Methodology article

Haplotype-based score test for linkage in nuclear families

Chung Mo Nam[1], Dae Ryong Kang[1] and Jinheum Kim*[2]

Address: [1]Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, Seoul, South Korea and [2]Department of Applied Statistics, University of Suwon, Suwon, South Korea

Email: Chung Mo Nam - cmnam@yumc.yonsei.ac.kr; Dae Ryong Kang - drkang@yumc.yonsei.ac.kr; Jinheum Kim* - jinhkim@suwon.ac.kr

* Corresponding author

Abstract

Background: To look for genetic linkage between angiotensin-I converting enzyme(ACE) gene and hypertension in a Korean adolescent cohort, we developed a powerful test using the covariances between marginal differences and their variances in a transmission/non-transmission table.

Results: We estimated haplotype frequencies using the parental and affected offspring's genotypes and then constructed a transmission/non-transmission table for the parental haplotypes transmitted to the offspring. We then proposed a test for checking the marginal homogeneity in the table. Because the cells in the table were dependent due to the uncertainty of the parental haplotypes, we adopted a randomization procedure to estimate the significance of the observed test statistic. Simulations show that our test performs well on a nominal level and has a monotone power, which increases as the relative risk increases. With our test, there was no evidence of genetic linkage between the ACE gene and hypertension in the Korean adolescent cohort.

Conclusion: We developed a score test for linkage and used simulations to demonstrate that our test performs well at a nominal level. Under some situations where the diversity of haplotypes is low, the proposed test gained a little power over the method based on only variances between marginal differences in a transmission/non-transmission table.

Background

For linkage and/or association studies based on haplotypes, molecular haplotyping can be done for each individual, but at a high cost. Instead, statistical methods such as Clark's algorithm [1], the EM algorithm [2], or Gibb's sampler method [3], are commonly used to reconstruct haplotypes. The likelihood ratio test, which is based on the difference between the sum of the log-likelihoods of the case and control groups and the log-likelihood of the combined data, is usually used for case-control association studies [4], while the TDT is used for linkage and/or association studies in nuclear families. The latter compares the frequency of transmitted parental haplotypes with that of non-transmitted parental haplotypes to an affected offspring. One of the difficulties with the TDT is that there is uncertainty in estimating haplotype frequencies from parental genotypes. Wilson [5] and Clayton and Jones [6] proposed that these uncertain families be discarded from the analysis, but this drop in families leads to a low power. Clayton [7] proposed an estimating procedure based on likelihood, but it is no longer robust for population admixture or population stratification. Zhao et al [8] extended Spielman and Ewens' method [9] to test for marginal homogeneity in the transmission and non-

transmission of haplotypes. Their method is advantageous over those of Wilson [5] and Clayton and Jones [6] because there is no discarding of unresolved families, and it is still robust for cases of population admixture or population stratification. Cordell and Clayton [10] modeled the nuclear-family data via a conditional logistic regression in which they based on either haplotypes if parental phase may be inferrable or genotypes otherwise. Horvath et al [11] proposed family-based association tests with tightly linked markers where the phase may be ambiguous and the parental genotype data is missing. Their weighted conditional approach extended the method provided by Rabinowitz and Laird [12] to multiple markers.

Here, assuming that the haplotype block is very tight so that there is no recombination, we propose a score test for linkage and investigate its performance through simulations. We also illustrate the use of the proposed test with a data set taken from the Kangwha study [13].

Methods and Results
Score test
Let $\{H_s H_u, H_t H_v\}$ denote the event in which the transmitted haplotype in the father is H_s and the non-transmitted haplotype is H_t, and simultaneously the transmitted haplotype in the mother is H_u and the non-transmitted haplotype is H_v. We designate $\{H_s H_u, H_t H_v\}$ as one haplotype group and define $\{H_s{}^g H_u{}^g, H_t{}^g H_v{}^g\}$ as haplotype groups corresponding to the set of genotypes $g = 1, ..., G$, where G is the number of distinct sets of genotypes across all markers. Here each set of genotypes refers to the observed genotypes for the individual markers of the two parents and the affected offspring. Suppose the total number of possible haplotypes is k.

Zhao et al [8] proposed a TDT statistic to test linkage between markers and disease genes. They obtained a set of the estimated haplotype frequencies, $\{\tilde{h}_i\}$, using the EM algorithm based on the parental genotypes only, and for the haplotype group $\{H_s H_u, H_t H_v\}$ compatible with the set of genotypes g, defined $\tilde{t}_g^{su,tv}$ as the estimated number of families in which the father with haplotypes $\{H_s, H_t\}$ transmits H_s and the mother with haplotypes $\{H_u, H_v\}$ transmits H_u. Here the value of $\tilde{t}_g^{su,tv}$ is given by

$$\tilde{t}_g^{su,tv} = n_g \frac{\tilde{h}_s \tilde{h}_t \tilde{h}_u \tilde{h}_v}{\sum \tilde{h}_{s^g} \tilde{h}_{t^g} \tilde{h}_{u^g} \tilde{h}_{v^g}}, \qquad (1)$$

where the summation is over all haplotype groups compatible with g, and n_g is the total number of families compatible with g. Then they constructed a $k \times k$ transmission/

non-transmission table $\tilde{\mathbf{T}} = \{\tilde{t}_{ij}\}$, where $\tilde{t}_{ij}$ is the estimated number of parents who have haplotypes $\{H_i, H_j\}$ and who transmit H_i to the affected offspring across all sets of genotypes. That is,

$$\tilde{t}_{ij} = \sum_g \left\{ \sum_u \sum_v \tilde{t}_g^{iu,jv} + \sum_s \sum_t \tilde{t}_g^{si,tj} \right\}, \qquad (2)$$

where the summation over g ranges from 1 to G and the summations over s, t, u, and v range from 1 to k. They derived a test applying Spielman and Ewens' method to the table $\tilde{\mathbf{T}}$.

In this work we propose a score test that extends Sham's method [14] by modifying the transmission/non-transmission table $\tilde{\mathbf{T}}$. As mentioned by Rohde and Fuerst [15], the haplotype frequency estimation done using the EM algorithm over independent parents may derive possible haplotype pairs for parents that are contradictory to the ones for the offspring. However, including the offspring's genotype in the haplotype frequency estimation makes it possible to exclude these misleading haplotype pairs, thereby improving the accuracy of the underlying haplotype pair and the accuracy of a transmission/non-transmission table. This is why we adopted the EM method proposed by Rohde and Fuerst [15] and Becker and Knapp [16] rather than the method proposed by Long, Williams, and Urbanek [2]. We first obtained estimated haplotype frequencies, $\{\hat{h}_i\}$, based on family trios instead of on independent parental genotypes. Replacing $\tilde{h}_i$, $i = 1, ..., k$, by $\hat{h}_i$ in equations (1) and (2), we denote $\tilde{t}_g^{su,tv}$, $\tilde{t}_{ij}$, and $\tilde{\mathbf{T}}$ by $\hat{t}_g^{su,tv}$, and $\hat{t}_{ij}$, and $\hat{\mathbf{T}}$, respectively. We let $\hat{t}_{i\cdot}$, $i = 1, ..., k$, be the row marginal totals, defined by $\hat{t}_{i\cdot} = \sum_{j=1}^k \hat{t}_{ij}$ and $\hat{t}_{\cdot j}$, $j = 1, ..., k$, the column marginal totals, defined by $\hat{t}_{\cdot j} = \sum_{i=1}^k \hat{t}_{ij}$. We denote a vector of the marginal discrepancies for haplotypes 1 to k - 1 by $\hat{\mathbf{\Delta}} = (\hat{t}_{1\cdot} - \hat{t}_{\cdot 1}, ..., \hat{t}_{k-1\cdot} - \hat{t}_{\cdot k-1})'$. Note that when there is no linkage, $E(\hat{\mathbf{\Delta}}) = (0, ..., 0)'$ because the matrix $\hat{t}_{i,j}$ is symmetrical under the null hypothesis of no linkage irrespective of a set of haplotype frequencies [8]. Letting $\hat{\mathbf{\Sigma}} = \{\hat{\sigma}_{ij}\}$ with diagonal elements $\hat{\sigma}_{ii} = \hat{t}_{i\cdot} + \hat{t}_{\cdot i} - 2\hat{t}_{ii}$ and off-

diagonal elements $\hat{\sigma}_{ij} = -(\hat{t}_{ij} + \hat{t}_{ji}); (i \neq j)$, we propose a score test statistic T_s defined by

$$T_s = \hat{\boldsymbol{\Delta}}' \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Delta}}.$$

Whenever all the parental haplotype phases are uniquely determined, T_s asymptotically follows a chi-square distribution with $(k - 1)$ degrees of freedom because $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix of $\hat{\boldsymbol{\Delta}}$ under no linkage (see Appendix for the details). We note, however, that the cells in $\hat{\mathbf{T}}$ are not independent because the contribution of a father or a mother is not limited to a cell because of its uncertainty as a haplotype pair, but reaches out to some cells according to the number of possible haplotype pairs and corresponding probabilities. That is why T_s may not have a chi-square distribution with $(k - 1)$ degrees of freedom. Thus, we adopted the randomization test procedure introduced by Zhao et al [8] to estimate the statistical significance of the proposed test instead of relying on an asymptotic chi-square test.

Simulation studies
In our simulations, a hypothetical genealogy with three biallelic genetic markers was considered. We denoted alleles at the unobservable disease-susceptibility locus as D and d. There are eight haplotypes consisting of bialleles at three loci, l_1, l_2, and l_3: $H_1 = (1,1,1)$, $H_2 = (1,1,2)$, $H_3 = (1,2,1)$, $H_4 = (1,2,2)$, $H_5 = (2,1,1)$, $H_6 = (2,1,2)$, $H_7 = (2,2,1)$, and $H_8 = (2,2,2)$. Here both H_7 and H_8 include the disease-susceptibility allele d, whereas the other six haplotypes include the wild-type allele D. For each $i = 1, ..., 8$, let h_i be the frequency of H_i. We set 4 types of haplotype distribution: $(h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8) = (.125, .125, .125, .125, .125, .125, .125, .125)$, $(.343, .147, .147, .063, .147, .063, .027)$, $(.343, .147, .210, .000, .210, .000, .063, .027)$, and $(.700, .000, .000, .000, .210, .000, .063, .027)$, denoted as 'E', 'UE1', 'UE2', and 'UE3', respectively. Note that the degree of LD between each marker at loci, l_1 and l_2, and the disease marker is high ($D' = 1$), while the LD between the marker at a locus l_3 and the marker for the disease is relatively low ($D' = 0$ for E and UE1, $D' = 0.15$ for UE2, and $D' = 1$ for UE3). The subjects with the d allele have experienced mutations at loci, l_1 and l_2, whereas the mutation at a locus l_3 happened irrespective of the disease marker allele. Both dominant and recessive models were considered for the mode of inheritance. We assumed Hardy-Weinberg equilibrium and random-mating and that the families were ascertained through one affected offspring.

For the dominant genetic models, if an offspring has at least one of the high risk haplotypes, H_7 and H_8, its penetrance equals to $c \times RR$; otherwise, the individual who does not carry the disease-susceptibility allele d may be affected with a probability of c. For the recessive genetic models, only subjects having both H_7 and H_8, H_7 homozygote, or H_8 homozygote may have the penetrance of $c \times RR$, and the rest may be affected with a probability of c. We set c as 0.1 and RR as 1, 2, 3, and 4. Here RR = 1 corresponds to level and the rest to power. We ascertained 100 families and 200 independent samples were generated for each simulation model. We estimated the significance level of the tests based on 100 randomly generated samples from each sample in the study of type I error rate and power. Table 1 summarizes the estimated type I error rates and powers of the proposed and Zhao et al's tests at a significance level of 0.05. Note that the standard error of the type I error rate estimate is 0.015. Additionally, entries in columns denoted by 'p_f' and 'p_o' represent the average percentages of perfectly estimated haplotypes for 100 families, respectively, when based on family trios and based on parents alone in estimating haplotype frequencies.

As mentioned by Rohde and Fuerst [15], the percentage of perfectly estimated haplotypes is higher when estimates are based on family trios rather than on parents alone. Therefore, using the offspring's genotype can reduce uncertainty in parental haplotypes. As expected, the lower the diversity of haplotypes, the higher the percentage of perfect haplotype estimates. For example, the percentage of perfectly estimated haplotypes for the UE1 case was higher than for the E case, and the percentage was higher for the UE2 or UE3 case than for the UE1 case. However, there was no difference in percentages between dominant and recessive models.

Based on our results with the proposed test, shown in entries corresponding to RR = 1, the estimated type I error rate seems to satisfy a nominal level of 0.05, regardless of the mode of inheritance and type of haplotype distribution. For the dominant models, the power of the test increases as RR increases from 2 to 4, irrespective of the type of haplotype distribution. In the recessive models, the power increases as RR increases for the E case, while the power equals to the nominal level for the UE1, UE2, and UE3 cases. This explains why the probability of a proband having the mutant homozygotes is very low at 0.006 for the UE1, UE2, and UE3 cases relative to 0.047 for the E case.

We also compared the proposed test and Zhao et al's test in terms of type I error rate and power. Both tests showed comparable performance at a significance level of 0.05. For the E case, Zhao et al's test was more powerful than our test; for the UE1, UE2, and UE3 cases, however, the

Table 1: Empirical type I error rates and powers of the proposed (T_s) and Zhao et al's (T_z) tests, and average percentages of perfectly estimated haplotypes for the dominant and recessive models according to 4 types of haplotype distribution (HD)

HD	RR	Dominant				Recessive			
		p_f	p_o	T_s	T_z	p_f	p_o	T_s	T_z
E	1	88	72	.060	.065	88	72	.045	.050
	2	88	73	.275	.280	88	72	.145	.110
	3	88	73	.635	.645	89	73	.275	.275
	4	89	74	.815	.820	89	72	.455	.460
UE1	1	92	79	.060	.065	92	79	.050	.040
	2	92	79	.240	.235	92	79	.060	.050
	3	92	79	.560	.545	92	79	.050	.055
	4	92	79	.845	.840	93	79	.065	.055
UE2	1	99	94	.050	.065	99	94	.075	.085
	2	98	92	.270	.250	99	93	.035	.040
	3	98	90	.610	.580	99	93	.065	.055
	4	97	89	.885	.860	99	93	.065	.060
UE3	1	100	100	.040	.055	100	100	.090	.075
	2	100	100	.375	.325	100	100	.055	.065
	3	100	100	.795	.780	100	100	.060	.055
	4	100	100	.960	.950	100	100	.070	.060

The rows corresponding to RR = 1 are for type I error rate, and the rows corresponding to RR = 2, 3 or 4 are for power. The columns corresponding to p_f and p_o are the average percentages of perfectly estimated haplotypes, respectively, when based on family trios and based on parents alone in estimating haplotype frequencies.

Table 2: Empirical and asymptotic type I error rates of the proposed (T_s) and Zhao et al's (T_z) tests for the dominant model according to all the combinations of 4 pairs of haplotype distributions (HDs) and 3 types of proportion for the two populations in sample size (PP)

Pair of HDs	PP	Empirical		Asymptotic	
		T_s	T_z	T_s	T_z
E & E-3	P11	.055	.045	.025	.020
	P13	.050	.045	.015	.030
	P31	.060	.035	.035	.040
UE1 & UE-3	P11	.060	.055	.025	.035
	P13	.055	.075	.020	.025
	P31	.050	.040	.030	.030
UE2 & UE-3	P11	.060	.060	.000	.015
	P13	.035	.030	.000	.005
	P31	.070	.055	.020	.035
UE3 & UE-3	P11	.045	.040	.005	.020
	P13	.030	.035	.000	.010
	P31	.050	.045	.005	.005

P11, P13, and P31, respectively, combine two populations at the ratio of 1 to 1, 1 to 3, and 3 to 1.

latter had better power than the former. It seems that as the diversity of haplotypes decreases, the off-diagonal entries of $\hat{\mathbf{T}}$ correspondingly become more unbalanced and thereby are as informative as the marginal totals and this results in our test being more powerful than Zhao et al's test.

Finally, we performed simulations to investigate the robustness of the proposed test in cases of population admixture and to determine the conservativeness of the asymptotic chi-squared test. To this end, the haplotype distribution of a new population to be mixed with E and UE1, UE2 or UE3, respectively, was taken as (h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8) = (.250, .000, .250, .000, .250, .000, .250, .000) and (.490, .000, .210, 000, .210, .000, .090, .000), denoted as 'E-3' and 'UE-3'. These new populations have no mutation at a locus l_3 in common. We set c as 0.2 for the 'E-3' and 'UE-3' cases. We considered three types of proportion for the two populations in the sample size; 1:1, 1:3, and 3:1, denoted as 'P11', 'P13', and 'P31', respectively. The total number of ascertained families in the mixed population was 100. Table 2 presents the empirical and asymptotic type I error rates of our test along with Zhao et al's test with the dominant model at a nominal level of 0.05. Based on various situations incorporated in the simulations, we have shown that both our proposed test and Zhao et al's test are robust to population admixture. As mentioned in Section Score test, we also determined that the asymptotic chi-squared tests were conservative through simulations.

Kangwha study

The Kangwha study was performed to analyze the natural history of BP in Koreans in order to determine important factors associated with changes in BP [13]. In 1986, we initially constructed a cohort of 430 6-year old children who were living in Kangwha Province, Korea. The size of the cohort increased to 715 in 1992 and 784 in 1995. Our case group included blood samples from 101 students(61 boys and 40 girls) who experienced at least once systolic BP measurement of more than 130 mmHg or diastolic BP measurement of more than 85 mmHg between the ages of 15 and 17. Among the 101 samples from the case group, we were also able to obtain blood samples from the par-

ents of 40 subjects and our analysis was based on these 40 probands and their parents.

Angiotensins are substances smaller than proteins that act as vasoconstricting agents, i.e., they cause blood vessels to narrow. Narrowing the diameter of blood vessels increases the blood pressure. ACE, which is located on chromosome 17q23 and has a length of 26 kb, converts angiotensin to its activated functional form, angiotensin-II. There have been several studies investigating the relationship between the ACE gene and high BP [17]. In our analysis we included 4 SNPs, A-240T and C-93T on the promoter, I/D on intron 16, and A2350G on exon 17.

We used the Long, Williams, and Urbanek's method to determine the frequencies of haplotypes comprised of these 4 SNPs and estimated the frequencies of haplotypes, TTDG, ACIA, TTIA, TCIA, ATIA, and ACIG as 41.3, 53.7, 3.1, 0.6, 0.6, and 0.6%, respectively. When estimated by Rohde and Fuerst's method, there were no differences in the frequencies, except that TTDG and ACIA had frequencies of 41.2 and 53.8%, respectively. Table 3 shows the results of our proposed test and Zhao et al's test. The values in the row denoted by 'all SNPs' in the first column of the table represent test statistics and their *p*-values when all 4 SNPs were considered. The *p*-values were obtained by 1,000 randomizations. Neither test showed genetic linkage between ACE and high BP at a significance level of 0.05, but the observed values for each test were quite different.

In addition we examined whether any of the 4 SNPs were redundant and if exclusion of those SNPs could affect the outcomes of the two tests. We used a measure of PDE (proportion of diversity explained by a SNP set selection) proposed by Clayton [18] to select so-called htSNPs from the 4 SNPs being considered. This PDE acts like the coefficient of determination in an ordinary regression model. Using the HTSNP procedure [19] on 176 samples of the control group, we identified the htSNP set of C-93T and A2350G with a PDE of 0.992. We repeated the two tests using only the two htSNPs and listed the corresponding results in the row denoted by 'htSNPs' in the first column of Table 3. As was in the case when using all SNPs, using

Table 3: Observed values (*p*-value) of proposed test (T_s) and Zhao et al's test (T_z) for assessing genetic linkage between ACE gene and hypertension

SNPs included	T_s	T_z
all SNPs (A-240T,C-93T,I/D,A2350G)	5.275 (0.232)	7.363 (0.152)
htSNPs (C-93T,A2350G)	3.318 (0.332)	4.314 (0.163)

just the htSNPs resulted in no evidence of genetic linkage between ACE and hypertension at a level of 0.05.

Discussion
Here we propose a score test for linkage between genetic markers and disease-susceptibility genes based on haplotypes. First, we estimated haplotype frequencies using genotypes from affected offspring and their parents. As in [8], we constructed a transmission/non-transmission table of parents' haplotypes to the offspring. Then we proposed a test which mimics Sham's method [14] to test a marginal homogeneity in the transmission/non-transmission table.

Simulations indicate that our test works well at a nominal significant level. Further, we found that the power of the test is monotone and increases as the relative risk increases for dominant models. For recessive models with an unequal haplotype distribution, however, the test was highly conservative due to a low probability of a proband having the mutant homozygotes. In comparison with Zhao et al's test, their test had better power than our test for E case, while our test was more powerful than their test for the UE1, UE2, and UE3 cases. This implies our test has an advantage over Zhao et al's test in the situations where the diversity of haplotypes is low. We also found that our proposed test as well as Zhao et al's test is robust to population admixture. Although both asymptotic chi-squared tests were conservative, Zhao et al's test seemed to be less conservative than our test.

Conclusion
We propose a score test for linkage between genetic markers and disease-susceptibility genes and show through simulations that our test performed well at a nominal significant level. Our test showed the better performance in power than Zhao et al's test under some situations where the diversity of haplotypes is low. For an application to a Kangwha adolescent cohort, there was no remarkable evidence of genetic linkage between ACE and hypertension.

Authors' contributions
JK and CN contributed to the formulation of the score test and performed the simulations to investigate its performance. DK applied our test to the Kangwha study data and interpreted the results. All authors read and approved the final manuscript.

Abbreviations
ACE: Angiotensin-I converting enzyme

BP: Blood pressure

EM: Expectation-Maximization

htSNP: Haplotype tagging single nucleotide polymorphism

RR: Relative risk

TDT: Transmission/disequilibrium test

Appendix
If the haplotype phases for each parent were identified, the observed set of genotypes is compatible with only one haplotype group. Then $\hat{t}_{ij} = t_{ij}$ and thereby $\hat{\boldsymbol{\Delta}} = (t_{1\cdot} - t_{\cdot 1}, ..., t_{k-1\cdot} - t_{\cdot k-1})'$, where t_{ij} is the number of parents with haplotypes $H_i H_j$ who transmit H_i to the affected offspring. Thus the table $\hat{\mathbf{T}} = \{t_{ij}\}$ can be simply considered as a transmission/non-transmission table corresponding to the case of a locus with k marker alleles because there are no ambiguities in the parental haplotypes. Let $n = 2\sum_{g=1}^{G} n_g$. Define p_{ij} as the probability of each parent with haplotypes $H_i H_j$ transmitting H_i to the affected offspring, and the marginal probabilities as $p_{i\cdot} = \sum_{j=1}^{k} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{k} p_{ij}$. Following the arguments of Stuart [20], $\sqrt{n}[\boldsymbol{\Delta} - E(\boldsymbol{\Delta})]$ asymptotically follows a $(k - 1)$-variate normal distribution with covariance $n\Sigma = \{n\sigma_{ij}\}$, where $E(\hat{\boldsymbol{\Delta}}) = E(n(p_{1\cdot} - p_{\cdot 1}), ..., n(p_{k-1\cdot} - p_{\cdot k-1}))'$ and $\sigma_{ii} = n[(p_{i\cdot} + p_{\cdot i} - 2p_{ii}) - (p_{i\cdot} - p_{\cdot i})^2]$ and for $i \neq j$, $\sigma_{ij} = -n[(p_{ij} + p_{ji}) + (p_{i\cdot} - p_{\cdot i})(p_{j\cdot} - p_{\cdot j})]$. From Sham and Curtis [21], under no linkage, $p_{i\cdot} = p_{\cdot i}$, $i = 1, ..., k$. Therefore, to test linkage we can use the marginal homogeneity test statistic. Under no linkage, we have $E(\hat{\boldsymbol{\Delta}}) = (0, ..., 0)'$ and, plugging the maximum likelihood estimates, $\{t_{ij}/n\}$, $\{t_{i\cdot}/n\}$, and $\{t_{\cdot j}/n\}$, of $\{p_{ij}\}$, $\{p_{i\cdot}\}$ and $\{p_{\cdot j}\}$, respectively into the entries of Σ, the estimated covariance matrix $n\hat{\boldsymbol{\Sigma}} = \{n\hat{\sigma}_{ij}\}$, where $\hat{\sigma}_{ii} = t_{i\cdot} + t_{\cdot i} - 2t_{ii}$ and for $i \neq j$, $\hat{\sigma}_{ij} = -(t_{ij} + t_{ji})$. Hence T_s asymptotically follows a chi-square distribution with $(k - 1)$ degrees of freedom.

References
1. Clark AG: **Inference of haplotypes from PCR-amplified samples of diploid population.** *Mol Biol Evol* 1990, **7**:111-122.
2. Long JC, Williams RC, Urbanek M: **An E-M algorithm and testing strategy for multiple-locus haplotypes.** *Am J Hum Genet* 1995, **56**:799-810.

3. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68:**978-989.
4. Zhao JH, Curis D, Sham PC: **Model-free analysis and permutation tests for allelic associations.** *Hum Hered* 2000, **50:**133-139.
5. Wilson SR: **On extending the transmission/disequlibrium test (TDT).** *Ann Human Genet* 1997, **61:**151-161.
6. Clayton D, Jones H: **Transmission/disequilibrium tests for extended marker haplotypes.** *Am J Hum Genet* 1999, **65:**1161-1169.
7. Clayton D: **A generalization of the transmission/disequilibrium test(TDT) for uncertain haplotype transmission.** *Am J Hum Genet* 1999, **65:**1170-1177.
8. Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenaauer DB, Sun F, Kidd KK: **Transmission/disequilibrium tests using multiple tightly linked markers.** *Am J Hum Genet* 2000, **67:**936-946.
9. Spielman RS, Ewens WJ: **The TDT and other family-based tests for linkage disequilibrium and association.** *Am J Hum Genet* 1996, **59:**983-989.
10. Cordell HJ, Clayton DG: **A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes.** *Am J Hum Genet* 2002, **70:**124-141.
11. Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM: **Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics.** *Genet Epidemiol* 2004, **26:**61-69.
12. Rabinowitz D, Laird N: **A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information.** *Hum Hered* 2000, **50:**211-223.
13. Suh I, Nam CM, Jee SH, Kim SI, Lee KH, Kim HC, Kim CS: **Twelve-year tracking of blood pressure in Korean school children : the Kangwha study.** *Yonsei Med J* 1999, **40:**383-387.
14. Sham P: **Transmission/disequilibrium tests for multiallelic loci.** *Am J Hum Genet* 1997, **61:**774-778.
15. Rohde K, Fuerst R: **Haplotying and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information.** *Hum Mutat* 2001, **17:**289-295.
16. Becker T, Knapp M: **Maximum-likelihood estimation of haplotype frequencies in nuclear families.** *Genet Epidemiol* 2004, **27:**21-32.
17. Keavney B, McKenzie CA, Connell JM, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, Farrall M: **Measured haplotype analysis of the angiotensin-I converting enzyme gene.** *Hum Mol Genet* 1998, **7:**1745-1751.
18. Clayton D: **Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci.** [http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf].
19. SAS Institute Inc: *SAS/Genetics 9.1 User's Guide, Cary* .
20. Stuart A: **A test homegeneity of the marginal distribution in a two-way classification.** *Biometrika* 1955, **42:**412-416.
21. Sham PC, Curtis D: **An extended transmission/disequilibrium test (TDT) for multi-allele marker loci.** *Ann Human Genet* 1995, **59:**323-336.