European Journal of Human Genetics (2011) 19, 915–920

www.nature.com/ejhg

ARTICLE

A novel approach for small sample size family-based association studies: sequential tests

Ozlem Ilk*[,1], Farid Rajabli[2], Dilay Ciglidag Dungul[3], Hilal Ozdag[3] and Hakki Gokhan Ilk[2]

In this paper, we propose a sequential probability ratio test (SPRT) to overcome the problem of limited samples in studies related to complex genetic diseases. The results of this novel approach are compared with the ones obtained from the traditional transmission disequilibrium test (TDT) on simulated data. Although TDT classifies single-nucleotide polymorphisms (SNPs) to only two groups (SNPs associated with the disease and the others), SPRT has the flexibility of assigning SNPs to a third group, that is, those for which we do not have enough evidence and should keep sampling. It is shown that SPRT results in smaller ratios of false positives and negatives, as well as better accuracy and sensitivity values for classifying SNPs when compared with TDT. By using SPRT, data with small sample size become usable for an accurate association analysis.
European Journal of Human Genetics (2011) **19,** 915–920; doi:10.1038/ejhg.2011.51; published online 23 March 2011

Keywords: transmission disequilibrium test; sequential probability ratio test; SNPs; simulation study; family-based association study

INTRODUCTION

A genome-wide association study ascertains through genome, potential genetic associations between genetic polymorphism and observable traits or disease.[1,2] Recent advances in DNA technology made it possible to build high-resolution single-nucleotide polymorphism (SNP) maps. A number of companies now offer mapping arrays that have been frequently used in genome-wide association studies.

Two commonly used approaches for genome-wide studies in complex diseases are case–control and family-based approaches. The problem of population stratification in a case–control design is solved in family-based approach.[3] For this purpose transmission/disequilibrium test (TDT) has been frequently used, which was initially proposed by Spielman *et al*.[4] At the time when this manuscript was written, this paper was cited by more than 2500 studies according to the Web of Science. However, this popular test requires large number of trios (mother–father–offspring) to attain a reasonable statistical power. Especially, for diseases with late age of onset, it is difficult to find alive parents.[5] Consequently, it is usually difficult to gather sufficient number of complete trios.[3]

Sequential tests, on the other hand, consider one sample at a time to decide on whether there is sufficient information in favor of one of the hypotheses or more samples are needed. They are especially useful when it is not practical to fix the sample size in advance. Increasing the sample size is frequently costly, either in monetary or non-monetary terms. One such instance occurs when the experiment requires breaking parts in a quality-control experiment. Another example involves collecting samples from both parents of a patient, in which the disease, for instance cancer, progress during late ages. Sample size has utmost importance in such studies as such diseases are usually developed in the elderly. Obtaining family-based information is almost impossible from these patients' parents because few of them are alive and willing to give samples.

The motivation of our study is that sequential tests have been used in clinical trials with a goal of decreasing cost and possible ethical issues.[6,7] However, they have not been applied in genome-wide association studies. In this paper, we propose a sequential hypothesis testing approach for family-based association studies. Our approach provides the detection of interesting genetic associations even with small sample sizes.

METHODS

Consider a situation in which a child is affected by a disease. For simplicity, assume that each family has only one affected offspring. Moreover, suppose that data from both parents of this child are available. In other words, it is assumed that we have complete mother–father–child trios with no missing cases in our sample. Our goal is to find the SNPs for which we can detect an association between the genetic marker and disease locus. However, a serious challenge is to find a sufficiently large number of trios. In this section, statistical methods are discussed briefly to overcome such a challenge. Methods that are described in this section are applied to a simulated data in the next section.

Transmission/disequilibrium test

First proposed by Spielman *et al*,[4] the TDT is a family-based association test. It uses heterozygous parents (possessing two different alleles for a single trait) for an allele to test the linkage disequilibrium and/or association between the genetic marker and a susceptible disease gene.[3]

Considering a biallelic marker situation and denoting the markers by M_1 and M_2, a classical TDT test makes use of a 2×2 table. In such a table, let n_{12} be the number of heterozygous (M_1M_2) parents who transmit M_1, but not M_2, allele to their affected offspring. In a similar manner, n_{21} is the number of heterozygous parents who transmit M_2, but not M_1, allele to their affected offspring. The other two cells, n_{11} and n_{22}, are non-informative, as the transmitted and non-transmitted alleles are 'tied'.

The null hypothesis, *Ho*, is that there is either no association or no linkage between the marker and the disease. Whittaker and Morris[8] discussed that $n_{12} \sim \mathrm{Bin}(n_{12}+n_{21}, \tau)$, where Bin is binomial distribution, and that under the

[1]Department of Statistics, Faculty of Arts and Sciences, Middle East Technical University, Ankara, Turkey; [2]Ankara University, Faculty of Engineering, Department of Electronics Engineering, Ankara, Turkey; [3]Genomics Unit Of Central Laboratory, Ankara University Biotechnology Institute, Ankara, Turkey
*Correspondence: Dr O Ilk, Department of Statistics, Faculty of Arts and Sciences, Middle East Technical University, Ankara 06531, Turkey. Tel: +90 312 210 5326; Fax: +90 312 210 2959; E-mail: oilk@metu.edu.tr
Received 6 September 2010; revised 15 February 2011; accepted 25 February 2011; published online 23 March 2011

no association assumption τ is equal to 0.5. In other words, under *Ho*, the probability for transmission of a marker from a parent to a child, τ, is near 0.5 under the Mendelian inheritance. In the presence of association, this probability deviates from 0.5. They have also provided that the score test (ie, $\left[\frac{\partial \ell(\tau)}{\partial \tau}\right]^2 / \frac{-\partial^2 \ell(\tau)}{\partial \tau^2}$, where $\ell(\tau)$ is the log-likelihood function) evaluated at τ=0.5 leads to $T = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$. This test is known as transmission/disequilibrium test (TDT). It is known that this test statistic has asymptotically χ^2 distribution with 1 degrees of freedom under H_0. However, if $n_{12}+n_{21}<25$, then the use of this asymptotic distribution is not advised.

Sequential tests

Let τ be the probability of transmitting allele M_1 from a heterozygous parent. Consider the hypothesis testing problem with a null hypothesis that H_0: $\tau=\tau_0$ *versus* the alternative H_1: $\tau=\tau_1$. Here τ_0 is different from τ_1. A sequential probability ratio test (SPRT) is, then, defined as the following.

$$\begin{cases} \text{Accept } H_0 \text{ if } \lambda_m \geq k_1 \\ \text{Reject } H_0 \text{ if } \lambda_m \leq k_0 \qquad\qquad \text{where,} \\ \text{Continue sampling if } k_0 < \lambda_m < k_1 \end{cases} \tag{1}$$

$$\lambda_m = \lambda_m(X_1, \ldots, X_m) = \frac{f(X_1, \tau_0) \ldots f(X_m, \tau_0)}{f(X_1, \tau_1) \ldots f(X_m, \tau_1)} \text{ for } m = 1, 2, \ldots$$

Here k_0 and k_1 are thresholds that depend on type I (α) and type II (β) errors. The approximate thresholds are given in equation (2).

$$k_0 = \frac{\alpha}{1-\beta},\ k_1 = \frac{1-\alpha}{\beta} \tag{2}$$

Using the fact that $n_{12} \sim \text{Bin}(n_{12}+n_{21}, \tau)$, the SPRT in equation (1) becomes,

$$\lambda_m = \frac{\tau_0^{n_{12}}(1-\tau_0)^{n_{21}}}{\tau_1^{n_{12}}(1-\tau_1)^{n_{21}}} = \left(\frac{\tau_0}{\tau_1}\right)^{n_{12}} \left(\frac{1-\tau_0}{1-\tau_1}\right)^{n_{21}} \text{ for } m = 1 \tag{3}$$

For later samples, the statistic in equation (3) is iterated, but now the frequencies n_{12} and n_{21} are calculated by using all available m samples.

As the association levels for τ_1 and $1-\tau_1$ are same, SPRT results are combined for these two levels. Specifically, following Wetherill,[9] we conclude that there is no association if tests for both τ_1 and $1-\tau_1$ values accept H_0. If one of them accepts H_1, then we conclude that there is association. By this way, we are converting one-sided SPRT to two-sided test. This is necessary for two reasons: to accommodate a better comparison with TDT, which is a two-sided test; and to combine τ_1 and $1-\tau_1$ values that correspond to the same level of association.

Sample size comparison

There is a rich literature on the discussion of sample size and power calculations for family-based studies. Simulation studies under different assumptions suggest that one should have hundreds and maybe even thousands of trios in a family-based association study to gain a reasonable power.[10] Sequential tests, on the other hand, do not require a fixed specific sample size to start with. Note that, in a sequential test, the sample size is also a random variable. Wald[11] provides formula for calculating the expected sample size, $E_\tau(n)$, for a binomial sequential test as a function of the success probability,τ. This formula is provided in equation (4), where $L(\tau)$ is the probability of accepting H_0 given the value of τ.

$$E_\tau(n) = \frac{L(\tau)\log(\beta/(1-\alpha)) + (1-L(\tau))\log((1-\beta)/\alpha)}{\tau\log(\tau_1/\tau_0) + (1-\tau)\log((1-\tau_1)/(1-\tau_0))} \tag{4}$$

Here α and β are type I and II errors, respectively; τ_0 and τ_1 are the probabilities of transmission under H_0 and H_1, respectively; and τ is the 'true' probability of transmission. Please note that $E_\tau(n)$ given in equation (4) represents the expected number of observations needed by the sequential test procedure. Therefore, it is not used to conclude in favor of H_0 or H_1.

The expected sample number curve usually increases as τ increases from 0 to τ_0 and decreases as τ goes from τ_1 to 1. Between τ_0 and τ_1, the expected number of samples usually increases up to some point, let us say, τ', and then starts decreasing. Therefore, the maximum expected sample size occurs either at τ_1 or near τ'. To obtain this maximum size, one should take the maximum of two values: the one obtained by replacing τ with τ_1 in equation (4) and the one obtained by using equation (5):

$$E_{\tau'}(n) = \frac{-(\log(\beta/(1-\alpha)))(\log((1-\beta)/\alpha))}{\log(\tau_1/\tau_0)\log((1-\tau_0)/(1-\tau_1))} \tag{5}$$

Some descriptive statistics of expected sample sizes are calculated from a theoretical perspective by using equations (4) and (5) and presented in Figure 1. Note that, the average in this figure refers to the average of the sample sizes taken for all possible τ values.

As τ_1 gets closer to the value under H_0 (0.5), the number of samples that is necessary to detect the association increases. That is anticipated, as one needs more evidence through data to distinguish between H_0 and H_1 in such situations. However, even when τ_1 is 0.4 or 0.6, the expected sample size ranges between 15 (minimum) and 305 (maximum). When one wants to test against τ_1=0.3 or 0.7, the maximum number of expected trios is 73. In other words, by using SPRT, we can expect to detect the SNPs associated with the disease with only 73 trios, when the probability for transmission of a marker allele from a parent to an affected offspring is $\leq$30 or $\geq$70%. On the other hand, much larger number of trios is reported to be required for TDT. According to TDT power calculator,[10] 547 trios are necessary to attain the same power (80%) at the same type I error (0.001) rate. Indeed, through simulation studies, it was reported that SPRT, on the average, requires smaller samples compared with a test with fixed sample size.[6]

Simulation study

The performance of the sequential test and TDT are evaluated with the same simulation data. It is assumed that there are 270 000 SNPs and at most 200 trios in the simulated data set. Considering the multiple testing problem, type I error is set to be very small (α=0.001). As earlier simulation runs provided inflated type I error estimates for TDT, but not for SPRT, Benjamini–Hochberg procedure is applied on TDT with a goal of controlling type I error rate. Power of the test is taken as 0.8 (ie, β=0.2).

To simulate matched pairs, the following scenario is used. In similar studies, data are usually summarized as (n_{12}, n_{21}, n–n_{12}–n_{21}). The concordant pairs (n_{11} and n_{22}) are non-informative for both TDT and SPRT test statistics, and hence one does not need their exact values. Moreover, one can summarize the data as the pair of (n_{12}, n_{21}), as, obviously, n–n_{12}–n_{21} is a function of this pair. Therefore, we only simulated (n_{12}, n_{21}) pairs. For a 2×2 matched-pairs data, the conditional distribution of n_{21} given an observed value of n_{12} is binomial.[12] Specifically, this conditional distribution is given in equation (6). In this equation, τ_{12} is the probability of transmitting allele M_1 from a parent, whereas τ_{21} is the probability of transmitting allele M_2.

$$n_{21}|n_{12} \sim \text{Bin}(n - n_{12}, \frac{\tau_{21}}{1-\tau_{12}}) \tag{6}$$

Therefore, for the first trio, we first generated n_{12} observations from binomial distribution with a sample size of 1 and success probability τ_{12}. Next, given this

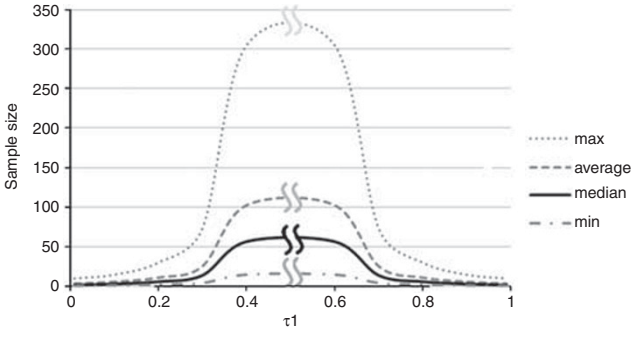

Figure 1 Minimum, maximum, median and average number of expected sample sizes required in SPRT to gain 80% power with a 0.1% type I error for different values of τ_1. Note that the sample size takes the value of infinity for τ_1=0.5.

observation, n_{21} is generated from equation (6) with $n=1$ to create a matched pair. For instance, when τ_{12} and τ_{21} are assumed to be 0.1 and 0.8, respectively, n_{12} is generated from a Bin(1, 0.1) and n_{21} is generated from Bin($1-n_{12}$, 0.8/0.9). In this case, we will most probably obtain a simulated pair of (0, 1) for (n_{12}, n_{21}). This process is repeated for the next trios.

Simulation is hold for different values of genotypic risk ratio (GRR). It was suggested that GRR can be approximated by n_{12}/n_{21}.[13,14] Therefore, the ratio of the probabilities (τ_{12}/τ_{21}) is used to obtain either one of the association levels: no significant association (GRR$<$1.5), moderate association (1.5$\leq$GRR$\leq$3.5) and high association (GRR$>$3.5). This classification according to GRR values are assigned by following Kharrat *et al*.[14] They reported that in complex diseases, most associated genes have low or medium GRR values (between 1.5 and 3.5). Under the null hypothesis, the informative cells, n_{12} and n_{21}, are equally likely, and hence the association ratio is close to 1. As this ratio shifts from 1, the association between the disease and the marker is expected to increase as well, and hence, we are more likely to reject H_0. In our study, all possible combinations of τ_{12} and τ_{21} are considered. Specifically, a sequence between 0.1 and 0.9 with increments of 0.1 is assigned to both τ_{12} and τ_{21}. This resulted in 45 different combinations. Although highly associated SNPs do not have biological sense, they are included to present the whole picture. Next, the ratio of these two probabilities is calculated to detect the corresponding association group.

Out of 270 000 SNPs, 78 000 are assumed to have no significant association. In other words, these SNPs are generated under H_0. To generate 'truly positive' SNPs, that is, SNPs under H_1, the existence of two levels of association is assumed. Although a total of 108 000 SNPs are generated with a moderate amount of association, another 84 000 SNPs are generated under the assumption of a high association. This simulation study is repeated 100 times. Both TDT and SPRT are applied to the simulated data. Using equation (2), we calculated the boundaries for SPRT as k_0=0.00125 and k_1=4.995. In both TDT and SPRT, the same nominal α (0.1%) and β (20%) values are used. Simulation study is hold in MATLAB and the related code is available on request.

We should warn the reader that the alternative hypothesis under TDT and SPRT may be different. For instance, even τ=0.500001 belongs to the alternative hypothesis in TDT. One should note, though, that testing against such an alternative will lead to the conclusion of H_1 with only very low power (as low as type I error). In other words, if a researcher uses such a value in the alternative, she/he will not be able to detect many associated SNPs even with enormous sample sizes. Moreover, the ones that are marked as associated will most probably be false-positive ones. We believe this is another advantage of SPRT. In SPRT, the researcher can focus on the alternatives of interest. Screening τ values that are far away from the value under H_0 will obviously be more practical. Unfortunately, TDT does not allow this. SPRT, on the other hand, is using a combination of say, 0.3 and 0.7. Although it seems as SPRT is detecting associated SNPs only at these two τ_1 values, it is detecting SNPs that are at a higher association level as well. A simple calculation of odds ratio shows us that, an SPRT with τ_1=0.7 attempts to detect the SNPs that are at least 2.3 (=(0.7/0.3)/(0.5/0.5)) times more likely to be associated compared with non-associated SNPs. On the other hand, an SPRT with τ_1=0.8 attempts to detect the SNPs that are at least four times more likely to be associated. Therefore, an alternative hypothesis of τ_1=0.3 or 0.7 is equivalent to an H_1 of $\tau_1\leq$0.3 or $\geq$0.7. This implies that SPRT skips the τ_1 values that are between 0.3 and 0.7 (except 0.5 under H_0), and focuses on the parameter space that we most expect to see the association. Note that, an SPRT with τ_1=0.500001 can also be constructed. However, this will attempt to detect the SNPs that are at least 1.00004 times more likely to be associated, and this is a waste of time.

RESULTS

Results of true and false positives and negatives are reported with a categorization of association level. Figure 2 presents the results for which the sequential test requires the continuation of sampling for different number of trios.

For 84 000 SNPs that are generated under the high association, the curve approaches to zero quite fast. Even with 50 trios, SPRT can classify $>$60% of these SNPs. For $>$80 trios, only 10% or less of the high-associated SNPs is left for which we cannot conclude to either H_0 or H_1. In other words, if one can afford only 80 trios, then by using SPRT she/he will be able to detect around 90% of the SNPs that are not associated or highly associated with the disease. Moreover, one can detect around 65% of the 'no significant association' SNPs with only 80 trios. SNPs with moderate association require considerably more samples to decide. Specifically, only 36% of them can be classified as either positive or negative with 80 trios. The rest of them, 64%, are still in the gray area. One needs to gather about 200 trios if she/he wants to classify $>$80% of these SNPs. This makes sense, as this group is generated under H_1, but has GRR values that are close to the ones generated under H_0. Therefore, algorithm needs more information through data to conclude one of the hypotheses. Nevertheless, it can be shown that a sequential test will certainly stop and conclude to either H_0 or H_1 in a finite number of samples.[15] Furthermore, with 40 trios and more, results calculated by SPRT are realized with accuracy over 90%, which is calculated according to equation (7):

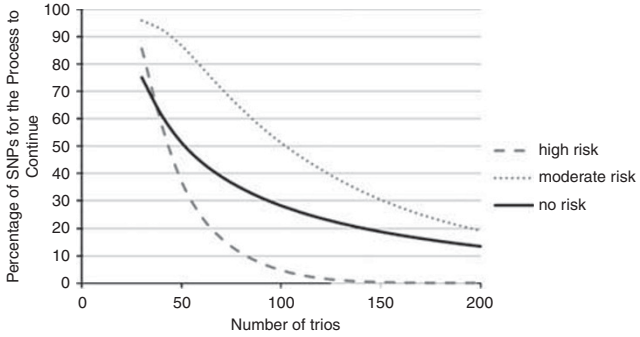

Figure 2 Percentage of SNPs for which the proposed algorithm requires more samples to decide for α=0.1% and β=20%.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{FP+FN+TP+TN}} \quad (7)$$

However, TDT reaches the 90% accuracy when 200 trios are available. Overall accuracy of SPRT and TDT tests by the number of trios are given in Table 1.

When data are generated under high association assumption, true-positive (TP) and false-negative (FN) percentages are presented for both methods in Table 2. This table also includes the sensitivity, which is calculated according to equation (8):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (8)$$

Similar percentages are provided in Table 3 for moderate association SNPs.

As it is seen from these tables, TP and FN columns for TDT add up to 100% (except for some rounding issues). However, for SPRT, the summation of three percentages, namely TP, FN and continue sampling, gives 100%. The information on the percentage of truly positive SNPs for which to continue sampling is clearly available only for SPRT method, as TDT classifies each SNP as either positive or negative. Because of this classification difference, the power (TP) of our test seems lower than the one for TDT. For instance, for high-association SNPs, when there are 80 trios available, the power for TDT is 94.4%, whereas it is 89.1% with SPRT. TDT dumps the rest of the power into FN, but SPRT distributes the rest into two categories (FN and continue sampling); SPRT therefore results in much lower false classifications. For instance, for 80 trios, the FN percentage for TDT is observed as 5.6%, whereas it is only 0.1% for SPRT. This difference in the classification also leads to a difference in the calculation of accuracy, sensitivity and specificity. Note that, these calculations for

SPRT do not involve suspicious SNPs (for which we need to continue sampling). For instance, the denominator of equation (7) is equal to the total number of SNPs for which the null hypothesis is rejected or failed to be rejected for both methods. However, this is equal to the total number of SNPs in the study (270 000) for TDT, whereas it is the total number of SNPs minus those that are suspicious for SPRT.

It is also interesting to see that as the number of trios increase, the power of SPRT increases, but the percentage of false negatives stay constant. False negatives captured by the sequential test stayed almost constant for larger number of trios, whereas the percentage of samples for which we need to continue sampling decreased considerably. This is because the algorithm carries the SNPs that were not classified beforehand to correctly classified group. In other words, SNPs for which we are advised to continue sampling are then correctly detected as positive as the number of trios increase.

For small number of trios, the powers of both TDT and SPRT are quite small, as one would expect. To obtain a power $>80\%$ with SPRT, one needs to obtain at least 70 trios to detect highly associated SNPs. To detect moderately associated SNPs with a power of 80%, one, unfortunately, needs >200 trios. However, it should be noted that the power for TDT is also low for this group of SNPs (Table 3).

It is striking how high the false-negative percentages for TDT are. Especially for moderate association SNPs, the percentages of FNs are $>80\%$ for small number of trios. That means, for small number of trios with moderate association SNPs, the use of TDT would mean being receptive to a great deal of wrong decisions.

Table 4 presents the TN, FP percentages and specificity for no significant association SNPs. Specificity is calculated according to equation (9):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{9}$$

It is clear that FP's for TDT are larger compared with the ones for SPRT even after multiple testing correction is applied to TDT. The nominal value of 0.1% is attained in SPRT when there are about 80 trios. Although after that point, this error rate increases, it stays well below the one for TDT. In other words, FP rates that are close to the nominal value are attained by SPRT even without a multiple testing correction. The specificities for both tests are observed to be high and close to each other.

Table 1 Overall accuracy of two tests by the number of trios for α=0.1% and β=20%

	Overall accuracy (%)	
Number of trios	*SPRT*	*TDT*
30	87.9	39.9
40	91.5	50.0
50	92.7	57.9
60	93.3	63.3
70	93.6	67.9
80	93.7	71.9
90	93.8	74.8
100	93.9	77.6
110	94.0	79.8
120	94.1	81.7
130	94.1	83.3
140	94.1	84.7
150	94.1	85.9
160	94.1	86.8
170	94.1	87.8
180	94.1	88.5
190	94.0	89.2
200	94.0	89.8

Abbreviations: SPRT, sequential probability ratio test; TDT, transmission disequilibrium test.

DISCUSSION AND CONCLUSION

Sample size is a crucial issue in genome wide family-based association studies. The collection of hundreds of trios that is necessary for TDT analysis is a challenging task. It needs a well-coordinated and collaborative effort. In addition, for late onset complex diseases it is

Table 2 Percentage of true positives, false negatives and sensitivity by the number of trios for high association SNPs

	TP (percentage of 'truly positive' SNPs for which H_0 is rejected), %		*FN (percentage of 'truly positive' SNPs for which $H_0$is mistakenly accepted), %*		*Sensitivity, %*	
Number of trios	*SPRT*	*TDT*	*SPRT*	*TDT*	*SPRT*	*TDT*
30	14.6	33.9	0.079	66.1	99.5	33.9
40	42.7	61.1	0.104	38.9	99.8	61.1
50	63.3	77.8	0.107	22.2	99.8	77.8
60	76.0	85.8	0.107	14.2	99.9	85.8
70	83.7	91.2	0.107	8.83	99.9	91.2
80	89.1	94.4	0.107	5.59	99.9	94.4
90	92.9	96.4	0.107	3.64	99.9	96.4
100	95.3	97.7	0.107	2.33	99.9	97.7
110	97.0	98.6	0.107	1.38	99.9	98.6
120	98.2	99.2	0.107	0.80	99.9	99.2
130	98.9	99.5	0.107	0.46	99.9	99.5
140	99.3	99.7	0.107	0.26	99.9	99.7
150	99.6	99.9	0.107	0.14	99.9	99.9
160	99.7	99.9	0.107	0.09	99.9	99.9
170	99.8	100	0.107	0.04	99.9	100
180	99.8	100	0.107	0.025	99.9	100
190	99.9	100	0.107	0.018	99.9	100
200	99.9	100	0.107	0.004	99.9	100

Abbreviations: FN, false negative; SNP, single-nucleotide polymorphism; SPRT, sequential probability ratio test; TDT, transmission disequilibrium test, TP, true positive.

Table 3 Percentage of true positives, false negatives and sensitivity by the number of trios for moderate association SNPs

	TP (percentage of 'truly positive' SNPs for which H_0 is rejected), %		*FN (percentage of 'truly positive' SNPs for which $H_0$is mistakenly accepted), %*		*Sensitivity, %*	
Number of trios	*SPRT*	*TDT*	*SPRT*	*TDT*	*SPRT*	*TDT*
30	0.1	1.1	4.0	98.9	2.1	1.1
40	1.5	5.3	5.7	94.7	21.3	5.3
50	6.2	12.0	7.0	88.0	46.7	12.0
60	12.8	19.4	8.0	80.6	61.7	19.4
70	19.8	26.7	8.8	73.3	69.3	26.7
80	26.5	34.1	9.4	65.9	73.7	34.1
90	32.7	40.0	10.0	60.0	76.6	40.0
100	38.0	46.0	10.4	54.0	78.5	46.0
110	42.8	51.1	10.7	48.9	80.1	51.1
120	47.3	55.5	10.9	44.5	81.3	55.5
130	51.4	59.5	11.1	40.5	82.2	59.5
140	54.8	62.9	11.3	37.1	82.8	62.9
150	57.7	66.1	11.5	33.9	83.4	66.1
160	60.4	68.7	11.6	31.4	83.9	68.7
170	62.8	71.3	11.7	28.7	84.3	71.3
180	64.8	73.5	11.9	26.5	84.5	73.5
190	66.7	75.6	12.0	24.4	84.8	75.6
200	68.4	77.3	12.1	22.7	85.0	77.3

Abbreviations: FN, false negative; SNP, single-nucleotide polymorphism; SPRT, sequential probability ratio test; TDT, transmission disequilibrium test, TP, true positive.

Table 4 Percentage of true negatives, false positives and specificity by the number of trios for no significant association SNPs

	TN (percentage of 'truly negative' SNPs for which H_0 is accepted), %		*FP (percentage of 'truly negative' SNPs for which H_0 is mistakenly rejected), %*		*Specificity, %*	
Number of trios	*SPRT*	*TDT*	*SPRT*	*TDT*	*SPRT*	*TDT*
30	24.5	100	0	0	100	100
40	38.3	100	0	0.008	100	100
50	48.6	100	0	0.027	100	100
60	55.7	100	0.015	0.046	100	100
70	60.9	99.9	0.027	0.081	100	99.9
80	65.2	99.8	0.058	0.154	99.9	99.8
90	68.6	99.8	0.127	0.223	99.8	99.8
100	71.6	99.6	0.192	0.358	99.7	99.6
110	74.0	99.4	0.281	0.608	99.6	99.4
120	75.8	99.2	0.385	0.792	99.5	99.2
130	77.7	98.9	0.515	1.088	99.3	98.9
140	79.1	98.7	0.638	1.335	99.2	98.7
150	80.4	98.2	0.781	1.777	99.0	98.2
160	81.6	97.8	0.950	2.154	98.8	97.8
170	82.7	97.5	1.065	2.546	98.7	97.5
180	83.7	97.0	1.212	2.962	98.6	97.0
190	84.5	96.5	1.350	3.538	98.4	96.5
200	85.3	96.0	1.496	3.977	98.3	96.0

Abbreviations: FP, false positive; SNP, single-nucleotide polymorphism; SPRT, sequential probability ratio test; TDT, transmission disequilibrium test, TN, true negative.

difficult to reach complete (alive) parent–child trios. The inability to collect the necessary number of samples disables many researchers to analyse small sample size data by TDT. With the advent of SPRT, data with small sample sizes become usable for an accurate association analysis. We therefore propose the use of sequential hypothesis testing to detect the associated markers with the disease. This approach is not assuming a fixed sample size at the beginning of the biological experiment. Instead, after each observed trio or a group of observed trios, the test clarifies whether there is enough evidence to detect the association or no association or one should continue to take more trios to correctly classify SNPs. The test is especially beneficial when the suggested number of trios for TDT-type tests cannot be achieved.

In this paper, it was shown, through simulation studies, that the resulting accuracy for at least 40 trios by using SPRT is over 90%. This means that if one can afford at least 40 trios, SPRT classify SNPs with 90% accuracy. It was also shown, that false negatives and false positives obtained by SPRT are much smaller than that of TDT, especially for small number of trios. Our test is clearly more conservative than TDT,

as it is less likely to reject the null hypothesis, either it is true or false. This leads to a smaller power when compared with the results of TDT, but in return leads to a smaller type I error. The power of TDT is superior when the association between the disease and the marker is moderate. However, we suggest the use of TDT in such situations only if the number of trios is >150. For smaller number of trios, the percentages of false negatives in TDT are huge, ranging from 34 to 99%. In short, especially for small to moderate amount of trios, SPRT results in smaller ratios of false positives and negatives, as well as better accuracy and sensitivity values for classifying SNPs when compared with TDT.

The limitation of the approach proposed here is that the sample size required increases as the probability in alternative hypothesis gets closer to the one in null hypothesis. In other words, the expected sample size in sequential test becomes closer to the one needed in TDT as the association between the marker locus and disease locus gets weaker. However, the approach is highly beneficial over TDT, as, unlike TDT it does not require the asymptotic approximation of a distribution; it provides a gray region instead of a black and white categorization; and it results in small number of false predictions.

An exact binomial test or permutation tests can be used instead of asymptotic TDT test statistic in small samples. However, exact binomial and permutation tests are still limited, as they 'classify' the SNPs into two: those that are associated and those for which we do not have enough evidence. A sequential test, on the other hand, 'classifies' SNPs into three: those that are associated, those that are not associated and those for which we do not have enough evidence. In other words, SPRT follows a 'better to be safe than sorry' approach, and does not insist on classifying SNPs into one of the two categories. It clearly states so when one needs to gather more trios. This is one of the main advantages of SPRT; it recommends us to continue to sample when there is not enough trios, whereas TDT and similar large sample size theory tests result in false negatives or positives.

From a statistical point of view, obviously, one would like to gather as much sample as possible, which will lead to a higher power in any test. However, we should keep in mind the expected net gain of sampling (ENGS), which is defined as the difference between the expected value of sample information and the cost of sampling. Decreasing the cost is crucial in genetics. It was reported that the ENGS in SPRT with a maximum number of n samples is generally higher than the one with a fixed sample size test of size n when the cost of sampling is same.[16] This is consistent with the fact that in real life challenges, data are only available sequentially.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work has been supported through a project by The Scientific and Technological Research Council of Turkey (TUBITAK-107S348).

1 Cordell HJ, Clayton DG: Genetic association studies. *Lancet* 2005; **366**: 1121–1131.
2 Kruglyak L: The road to genome-wide association studies. *Nat Rev Genet* 2008; **9**: 314–318.
3 Evangelou E, Trikalinos TA, Salanti G, Ioannidis JP: Family-based *versus* unrelated case-control designs for genetic associations. *PLoS Genet* 2006; **2**: 1147–1155.
4 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–516.
5 Wang D, Sun F: Sample sizes for the transmission disequilibrium tests: TDT, S-TDT and 1-TDT. *Commun Stat Theory Methods* 2000; **29**: 1129–1142.
6 van der Lee JH, Wesseling J, Tanck MWT, Offringa M: Efficient ways exist to obtain the optimal sample size in clinical trials in rare diseases. *J Clin Epidemiol* 2008; **61**: 324–330.
7 van der Tweel I, van Noord PAH: Early stopping in clinical trials and epidemiologic studies for 'futility': conditional power *versus* sequential analysis. *J Clin Epidemiol* 2003; **56**: 610–617.
8 Whittaker JC, Morris AP: Family-based tests of association and/or linkage. *Ann Hum Genet* 2001; **65**: 407–419.
9 Wetherill GB: *Sequential Methods in Statistics*. London and New York: Chapman and Hall, 1975.
10 Chen WM, Deng HW: A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genet Epidemiol* 2001; **21**: 53–67.
11 Wald A: *Sequential Analysis*. New York: John Wiley & Sons, 1947.
12 Berger RL, Sidik K: Exact unconditional tests for 2×2 matched-pairs design. *Stat Methods Med Res* 2003; **12**: 91–108.
13 Schaid DJ: Likelihoods and TDT for the case-parents design. *Genet Epidemiol* 1999; **16**: 261–273.
14 Kharrat N, Ayadi I, Rebai A: Sample size computation for association studies using case-parent design. *J Genet* 2006; **85**: 187–191.
15 Bain LJ, Engelhardt: *Introduction to Probability and Mathematical Statistics*, 2nd edn. Boston: PWS-KENT Publication, 1992.
16 Winkler RL: *An Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart, and Winston, 1972.