markers (there being two possible homozygotes at the other marker locus) and two 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 D's 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 expectations of the marginal totals $E(n_{i.})$ and $E(n_{.j})$ then become the same. However, the converse is not true. Marginal homogeneity can occur *without* symmetry. Zhao *et al.* (2000) uses a test for marginal homogeneity for *h* x *h* transmission/non-transmission table with *h* possible haplotypes that does not imply symmetry below and above diagonal unless *h* = 2 as with a single marker. We need therefore here a test for symmetry. Following Bowker (1948), the test of symmetry in 4 x 4 contingency table can be performed with the help of the statistic

$$\gamma^2 = \Sigma\,\Sigma\,(n_{ij} – n_{ji})^2/(n_{ij} + n_{ji})$$

$$\text{for } 2 \le i \le 4,\ 1 \le j \le (i–1). \qquad (9)$$

This statistic follows a chi-square distribution with 6 d.f., being 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 as well as the 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}, therefore, the chi-square with one d.f. would be

$$\gamma^2 = (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji})$$
$$\text{for } 2 \leq i \leq 4,\ 1 \leq j \leq (i\text{–}1), \qquad (10)$$

which tests for the concerned linkage and is known as *McNemar* test. On the null hypothesis of no linkage between the disease gene and the markers, each of these expectations of the *difference* would be zero. These component chi-squares, each with 1 d.f. are given below:

$$\gamma^2 \text{ (AB/Ab)} = (n_{21} - n_{12})^2/(n_{21} + n_{12}) \qquad (11)$$

$$\gamma^2 \text{ (AB/aB)} = (n_{31} - n_{13})^2/(n_{31} + n_{13}) \qquad (12)$$

$$\gamma^2\text{(AB/ab)} = (n_{41} - n_{14})^2/(n_{41} + n_{14}) \qquad (13)$$

$$\gamma^2 \text{ (Ab/aB)} = (n_{32} - n_{23})^2/(n_{32} + n_{23}) \qquad (14)$$

$$\gamma^2 \text{ (Ab/ab)} = (n_{42} - n_{24})^2/(n_{42} + n_{24}) \qquad (15)$$

$$\gamma^2 \text{ (aB/ab)} = (n_{43} - n_{34})^2/(n_{43} + n_{34}) \qquad (16)$$

For testing linkage between the disease gene and *both* the markers, we have to consider only the case where the parental genotype is doubly heterozygous, either in the coupling phase or in the repulsion phase of linkage. The corresponding chi-square tests, each with 1 d.f. would be given by relations [13] and [14]. 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 disequilibrium exists.

(c) Power of the Tests

Under the alternative hypothesis that the DSL is linked with *both* the markers and *at least* one of the three disequilibrium coefficients is non-zero, each of the chi-squares follows approximately a non-central chi-square distribution with 1 d.f. and with approximate non-centrality parameters λ_1 for AB/ab and λ_2 for Ab/aB obtained by replacing the observed counts *n* by their expected values in the formulae of chi-squares (Meng and Chapman, 1966; Deng and Chen, 2001) and given by

$$\lambda_1 = [E(n_{14}) - E(n_{41})]^2/[E(n_{14}) + E(n_{41})]$$
$$= 4N^2 [(1 - 2r_1)C_1 + (1 - 2r_2)C_2]^2 (C/\phi)^2/S_1, \qquad (17)$$

$$\lambda_2 = [E(n_{23}) - E(n_{32})]^2/[E(n_{23}) + E(n_{32})]$$
$$= 4N^2 [(1 - 2r_1)C_1 - (1 - 2r_2)C_2]^2 (C/\phi)^2/S_2, \qquad (18)$$

where S_1 and S_2 are functions of *p's, r, D's*, and C/ϕ.

The power (probability of rejecting a false hypothesis) of the chi-square tests can be studied directly by considering the corresponding non-centrality parameters λ's.

It may be seen that λ's will be zero when $D_{1d} = D_{2d} = D_{12d} = 0$. This means, as already noted earlier, that TDT has no power unless at least one of the D's is non-zero, i.e. there is association between the genes and *at least* one of the marker loci and those at the DSL.

To study the effect of linkage on the power of the test, we evaluate power corresponding to different values of λ_1 for $N = 200$, $D_{1d} = D_{2d} = D_{12d} = 0.1$, $D_{12} = 0.1$, $r_2 = 0.05$, $p_1 = p_2 = 0.2$ and (C/ϕ) under different MOIs, for different values of r_1 for each of the two combinations (0.2 and 0.5) of the disease gene frequencies. The results are given in Table 4.

It is apparent from Table 4 that the effect of linkage is to increase the power in all cases considered. In other words, the tighter the linkage (smaller values of r_1) the higher is the power. Further, when we compare the power for different values of p_d, at a given value of r_1, it is found to be higher for $p_d = 0.2$ compared to that for $p_d = 0.5$ except in the case of recessive MOI. In the case of recessives, however, the power is lower with a smaller value of p_d. In all the cases, the power values are smaller in Table 2 compared to those in Table 4 indicating the superiority of TDT with a pair of linked marker loci in terms of increase in power of the test. It may be

Table 4: Power for different recombination probability (r_1) between the marker A-a and the disease locus for each of the two combinations of the disease gene frequencies (p_M) under different MOIs when $N = 200$, $g_2 = 2$, $D_{1d} = D_{2d} = D_{12d} = 0.1$, $D_{12} = 0.1$, $r_2 = 0.05$, $p_1 = p_2 = 0.2$ and $\alpha = 0.05$

	Additive		Recessive P_M		Dominant		Multiplicative	
r_1	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
0.45	0.243	0.176	0.093	0.229	0.408	0.144	0.210	0.185
0.30	0.314	0.224	0.110	0.295	0.522	0.179	0.269	0.235
0.20	0.360	0.255	0.121	0.338	0.590	0.203	0.309	0.269
0.10	0.405	0.287	0.132	0.380	0.650	0.227	0.348	0.303
0.04	0.431	0.306	0.139	0.405	0.685	0.242	0.371	0.323
0.01	0.444	0.316	0.142	0.418	0.700	0.249	0.382	0.333

noted that the power varies from 0.050 to 0.489 in Table 2 and from 0.093 to 0.700 in Table 4. These are somewhat low, never touching 80% mark, due to sample size (number of triads) being 200 only. Sample sizes needed to achieve 80% power with $\alpha = 0.05$ were obtained in Narain (2009). For $g_2 = 2$, the values were 475, 2038, 254, and 570 respectively for additive, recessive, dominant and multiplicative MOIs. The corresponding values for single marker case were, as expected, higher being 813, 3668, 420 and 985 respectively. At the significance level of 5×10^{-8} used in TDT for fine mapping and genome-wide association studies, the sample sizes required would be even several times larger than these values.

More than Two Pairs of Markers

The above theory of TDT with haplotypes consisting of only a pair of linked markers is capable of generalization to haplotypes with an arbitrary number of linked markers. For instance, let us consider the case of haplotypes with three markers (SNPs) on such a haplotype, each with two alleles, denoted by *A-a, B-b,* and *C-c* so that there are eight types viz. *ABC, ABc, AbC, Abc, aBC, aBc, abC,* and *abc*. Consider a sample of N triads, each containing two parents and their one affected child. The triads are genotyped for the three markers. We assume that their phases have been determined by pedigree analysis/molecular haplotyping or inferred by statistical methods/ computer algorithms. We consider the frequencies of the eight types of markers transmitted from a given parent of a specified genotype to the affected offspring with those of the types *not* transmitted. The resulting data in a 8 x 8 table containing the marker types transmitted (T) or not transmitted (NT) from parents to affected offspring in a sample of 2N parents of N affected offspring could be used to develop the four chi-square tests each based on 1 d.f. on the null hypothesis of no linkage between the disease-susceptibility gene and all the three markers together. However, in order to determine the power of these tests we need to work out the expected values of the frequencies in the tables generated. This depends on the population genetics model of a four loci system. Mathematically, this is quite involved but going by the results obtained for one marker (Narain, 2007b) and two markers (Narain, 2009) as discussed above, it is apparent that the power of tests for three markers, in terms of the non-centrality parameters, is expected to be higher than for the two markers case.

The above considerations indicated that we can generalize the tests to a set of k markers. In this situation with biallelic SNP markers, we have $N = 2^k$ types of possible haplotypes. Therefore we need consider a N x N table of transmission events in which N diagonal elements will correspond to homozygous transmissions, not relevant to the problem of testing linkage, and $N(N\text{-}1)$ off-diagonal elements resulting in $N(N\text{-}1)/2$ matched pairs of transmissions. This gives *N/2* chi-square tests, corresponding to different phases of the *k*-tuple heterozygote, each based on 1 d.f. on the null hypothesis of no linkage between the disease-susceptibility gene and all the *k* markers jointly.

Validation of the Theory

To validate the theory discussed above simulated data were obtained from Dr. Aravinda Chakravarti of Johns Hopkins University School of Medicine, Baltimore, USA and Dr. David J. Cutler of Emory University School of Medicine, Athens, USA. These pertained to 50 markers for 50 trios after phase reconstruction. Pedigrees of triads were developed for a data set on 500 trios with four SNPs say *A*, *B*, *C*, and D each with two alleles. Transmissions numbers were determined for them, there being a total of 1,000, each trio being counted twice for the parent transmitting and not-transmitting the concerned haplotype along with the disease gene. The resulting samples gave rise to chi-square tests, each with 1 d.f., for various situations as described below.

For testing linkage with single markers, appropriate cell numbers in the respective 4 x 4 contingency table were pooled and relation [5] was used to get the corresponding chi-squares. The results are presented in Table 5. These results indicate that none of the chi-squares is significant. None of the four markers is therefore associated with the disease.

It may be noted that chi-squares for single marker case can also be worked out for the situation when singly heterozygous parents are considered with homozygotes at the other marker locus, there being two possible cases. For such chi-squares, relations similar to [11, 12, 15], and [16] pertaining to the pair of markers A-a & B-b could be used. However results for such cases are not presented here.

Table 5: Chi-square tests with 1 d.f. for four Single Marker Case

Marker	Sample size	γ^2 values
A-a	211	1.36967
B-b	497	2.19115
C-c	470	2.45957
D-d	284	3.16901

The chi-square results for the six pairs of markers are given in Table 6.

Table 6: Chi-square tests with 1 d.f. for six Pairs of Markers

Pair	Coupling		Repulsion	
	Sample size	γ^2 values	Sample size	γ^2 values
A-a & B-b	-	-	128	0.125
A-a & C-c	94	0.38298	23	2.13044
A-a & D-d	-	-	41	8.80488**
B-b & C-c	40	0.1	207	3.01932
B-b & D-d	67	3.35821	72	0.22222
C-c & D-d	87	0.93103	60	1.06667

**Significant at 1% level (≥ 6.635)

Only one pair A-a & D-d in the repulsion phase happens to be significant at 1% level. These results show that individually markers A-a as well as D-d are not associated with the disease but taken together they are. This is what the theory predicted and emphasized the necessity of a pair of makers bracketing the gene (a sort of interval mapping) to be considered for gene mapping using TDT.

Discussion

The theory of TDT with an arbitrary number of marker loci has been developed. In particular the case of two linked marker loci with first and second order disequilibria has been studied in detail. A new test statistic for testing linkage with 1 d.f., based on the test for symmetry, has been discussed. It uses data only on the doubly heterozygous parents who transmitted the given haplotype to their affected offspring. The power of the test has also been discussed in terms of the 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 the parameters pertaining to gene frequency and linkage. But the underlying tests for linkage still hold. The theory has been validated with the help of simulated data on four markers with 500 families of father-mother- one affected child (trios).

The major assumption in this study is that the two-loci haplotypes are known in the parents. The traditional method to determine haplotypes is either pedigree analysis or molecular haplotyping. Both of these methods require a lot of work in 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 to reconstruct the haplotypes. 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. However, a better way would be to generalize the test proposed in this paper to the 'sib TDT' or S-TDT type procedure where data consist of marker genotypes of the offspring only, both affected and unaffected, for each family.

There is one issue with TDT even in a single marker case that does not seem to have been addressed in the literature. The theory well established for this case and most widely used is based on a model of marginal QTL of a single locus whereas for most complex traits several QTLs are expected to be involved. We may consider two adjacent linked QTLs with the risk of duplicate dominant epistasis type, a sort of duplicate gene action that when extended to infinitely large number of genes results in quantitative genetic variation. With transmission information on a given marker locus the model of null hypothesis H_0 : (1-2r)D = 0 [from relation (3)] discussed above would then need to be replaced by H_0 : (1-r_1-r_2)D_{1d1d2} = 0 where r_1 and r_2 are linkage parameters between the marker and the two respective QTLs and D_{1d1d2} is the non-zero disequilibrium coefficient between the marker and the haplotype consisting of the two QTLs, assuming that the marker and each of the two QTLs separately are in linkage equilibrium so that the corresponding disequilibrium coefficients are zero. Under this new model the null hypothesis tests r_1 = 1/2 and r_2 = 1/2. Preliminary investigation on the power of such a test shows that it gets somewhat reduced compared to that under the standard model. Further properties of such a model remain to be determined.

An interesting issue is how does power of TDT compares with that of case-control studies when we incorporate two flanking markers into the model. As already pointed out in Section 2, the case-control studies with one marker locus *A*-*a*, involve comparing gene or genotypic frequencies in control (individuals unaffected by the disease) with those in the case (individuals affected by the disease). When the difference is significant the marker is said to be associated with the disease. If $p_{unaff.}$ and $p_{aff.}$ denote respectively the frequencies of *A* in the two situations, the expected value of the difference between them is $x = DC/\phi(1\text{-}\phi)$ that leads to the usual γ^2 with 1 d.f. as well as the power function in terms of the non-centrality parameter which depends on x^2. When the allelic gene substitution effect *C* for the character *penetrance* is negligible, *x* would be zero in spite of a large *D* indicating that the disease association with the marker is only possible to be detected on the condition that the penetrance is heritable. A power comparison with TDT has been quoted in Laired and Lange (2006) based on the studies by McGinnis (2000) and McGinnis *et al.* (2002) involving 200 cases, 200 controls and 200 trios. For rare disease (ϕ = 0.1%), the TDT is more powerful than case-control design but for common disease (ϕ = 14%), case-control designs are slightly more powerful than TDT.

When we consider two markers *A*-*a* and *B*-*b* we have to compare four haplotype frequencies for *AB*, *Ab*, *aB*, and *ab* between the case and control. It leads to four x's

$$x_1 = \delta_1 C/\phi(1\text{-}\phi);\ x_2 = \delta_2 C/\phi(1\text{-}\phi);\ x_3 = -\delta_3 C/\phi(1\text{-}\phi);\ x_4 = -\delta_4 C/\phi(1\text{-}\phi)$$

where

$$\delta_1 = p_1 p_2(D_{1d}/p_1 + D_{2d}/p_2 + D_{12d}/p_1 p_2);$$

$$\delta_2 = p_1 q_2 (D_{1d}/p_1 - D_{2d}/q_2 - D_{12d}/p_1 q_2)$$

These lead to a chi-square with 3 d.f. that can be partitioned orthogonally into three chi-squares each with 1 d.f. with corresponding contrasts whose expected values are

$$y_1 = x_1 + x_2 - x_3 - x_4 = D_{1d}\ C/\phi(1\text{-}\phi)$$

$$y_2 = x_1 - x_2 + x_3 - x_4 = D_{2d}\ C/\phi(1\text{-}\phi)$$

$$y_3 = x_1 - x_2 - x_3 + x_4 = (p_1 - q_1)(p_2 - q_2)$$
$$[D_{1d}/(p_1 - q_1) + D_{2d}/(p_2 - q_2) + D_{12d}/$$
$$(p_1 - q_1)(p_2 - q_2)]\ C/\phi(1\text{-}\phi)$$

The chi-square corresponding to y_1 thus tests for disease association with the marker locus *A-a*, involving only a first order disequilibrium, that for y_2 with the marker locus *B-b*, involving only a first order disequilibrium and that for y_3 with both the markers, involving a combination of the two first and one second order disequilibria. It is interesting to note that when the two marker gene frequencies p_1 and p_2 are each equal to half, as in F_2 of a cross between two pure strains, the third component y_3 reduces to $D_{12d}\ C/\phi(1\text{-}\phi)$, indicating that this component tests for the association of the disease jointly with the second order disequilibrium D_{12d} only. The power of the test for case-control design in a two flanking markers setting can thus be obtained from the non centrality parameter of the chi-square derived from y_3 and can be compared with that for the TDT discussed in this paper.

Acknowledgements

The author is grateful to Dr. Aravinda Chakravarti of Johns Hopkins University School of Medicine, Baltimore, USA and Dr. David J. Cutler of Emory University School of Medicine, Athens, USA. for making available the simulated data on trios for a large number of markers used in this study. The author is also grateful to two referees for valuable suggestions that led to improvement in the presentation. This work was supported by the Indian National Science Academy, New Delhi, India under their program ‘INSA Honorary Scientist’.

References

Bowker A H (1948) A test for symmetry in contingency tables *J Am Statist Assoc* **43** 572-574

Deng H W and Chen W M (2001) The power of the transmission disequilibrium test (TDT) with both case-parent and control-parent trios *Genet Res Camb* **78** 289-302

Edwards A W F (2000) Foundations of Mathematical Genetics. Second Edition, Cambridge University Press, Cambridge, UK, New York, USA

Ewens W J and Spielman R S (1995) The transmission/disequilibrium test: history, subdivision, and admixture *Am J Hum Genet* **57** 455-464

Laired N M and Lang C (2006) Family-based designs in the age of large-scale gene-association studies *Nature Reviews Genetics* **7** 385-394

Lin S, Chakravarti A and Cutler D (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies *Nature Genetics* **36** 1181-1188

McGinnis R (2000) General equations for P_t, P_s, and the power of the TDT and the affected sib-pair test *Am J Hum Genet* **67** 1340-1347

McGinnis R, Shifman S and Darvasi A (2002) Power and the efficiency of the TDT and case-control design for association scans *Behav Genet* **32** 135-144

Meng R C and Chapman D G (1966) The power of chi-square tests for contingency tables *J Am Stat Assoc* **61** 965-975

Narain P (1990). Statistical Genetics. (New York: John Wiley) (Wiley Eastern Ltd., New Delhi, reprinted in 1993). Published by the New Age International Pvt. Ltd., New Delhi in 1999 & reprinted in 2008)

Narain P (2000) Genetic diversity: conservation and assessment. *Current Science* **79(2)** 170-175

Narain P (2003) Evolutionary genetics and statistical genomics of quantitative characters *Proc Indian Natn Sci Acad Biological Sci* **B69(3)** 273-352

Narain P (2005) Mapping of Quantitative Trait Loci (QTL) *The Math Student* **74(1-4)** 7-18

Narain P (2007a) Transmission disequilibrium test (TDT) used in human genetics *The Math Student* **76(1-4)** 47-58

Narain P (2007b) A theoretical treatment of interval mapping of a disease gene using transmission disequilibrium tests *J Biosci* **32(7**) 1317-1324

Narain P (2009) Transmission Disequilibrium Test (TDT) for a pair of linked marker loci *Comput Stat Data Anal* **53(5)** 1883-1893

Nielsen D M, Margaret G E, Zaykin D V and Weir B S (2004) Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations *Genetics* **168** 1029-1040

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

Weir B S (1996) Genetic Data Analysis II. Methods for Discrete Population Genetic Data. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, USA

Zhao H, Zhang, Merikangas K R, Trixler M, Wildenauer D B, Sun F and Kidd KK (2000) Transmission/Disequilibrium tests using multiple tightly linked markers *Am J Hum Genet* **67** 936-946.