Backward Haplotype Transmission Association (BHTA) Algorithm—A Fast Multiple-Marker Screening Method

Shaw-Hwa Lo Tian Zheng

Department of Statistics
Columbia University
New York, NY

Running title: BHTA Multiple-Marker Screening Method

Correspondence to: Shaw-Hwa Lo, Ph.D.
Department of Statistics
Mail Code 4403
2990 Broadway
Columbia University
New York, New York 10027
Tel: (212) 854-3639
Fax: (212) 663-2454
slo@stat.columbia.edu

Abstract

The mapping of complex traits is one of the most important and central areas of human genetics today. Recent attention has been focused on genome scans using a large number of marker loci. Because complex traits are typically caused by multiple genes, the common approaches of mapping them by testing markers one after another fail to capture the substantial information of interactions among disease loci. Here we propose a backward haplotype transmission association (BHTA) algorithm to address this problem. The algorithm can administer a screening on any disease model when case-parent trio data are available. It identifies the important subset of an original larger marker set by eliminating the markers of least significance, one at a time, after a complete evaluation of its importance. In contrast with the existing methods, three major advantages emerge from this approach. First, it can be applied flexibly to arbitrary markers, regardless of their locations. Second, it takes into account haplotype information; it is more powerful in detecting the multifactorial traits in the presence of haplotypic association. Finally, the proposed method can potentially prove to be more efficient in future genomewide scans, in terms of greater accuracy of gene detection and substantially reduced number of tests required in scans. We illustrate the performance of the algorithm with several examples, including one real data set with 31 markers for a study on the *Gilles de la Tourette* syndrome. Detailed theoretical justifications are also included, which explains why the algorithm is likely to select the “correct” markers.

Keywords: haplotype, transmission disequilibrium, screening, complex traits, epistatic loci, marker selection, genome scan.

Introduction

Novel statistical designs and association methods have been developed during the past two decades for identifying important genes that are responsible for a number of inherited human diseases [1]–[11]. However, success has been largely restricted to simple Mendelian diseases [12, 10, 13]. For more common and complex human disorders, the outcomes from similar statistical methods for detecting susceptibility loci are inconsistent and less encouraging [13, 11].

Recent studies of complex traits mapping have been focused on whole-genome scans, which are usually carried out on a dense map of polymorphic markers for the localization of the disease susceptibility loci. Such a search typically involves individual tests using a statistic that tests linkage/association repeatedly, either at each marker locus in turn or by combining a few nearby markers throughout the genome. Though this strategy of marker-wise search has the advantage of being easy to implement, it has the drawback of ignoring useful information. This is especially true when dealing with a multifactorial trait due to many epistatic genes, each of modest effect. Since complex traits may be governed by an unknown number of genes interactively [14, 15, 16, 17], the common approaches, using only marginal information from each marker, cannot capture the important information of the interactions among the disease susceptibility loci [18]. Thus, it is less likely for these methods to have adequate power to detect the mutated genes. On the other hand, detailed studies on all the markers in a genome scan and their possible interactions will result in overwhelming complexity, which can not be handled using current genetic data sets [19]. To address this dilemma, new methods were proposed recently. A 2-stage approach was described in [19]: (1) selecting a subset of markers that are important for the trait studied, and (2) modeling main effects and interactions of the selected markers on the trait of interest. For stage (1), a nested bootstrap procedure was proposed to achieve the marker selection. This approach involved a selection algorithm based on a function (sum as an example used in the papers) of K statistics, $T_1, T_2, \dots, T_K$, which represents measure of association between K individual

markers. T_i is calculated based upon genotype information of i^{th} marker only. The other approach [20] is developed in such a way that allelic association (for example, measured by T_i) and Hardy-Weinberg disequilibrium information (measured by U_i at each marker locus) are combined to form a “set association” score, $\sum_i T_i U_i$, which evaluates the sets of markers under study in terms of association to the disease trait. Although these two functions take into account some dependence through statistics T_i , $1 \leq i \leq K$ and U_i , $1 \leq i \leq K$, they do not use the information of functional interactions among those markers. Therefore, like most of the marker-wise methods, only those markers with strong marginal effects may have a chance to be included in the selected set. Important markers with weaker marginal effects are likely to be missed.

As more and more researchers are realizing, haplotype transmissions contain valuable information not included in individual markers (eg. [21, 22, 23]). It can be demonstrated that methods based on haplotype transmission disequilibrium not only continue taking advantage of linkage/disequilibrium between the markers and possible disease genes closeby, but are also capable of dealing with epistatic disease genes. In view of this, we introduce in the present paper an efficient large-scale haplotype-based marker screening algorithm for case-parent trio data. Our algorithm identifies a group of important markers that show signs of linkage/disequilibrium to the disease genes, in which complex interactions exist among themselves. This algorithm can be applied to arbitrary marker loci regardless of where they are. It takes haplotype transmission information into account during each iteration and operates in a backward fashion. Therefore, we have termed this procedure the “backward haplotype transmission association (BHTA) algorithm.” The BHTA algorithm can be described as a “marker reduction” process, since in each iteration the algorithm intends to eliminate the least important marker one at a time. The algorithm uses primarily two statistics based on haplotype transmissions. The first statistic, HTD, is calculated based on aggregated counts of haplotype transmissions that measure the disease information contained in the current marker set. A second statistical score, HTA, is used to evaluate and determine the importance of each marker that remains in the current marker set under study. The algorithm

moves on to the next cycle after the marker of least significance has been deleted and stops when all the markers remaining are important.

Based on the BHTA algorithm, we suggest in this paper a selection procedure that consists of two main steps: (1) run BHTA on a random subset of original markers, with each run resulting in a smaller set of markers, and (2) repeat step 1 B times (B is fairly large, typically $B \geq 5000$), which will produce B sets of selected markers. The final selected set comprises those markers that occur at significantly higher frequencies compared with remaining markers in step 2. We evaluate the performance of this BHTA-based marker selection procedure using simulated data under a number of complex disease models. The results show that the proposed method manages to detect the interacting gene loci when marker-wise methods fail. We also illustrate with an application of our BHTA marker selection procedure on a study of *Gilles de la Tourette syndrome*.

Primary statistics used in BHTA

Association studies compare frequencies of genetic variants between patients and controls—either unrelated or family based—for assessing evidence of linkage disequilibrium due to true linkage. In this paper, we will present the BHTA algorithm under the case-parent trio design, that is, untransmitted genetic information will be gathered to serve as internal controls. It is possible to extend the method to other family-based and classic case-control data. A brief discussion of this issue is included in the final section of this paper.

For each of a sample of diseased children, there are two parent-child haplotype transmission pairs involved. One pair is between the patient and his or her mother and the other is determined by the father. Each parent-child pair furnishes two haplotypes, one (h_t) transmitted to the child at birth, and the other (h_u) not transmitted. These four gametic haplotypes, two from the mother and two from the father, can be inferred using both the parents' and child's genotypes. There are possible transmission ambiguities, when both par-

ents and the child are all heterozygous for the same alleles at the same locus. Although one can impute the missing indexes using more sophisticated methods, and possibly more efficiently by taking into account the information from nearby markers, we have chosen and implemented here, a simple solution to resolve this problem: we randomly assign either of the two possible transmission phases by flipping a coin. By doing so, the marginal results at each locus are not changed, while transmission information concerning the functional interaction might be slightly affected [8]. Further discussion on this issue can be found in the discussion section. For the sake of easy illustration of key issues, we assume in this section that there are no ambiguities.

Haplotype transmission disequilibrium

The transmission probability differences between the two haplotypes of a parent of an affected child can be rewritten into a sum of products as follows:

$$\begin{aligned}
 & P \left(\begin{array}{c} h_1 \text{ trans.} \\ h_2 \text{ untrans.} \end{array} \middle| \text{affected child} \right) - P \left(\begin{array}{c} h_2 \text{ trans.} \\ h_1 \text{ untrans.} \end{array} \middle| \text{affected child} \right) \\
 = & \sum_{\substack{\text{transmitted} \\ \text{disease genes}}} \left[P \left(\begin{array}{c} h_1 \text{ trans.} \\ h_2 \text{ untrans.} \end{array} \middle| \begin{array}{c} \text{transmitted} \\ \text{disease genes} \end{array} \right) - P \left(\begin{array}{c} h_2 \text{ trans.} \\ h_1 \text{ untrans.} \end{array} \middle| \begin{array}{c} \text{transmitted} \\ \text{disease genes} \end{array} \right) \right] \\
 & \times P \left(\begin{array}{c} \text{transmitted} \\ \text{disease genes} \end{array} \middle| \text{affected child} \right)
 \end{aligned} \tag{1}$$

The first factor in the product is a difference that contains all relevant information of the linkage/association between the markers and the disease genes. The second factor is the conditional probability to transmit disease genes given the disease status of the child, which can take any form of interactions among the susceptibility loci. If the markers have no linkage/association to the disease, the transmission of the haplotypes should especially be independent of the disease status of the child. Therefore, the sum of products in (1) equals zero when transmission equilibrium is achieved. Thus, association studies that compare

haplotype transmissions should be able to capture the complex nature of non-Mendelian diseases.

For simplicity, suppose that m markers are being studied, each with two alleles only. This generates $H = 2^m$ haplotypes, denoted as $\{h_1, h_2, \dots, h_H\}$. Given n diseased children in the data set, there are $2n$ parent-to-patient transmission pairs: two haplotypes are observed for each pair—one transmitted to the patient and one untransmitted, denoted by $h_t^{(l)}$ and $h_u^{(l)}$ respectively for the l^{th} pair. Let the aggregated transmission counts of haplotypes be

$$\begin{aligned} n_i^t &= \#(h_t^{(l)} = h_i) \\ n_i^u &= \#(h_u^{(l)} = h_i), \end{aligned} \tag{2}$$

where “ $\#$ ” stands for “count.” To measure the disease information contained in the markers, we propose to use the following haplotype transmission disequilibrium (HTD) score,

$$\begin{aligned} \text{HTD} &= \sum_{i=1}^H (n_i^t - n_i^u)^2 \\ &= \sum_{h_i \text{ observed}} (n_i^t - n_i^u)^2. \end{aligned} \tag{3}$$

It can be shown that, under the null hypothesis of no linkage or in the absence of LD, HTD has an expectation equal to the trace of the Fisher’s information matrix, parametrized by haplotype relative risks (see appendix A). This explains why we have chosen HTD to serve as a measure of disease information when current markers are evaluated. Comparing the form of HTD and the extended TDT score proposed in [6], where haplotypes are treated as “alleles”, the major difference between the two is that the HTD takes a simpler form, without the standardization that is required for the extended TDT.

The idea of marker selection is to pick out markers that contribute the least information (regarding the trait) in a current data set, one at a time. HTD as an information measure provides a way to achieve this. Let S_M denote the current set of k markers, $S_M = \{M_1, M_2, \dots, M_k\}$. To evaluate the importance of M_r , $1 \leq r \leq k$, consider $S_M^r = S_M \setminus M_r = \{M_1, M_2, \dots, \overset{\vee}{M}_r, \dots, M_k\}$, the r^{th} -deleted marker set. Let $\mathfrak{H}_r = \{h_1, h_2, \dots, h_H\}$ be

the set of haplotypes corresponding to S_M^r , that is, the haplotypes formed by $k - 1$ markers, with M_r excluded.

Suppose that the two alleles of the r^{th} marker are a_r and b_r . Denote the numbers of transmissions of the enlarged haplotypes $^{h_i}|_{a_r}$ and $^{h_i}|_{b_r}$ by $n_i^t(a_r)$ and $n_i^t(b_r)$, respectively. Define $n_i^u(a_r)$ and $n_i^u(b_r)$ similarly for non-transmissions of the enlarged parental haplotypes to the offspring.

It is easy to see that the transmission counts after and before the deletion of M_r must satisfy

$$\begin{aligned} n_i^t &= n_i^t(a_r) + n_i^t(b_r) \\ n_i^u &= n_i^u(a_r) + n_i^u(b_r). \end{aligned} \quad (4)$$

The information contained in S_M (before deletion) can then be measured naturally by

$$\text{HTD}(k) = \sum_{h_i \in \mathfrak{H}_r} (n_i^t(a_r) - n_i^u(a_r))^2 + (n_i^t(b_r) - n_i^u(b_r))^2. \quad (5)$$

After deletion, the information left in the data set based on the r^{th} -deleted marker set S_M^r with corresponding haplotypes transmissions can be written as,

$$\begin{aligned} \text{HTD}^r(k) &= \sum_{h_i \in \mathfrak{H}_r} (n_i^t - n_i^u)^2 \\ &= \sum_{h_i \in \mathfrak{H}_r} (n_i^t(a_r) + n_i^t(b_r) - n_i^u(a_r) - n_i^u(b_r))^2 \\ &= \sum_{h_i \in \mathfrak{H}_r} (n_i^t(a_r) - n_i^u(a_r))^2 + (n_i^t(b_r) - n_i^u(b_r))^2 \\ &\quad + 2(n_i^t(a_r) - n_i^u(a_r))(n_i^t(b_r) - n_i^u(b_r)) \\ &= \text{HTD}(k) + 2 \sum_{h_i \in \mathfrak{H}_r} (n_i^t(a_r) - n_i^u(a_r))(n_i^t(b_r) - n_i^u(b_r)). \end{aligned} \quad (6)$$

From this equation, one finds that the amount of information lost by deleting marker M_r

can be expressed as the difference between (5) and (6), that is,

$$\Delta\text{HTD}^r(k) = 2 \sum_{h_i \in \mathfrak{H}_r} (n_i^t(a_r) - n_i^u(a_r))(n_i^t(b_r) - n_i^u(b_r)). \quad (7)$$

If HTD is reduced substantially (more than $2 \sum_{h_i \in \mathfrak{H}_r} n(\frac{h_i}{a_r} \parallel \frac{h_i}{b_r})$ —the number of parents with genotype $\frac{h_i}{a_r} \parallel \frac{h_i}{b_r}$, which is the amount of HTD drop due to the expected dimensional loss under the null hypothesis) as we delete the marker M_r , the marker is regarded as important, otherwise what we are removing is mostly noise.

Haplotype transmission association

To track the changes of the HTD score due to the deletion of the marker M_r , a slightly modified statistic, the haplotype transmission association (HTA) is defined:

$$\text{HTA}^r(k) = \sum_{h_i \in \mathfrak{H}_r} (n_i^t(a_r) - n_i^u(a_r))(n_i^t(b_r) - n_i^u(b_r)) + \sum_{h_i \in \mathfrak{H}_r} n(\frac{h_i}{a_r} \parallel \frac{h_i}{b_r}), \quad (8)$$

which is half of $\Delta\text{HTD}^r(k)$ plus an adjusting term $\sum_{h_i \in \mathfrak{H}_r} n(\frac{h_i}{a_r} \parallel \frac{h_i}{b_r})$ whose magnitude is negligible. The reason for this adjustment is that the modified score $\text{HTA}^r(k)$ will have an expectation 0 under the null hypothesis of no linkage/LD¹.

$\text{HTA}^r(k)$ directly measures the haplotype information loss due to the deletion of r^{th} marker. Therefore, the larger the negative value of $\text{HTA}^r(k)$, the more important the marker M_r is. The positive value of the HTA score indicates that the deleted marker is less important. The rationale behind the algorithm is supported by the general properties of HTA statistics. They are presented under three scenarios. Mathematical justifications of these properties can be found in Appendix B.

Properties of $\text{HTA}^r(k)$ under three scenarios:

¹In fact, the adjusting term $\sum_{h_i \in \mathfrak{H}_r} n(\frac{h_i}{a_r} \parallel \frac{h_i}{b_r})$ carries no information for linkage/LD and represents an fixed amount of value loss caused by the deletion of M_r . A detailed explanation and justification can be found in Appendix B.

1. Suppose that M_r is in transmission disequilibrium with the disease, while the remaining markers are independent of the disease loci.

Then, one of the two alleles associated with M_r is more likely to be transmitted to the affected child than the other allele. The transmission count differences of allele a_r , $(n_i^t(a_r) - n_i^u(a_r))$, will have a negative correlation with the corresponding differences of the allele b_r , $(n_i^t(b_r) - n_i^u(b_r))$, for $i = 1, 2, \dots, H$. As a result, the observed association, $HTA^r(k)$, will have a negative mean, reflecting the importance of M_r .

2. Suppose that M_r has no linkage/LD to the disease.

Then, the transmission of the allele (a_r or b_r) is independent of the trait. In this case, the differences of the transmitted and untransmitted haplotypes are determined by the $h_i \in \mathfrak{H}_r$ defined by the markers in the r^{th} -deleted marker set S_M^r . The expected value of $HTA^r(k)$ will be always nonnegative and will be strictly positive if some markers in S_M^r are in linkage disequilibrium with the disease.

3. Suppose that the marker M_r is linked not only to the disease but also shows interaction with other markers in the r^{th} -deleted marker set S_M^r .

Then, if the marker M_r contains important linkage information or plays a critical role in the interactive set of markers, the expected scores will be negative and the marker will be retained. On the other hand, if M_r brings comparable lower risk to the interaction among the markers, the corresponding HTA score will result in a positive value which may lead to its deletion from the current marker set.

Implementing BHTA and Practical Issues

In this section, BHTA marker selection algorithm is developed based on the properties of HTA statistic. Demonstrative figures are included and they are obtained by using simulations. For detailed information about the simulations and disease models, the readers are referred to example 1 (next section).

As noted in the previous section, negative HTA scores indicate the importance of the marker. The algorithm intends to delete the least important marker one at a time, where the HTA score is positive, and at its maximum. The role of the HTA score in BHTA is described in Figure 1.

Figure 1 goes here — BHTA flow chart

From Figure 1, the BHTA screens out markers with little information regarding linkage/LD, and stops when all remaining markers show evidence of transmission disequilibrium.

To best understand the relation between the deletion and information changes, consider an example as follows: Figure 2 shows a typical application to a simulated data containing 200 patients. 30 diallelic markers are included in the study. The simulation is constructed with 6 markers (indicated by filled symbols of different shapes) in linkage disequilibrium with three interacting disease genes.

Figure 2 goes here — information flow plot.

Initially, before any marker is deleted, the disease information measured by HTD is low. This is because the amount of information has been swamped by the noises and large dimensions due to these markers. As the BHTA screens out more and more markers (after 15 markers were deleted), the information begins to grow. One notes that before the HTD rises, the markers are deleted randomly. The algorithm will stop at the peak of HTD score, which in this case is 5420. If the deletion process is forced to continue, the information would start to drop sharply. For comparison, we also include in Figure 2 an information flow plot for a set of markers without linkage/disequilibrium to the disease. It is easy to see that there is no important information contained in those markers since the HTD curve stays low throughout all deletions and eventually, BHTA returns no markers.

In practice, the number of markers and their possible interactions included in a genome scan is often larger than the number of observations. This moderate size of observations will cause a serious problem of sparseness in the haplotype data when dealing with many

markers simultaneously. This phenomenon is already observed in Figure 2 even when the number of markers is only 30. The fact is when the number of possible haplotypes is much larger than the number of observations, markers are deleted randomly. This is because the counts of haplotype transmissions are mostly 0's or 1's, hence the HTA scores based on these counts all cluster around zero, which makes the selection non-informative. For this reason, we suggest in the following a two-step marker selection procedure to fix this problem. In what follows, this two-step procedure is applied to all illustrated examples throughout the remaining paper. It should be pointed out that this two-step procedure is similar to the nested bootstrap procedure proposed in [19], when dealing with bootstrapping.

Two-step BHTA marker selection procedure:

Step 1 Randomly select k (30, for instance) markers out of the original set of K markers. Run BHTA on the selected markers and record the markers returned.

Step 2 Repeat step 1 B times (B typically ≥ 5000 ; (see Appendix C for a guideline on selecting values for B)). Markers that are returned more frequently than others will be selected in the resulting set. The criterion used in the present study is based on the distribution of the returning frequencies of all markers. Markers whose returning frequencies are more than the 3rd quartile plus 1.8 times IQR (inter-quartile range) will be selected in the resulting set. This criterion is equivalent to 3.1 standard deviation from the mean, which is corresponding to 10^{-3} type I error rate.

To understand the two-step BHTA selection procedure, it helps to examine the probability of deletion for each marker at each round (iteration) of one BHTA screening. At the i^{th} round, the probability to be deleted for marker k , $d(k, i)$, can be estimated by

$$\hat{d}(k, i) = \frac{\#(M_k \text{ is deleted in } i^{th} \text{ round})}{\#(M_k \text{ is not deleted before } i^{th} \text{ round})},$$

where the numerator is the number of times, out of B times, that marker M_k is deleted in

the i^{th} round. The denominator is defined analogously.

If we assume random deletion (no linkage/LD information), the probability of deletion $d_0(k, i) = 1/m$, while $m = K - i + 1$ (K is the number of markers in the original set). The relative probabilities of deletion, $\hat{d}(k, i)/d_0(k, i)$, are calculated for all markers at each round and the corresponding average values for associated and unassociated markers are plotted separately in Figure 3 for comparison.

Figure 3 goes here — average relative probability of deletion (alternative vs. null)

In the left plot of Figure 3, the relative probability of deletion for 30 markers is obtained using $B = 5,000$ independent repetitions of step 1. In the first half of the screening, 15 markers are picked out randomly, thus the probability of deletion is similar between the associated and unassociated markers. After the 15th deletion, the BHTA algorithm takes control for the 15 markers remaining. A sharp drop in the probability to be deleted is observed for the associated markers, while for the unassociated markers the probability of deletion becomes higher. In other words, if a marker is in linkage/disequilibrium with the disease susceptibility loci, as long as it survives the first 15 deletions, it will have a greater probability to be selected by the BHTA and retained in the final group of important markers. On the other hand, a marker without linkage/disequilibrium to the trait will eventually be removed from the final marker set even if it is included by chance in the second group of 15 markers. The right panel of Figure 3 displays markers' average relative probability of deletion when none of them contain information regarding the disease. No drops or increases appear in this plot. Thus, we can conclude that the differences in the probability of deletion between associated markers and unassociated markers in the left plot are due to the differences in disease information contributed by them. By repeating the screening a large number of times, these differences can easily be observed.

Returning frequencies of 30 markers are plotted in Figure 4. Selection results are based on the IQR-based criterion proposed above. The two-stage BHTA marker selection procedure manages to pick out the important markers covering all three disease loci, while sifting out

the unassociated markers.

Figure 4 goes here marker selection results from one simulation.

Illustrated examples and further discussion

Complex traits, which are also described as polygenic, multifactorial, or non-Mendelian, offer a challenge to the field of genetic mapping because of their incomplete penetrance, genetic heterogeneity, polygenic inheritance and other numerous unknown complexities [12] [17]. Here we list a few of these possibilities and demonstrate the potential of our methods in dealing with complex traits by assessing its performance under various scenarios.

We present in this section four data examples. The purpose of these examples is both to demonstrate and to evaluate the proposed BHTA method and to provide practical illustrations of its use. We start with three hypothetical complex models, followed by an application to real data set on *Gilles de la Tourette syndrome* [28]. The first example is a continuation from the previous section. We give further details on the disease models and explanation of simulation details of data generations. Some computational issues and further discussions are also included.

Example 1: Epistasis, HRR model displaying no marginal effects (continued)

It is believed that most complex traits are caused by multiple genetic variants at disjoint genome regions, with possible epistatic interactions among them. This kind of polygenic inheritance is commonly found in mice and other lab animals through crossing designs, while it is harder to identify in humans [14, 16]. Culverhouse *et al* [17] argued and demonstrated that, for some extreme hypothetical epistatic models, traditional association methods failed due to the absence of marginal effect. In this example, we simulate data sets under a three-

locus disease model using haplotype relative risks (HRR) (Table 1), which balances out marginal risk at each susceptibility locus. It is not clear how realistic these types of models are, nevertheless, it is illuminating to examine the performance of BHTA under this difficult and extreme situation.

Table 1 goes here—haplotype relative risk example.

In the current example, 30 markers are simulated, among which 2 markers are in linkage/association with the disease gene at each susceptibility locus, making a total of 6 associated markers. Data sets of genetic information of 200 patients and their parents can be easily generated under this model. For each patient-parent trio, the simulation starts with the patient’s haplotypes on the disease loci, given the affected status. By using pre-specified parameters such as gene-to-marker linkage/LD (recombination fraction, linkage disequilibrium) and markers’ population frequencies, the patient’s haplotypes on the marker loci (transmitted) can then be generated based on the disease haplotypes. Untransmitted haplotypes are simulated using only population parameters of the markers and inter-marker correlations. Numerical specification of the model parameters for each simulated study can be found in the notes of the corresponding figures and tables.

As mentioned above, the model is selected so that at each disease locus no marginal effects can be observed. Thus, there is no information that can be utilized by marker-by-marker association methods. The screening results and analysis have been demonstrated in the discussion of the previous section. BHTA screening has a dominant advantage in dealing with this epistatic model via greater efficiency in the usage of information.

Example 2: Modifier gene models

In this subsection, we demonstrate the performance of BHTA methods under a less extreme disease model, and compare it to traditional association methods represented by marker-by-marker TDT.

Genetic modifiers are genetic variants that alter the phenotype of a target gene, which are quite common in humans. Modification can cause enhanced, reduced, novel disease phenotypes, or normal phenotypes [16]. Usually, modifier genes only contribute mildly in deciding the traits of interest. As a result, there are only a few cases for which modifier genes have been identified [16]. An example is given in the results from a recent linkage analysis on Bardet-Biedl syndrome (BBS) [24]. In the study, the authors found that the BBS patients needed three mutant alleles on the disease’s six susceptibility loci in order to be affected (two homozygous copies at one locus, another copy at any of the remaining five loci) [25]. The results they obtained from pedigree data were consistent with their triallelic hypothesis. It might be termed as “recessive inheritance with a modifier of penetrance” [25]. In this case, each disease gene indeed contributes mild effect to the penetrance of the disease. Thus, conventional methods measuring the marginal effects may fail in detecting these important genes.

Motivated by the BBS’s triallelic inheritance pattern, we assume a disease model with 3 responsible genes, where each patient needs two copies of the mutated allele at one locus and another mutant at any of the other two loci in order to be affected. 100 diallelic markers are simulated, with 3 markers at each disease locus in linkage/disequilibrium with the disease genes (a total of 9 markers are associated with the disease). Figure 5 displays the performance of a marker-by-marker scan using individual TDTs and selection results from BHTA methods.

Figure 5 goes here — modifier genes model performance.

From the results above, it is obvious that the substantial advantages of BHTA method lie not only in increased power at each individual disease locus, but also in the higher probability of joint detection of all three susceptibility loci (in marker-wise methods it is very low due to the mild marginal effects of each gene). It is concluded from this example that the BHTA procedure is much more powerful than marker-by-marker scan methods in terms of overall detection and selection.

Example 3: Polygenic admixture models

The purpose of this example is to compare the performance of the proposed BHTA method with standard methods when the marginal effects are substantial.

Patients of a complex trait may present similar disease phenotypes that result from mutations at different genetic loci, which are not necessary to have functional interactions. Thus, data collection based on the patients' disease status will lead to a mixture of possible disease genotypes, which may affect the marginal information at each locus. Here we consider a simple example of such polygenic admixture. Assume a disease with 3 susceptibility loci, where each patient needs only one locus that is homozygous of the mutated allele in order to have a high penetrance of the disease. The 3 disease loci are simulated to be physically independent on the genome. 30 diallelic markers are simulated, with 3 markers are associated with the 3 disease genes respectively. This is a model with merely strong marginal effect since the 3 disease genes are assumed to have no functional interaction. It is interesting to exam the performance of BHTA under such a model, while comparing to traditional association methods, again represented by marker-by-marker TDTs. Figure 6 displays the selection results from 1000 simulations of this disease model, each with 150 patients.

Figure 6 goes here polygenic admixture model performance.

As shown in Figure 6, although the disease model contains only marginal effects (no interaction among the disease loci at all), our proposed method still shows a substantial advantage over the marker-wise methods.

Example 4: Application to Tourette Syndrome

Finally, we apply the BHTA marker selection to a study on *Gilles de la Tourette* syndrome recently conducted by Simonin *et al* [28].

Tourette's syndrome is a chronic neuropsychiatric disorder with onset in childhood, char-

acterized by multiple, fluctuating motor and vocal tics of variable severity [30]. In [28], five genomic region 2p, 8q, 11q, 20q and 21q were investigated using a sample of 91 Afrikaner nuclear families, as an attempt to confirm the findings from a previous case-control study. Single marker linkage and association tests were carried out on a total of 31 markers, and multiple-marker methods were used for 2 or 3 adjacent markers. From the results published in [28], significant results have been found in all five regions, in either the previous case-control studies or the association study carried out in [28], while by using only tests conducted marker-wise or on near-by markers, it is hard to identify possible polygenic interaction among genes from disjoint genomic regions, such as regions on different chromosomes.

We reinvestigate the data of [28] using the methods we propose in the present paper. We found eight markers to be important: D2S440, D8S270, D8S559, GATA8B01, D11S1377, D11S4464, D20S1085, and D20S469. Further analysis indicates that these eight markers can be divided into three groups, with each groups showing strong evidence of linkage/LD, with respective p-values 0.003, 0.0017, and 0.004. The p-values are obtained by the extended TDT method proposed in [6]. The results are listed in Table 2.

Table 2 goes here—TS results summary.

Our findings confirm the results from the previous papers [28, 29, 30]: strong linkage/association evidence is found pointing to the genomic regions of 2p11, 8q22, and 11q23-24. And yet the evidence of linkage/LD based on our method is significantly stronger than previously found. A reason for that is the haplotype-based methods are able to extract and to use more information from the same data. In addition to this, our results strongly suggest epistatic interactions between 8q and 20q (group 1 and 2 in Table 2). These findings are particularly interesting because the regions on chromosome 20 have been suspected in [28], but their attempt to reconfirm the previous evidence for association with GTS was unsuccessful.

Reduced Computational Complexity

Test statistics using all possible genotypes (haplotypes) are not computable in a study using 30 markers or more. Even if we conservatively assume that each of the markers only have 2 alleles, we will be dealing with 2^{30} possible haplotypes and 2^{60} possible genotypes, which is beyond the capacity of most personal computers. The BHTA method only uses observed haplotypes to construct the test statistics, which reduces the calculation to the order of the sample size. Without the reconstruction of high-dimensional matrix or statistic, the calculation of the HTA score in BHTA is simple and time-efficient. Running a screening for 30 markers using 200 patients took no more than 2 seconds. The two-step marker selection procedure takes about 10 minutes for 30 markers, and 24 hours for 500 markers. It is believed that by using better computational design and more developed computing facilities, screening time will be further reduced.

Algorithm Summary

Much of the current efforts for detecting genetic factors of complex traits in human diseases involve large numbers of markers. We have proposed in this article a backward haplotype transmission association algorithm method to identify important markers. The algorithm is designed in such a way that a smaller set of interesting and important markers that are in linkage/transmission disequilibrium with the disease traits can be identified from a very large set of markers.

Unlike the existing methods in genome scans, the proposed BHTA has the capacity to detect those markers with interactions (covariances) among themselves, even though the individual markers show no sign of transmission disequilibrium at all. Also, this algorithm can handle arbitrary marker loci across different chromosomes. Both these aspects are important and cannot be captured by existing methods in the literature.

In this paper, for simplicity, we present the BHTA algorithm based on diallelic markers

only (suitable for SNPs), but it can be generalized quite straightforwardly to markers with more than two alleles. For example, if the r^{th} -deleted marker consists of three alleles, an a_r , b_r , and c_r , one needs only replace the cross product term appended in (8) by the sum of three cross product terms involving (a_r, b_r) , (a_r, c_r) , and (b_r, c_r) , and then proceed with the same algorithm as described in the article. As one referee pointed out that according to [31], the likelihood ratio test proposed in [31] becomes more conservative as the number of alleles at a marker increases. Therefore, it is natural to pose the following questions: 1. to what extent the effect of multi-allelic markers is on the proposed BHTA analysis? 2. How to set a new IQR-based criterion to assess markers' importance? The authors do not have answers presently but feel that those are important issues certainly deserve further investigation.

In the earlier section where we introduced the primary statistics, a random phase assignment was proposed and implemented in this article. To further understand the impact of this strategy, we have used the data of example 3 to conduct a comparison study to assess the differences between the results obtained from the random phase assignments versus from the known phases. Table 3 displays the comparison result.

Table 3 goes here—Impacts of random phase assignments

From this example, we speculate that, in general, the impacts due to the random phase assignments are small, but a more detailed study is needed to confirm this speculation.

Although we have presented our findings based on small to moderate marker sizes, from 30 to 100, we have run (not shown here) both examples 1 and 2 based on a larger number, $K = 500$, of markers, and the results are very similar to what is presented here. The only difference is, instead of using $B = 5000$ repetitions in step 2, one needs to increase B to 20000 or so in order to see a clear separation. The reason for this is obvious: for each marker, the probability of being included in the k markers (step 1) out of 500 markers is smaller than out of 100 markers. Thus more runs are required in order to capture the real differences. The good news is that the number of repetitions B required maintains a linear relationship with the number of candidate markers, given the same size of data (see Appendix C). We

anticipate that, with 5000 markers, one needs to use roughly $B = 200000$ runs in step 2 to observe a real difference, which is not out of reach if one runs it on a parallel cluster network of 10 pc's (the repetitions of step 1 in step 2 are independent).

The procedure developed here can be extended to family data without parental genotypes or to the traditional case-control data. In the case of absence of parental information, however, unaffected individuals are required to serve as controls. The key idea for both types of data is to form a generalized haplotype by grouping possible genotypes. Various extensions are currently under investigation by these authors.

The proposed method not only shows strong promise in the examples presented here, it works well for many other simulated examples that we have carried out (not presented). We are currently investigating a large scale study involving thousands of markers, of which nearly a hundred are associated with the trait. When the number of disease loci is large (larger than 15, say), the proposed two-step procedure based on random subsets is still able to capture those loci as long as non-trivial interactions exist among some smaller subsets of markers. Theoretically, one can construct hypothetical examples where no subset with fewer than 15 markers contains any disease information (i.e., neither marginals nor interactions), which means that the haplotype risks of the disease concerning marker subsets smaller than 15 are completely balanced out. In this case, our proposed procedure may fail to detect them. However, we believe that it is highly unlikely for a situation like this to occur in reality.

In view of the many difficulties in genome scans for complex disease reported in recent literature [10, 13, 17], the proposed method and its extensions may offer a realistic way to tackle the challenging problems that human geneticists are facing today.

Acknowledgement

The authors thank Jürg Ott for providing the GTS data set, his papers [19, 20] and invaluable suggestions that greatly improve the presentation of this work. We thank Andrew Gelman for useful comments on an earlier draft. We also appreciate the invaluable comments from two anonymous reviewers, which are both insightful and constructive. The research is partially supported by NSF Grant DMS-00-71930.

Appendix

A Justification of the HTD statistic

Given n diseased children in the data set, there are $2n$ parent-child transmission pairs. Assume k markers $S_M = \{M_1, M_2, \dots, M_k\}$ are in the data set. Let i index possible haplotypes, that is, $h_i \in \mathfrak{H}_r = \{h_1, h_2, \dots, h_H\}$, $H = 2^k$. Define,

$$\begin{aligned} p_i &: \text{population frequency of haplotype } h_i, \\ r_i &: \text{relative risk of disease given haplotype } h_i, \end{aligned}$$

and they satisfy

$$\begin{cases} \sum_i p_i = 1, \\ \sum_i r_i p_i = 1. \end{cases} \quad (9)$$

Conditional probability of parental haplotype transmissions given the disease status of the child can be expressed by a matrix:

$$TP = [P_{trans}(i, j)]_{H \times H},$$

where each element $P_{trans}(i, j)$ in the matrix is defined as the probability to observe one parent with h_i as the transmitted haplotype and h_j as the untransmitted haplotype. Simple calculations yield,

$$\begin{aligned} P_{trans}(i, j) &= \text{Prob}(h_i \text{ trans.}, h_j \text{ untrans.} \mid \text{affected child}) \\ &= \frac{p_d r_i p_i p_j}{\text{Prob}(\text{parent of patient})}, \end{aligned} \quad (10)$$

where p_d is the population prevalence of the disease. It is easy to show that,

$$\begin{aligned} P_{pop} &\stackrel{\text{def}}{=} \text{Prob}(\text{parent of patient}) \\ &= p_d \sum_{i,j} r_i p_i p_j = p_d \end{aligned} \quad (11)$$

Under the null hypothesis of no linkage/association to the trait, that is the haplotypes carry no information regarding the disease's susceptibility:

$$H_0 : r_1 = r_2 = \dots = r_H = 1,$$

it is easy to see that matrix TP is symmetric, and *haplotype transmission equilibrium* is achieved. Deviation from this symmetry—*haplotype transmission disequilibrium*—yields under the alternative hypothesis that not all haplotypic relative risks are equal to 1.

Following the parameterization, the likelihood function can be written as

$$P(x; \tilde{p}, \tilde{r}) = \binom{2n}{n_{11}, n_{12}, \dots, n_{H-1H}, n_{HH}} \prod_{i,j} P_{trans}(i, j)^{n_{ij}}, \quad (12)$$

where

$$\begin{cases} x = \{n_{ij}, i = 1, 2, \dots, H, j = 1, 2, \dots, H\}, \\ \tilde{p} = \{p_1, p_2, \dots, p_H\}, \tilde{r} = \{r_1, r_2, \dots, r_H\}, \end{cases}$$

and n_{ij} is the number of transmission pairs with h_i transmitted and h_j untransmitted.

Then the log likelihood is

$$\log(P(x; \tilde{p}, \tilde{r})) = C(x) + \sum_{ij} n_{ij} \log(r_i p_i p_j) - 2n \log(P_{pop}). \quad (13)$$

We therefore have,

$$\frac{\partial \log(P(x; \tilde{p}, \tilde{r}))}{\partial r_i} = \frac{n_i^t}{r_i} - 2n \frac{p_i p_d}{P_{pop}}, \quad i = 1, 2, \dots, H, \quad (14)$$

satisfying $\sum_i r_i p_i = 1$, which leads to $P_{pop} = p_d$. Thus, (14) can be simplified as

$$\frac{\partial \log(P(x; \tilde{p}, \tilde{r}))}{\partial r_i} = \frac{n_i^t}{r_i} - 2n p_i, \quad i = 1, 2, \dots, H. \quad (15)$$

Similarly we have

$$\frac{\partial \log(P(x; \tilde{p}, \tilde{r}))}{\partial p_i} = \frac{n_i}{p_i} - 2n(r_i + 1), \quad i = 1, 2, \dots, H. \quad (16)$$

As defined in the previous section, recall that n_i^t is the number of transmission pairs that h_i is the transmitted haplotype and n_i^u is the number of times that h_i is untransmitted. Denote $n_i = n_i^t + n_i^u$ as the number of h_i observed in the data set, transmitted or not.

The values of p_i 's are estimated by the maximum likelihood estimators as follows,

$$m.l.e(p_i) = \frac{n_i^u}{2n}, \quad i = 1, \dots, H$$

(by solving

$$\left\{ \begin{array}{l} \frac{n_i^t}{r_i} + \lambda p_i = 0, \quad i = 1, 2, \dots, H, \\ \frac{n_i}{p_i} + \lambda r_i + \eta = 0, \quad i = 1, 2, \dots, H, \\ \sum_i p_i = 1, \quad \sum_i r_i p_i = 1 \end{array} \right. \quad (17)$$

Thus, the Fisher's information matrix for the relative risks under the null hypothesis can be found:

$$\begin{aligned} I^{(0)}(\tilde{r}_0) &= \left[E_0 \left(\frac{\partial \log(P(x; m.l.e(\tilde{p}), \tilde{r}_0))}{\partial r_i} \frac{\partial \log(P(x; m.l.e(\tilde{p}), \tilde{r}_0))}{\partial r_j} \right) \right]_{H \times H} \\ &= \begin{bmatrix} E_0(n_1^t - n_1^u)^2 & \dots & \dots & \dots \\ \dots & E_0(n_2^t - n_2^u)^2 & \dots & \dots \\ \dots & \dots & \ddots & \dots \\ \dots & \dots & \dots & E_0(n_H^t - n_H^u)^2 \end{bmatrix} \end{aligned} \quad (18)$$

It follows that the expectation of proposed score HTD (3) is indeed equal to the trace of the information matrix above.

B Properties of the HTA statistic

Assuming that we are testing marker M_r in $S_M = \{M_1, M_2, \dots, M_k\}$, $r \leq k$, let i index possible haplotypes expanded by $k - 1$ non-deleted markers S_M^r . $k - 1$ markers in total can define $2^{(k-1)}$ haplotypes $\mathfrak{H}_r = \{h_1, h_2, \dots, h_{2^{(k-1)}}\}$. Since the data takes the form of parent-child pairs, we index individual pairs by l , and use $n_p = 2n$ to denote the total number of involved pairs. Let $\delta_l^i(a_r)$ be the indicator of transmission of $\frac{h_i}{a_r}$ from the parent to the diseased child in the l^{th} pair; and $\eta_l^i(a_r)$ the indicator of non-transmission of $\frac{h_i}{a_r}$ from the parent, while a_r is one of the two alleles corresponding to the marker (M_r) being studied for the effects of its deletion. Denote the other allele of this diallelic marker by b_r , with $\delta_l^i(b_r)$ and $\eta_l^i(b_r)$ being the corresponding indicators similarly defined.

Using these indicators, the statistics used in the test can be written as:

$$\begin{aligned}
 n_i^t(a_r) &= \sum_{l=1}^{n_p} \delta_l^i(a_r) \\
 n_i^u(a_r) &= \sum_{l=1}^{n_p} \eta_l^i(a_r) \\
 n_i^t(b_r) &= \sum_{l=1}^{n_p} \delta_l^i(b_r) \\
 n_i^u(b_r) &= \sum_{l=1}^{n_p} \eta_l^i(b_r).
 \end{aligned} \tag{19}$$

Likewise, the difference between haplotype transmission disequilibrium scores can be rewritten as:

$$\begin{aligned}
 \Delta \text{HTD}^r(k) &= 2 \sum_{h_i \in \mathfrak{H}_r} \left[\sum_{l=1}^{n_p} (\delta_l^i(a_r) - \eta_l^i(a_r)) \sum_{l=1}^{n_p} (\delta_l^i(b_r) - \eta_l^i(b_r)) \right] \\
 &= 2 \sum_{i=1}^{2^{(k-1)}} \left[\sum_{l=1}^{n_p} (\delta_l^i(a_r) - \eta_l^i(a_r)) \sum_{l=1}^{n_p} (\delta_l^i(b_r) - \eta_l^i(b_r)) \right]
 \end{aligned} \tag{20}$$

The second equation in (20) follows from the fact that the products

$$\left[\sum_{l=1}^{n_p} (\delta_l^i(a_r) - \eta_l^i(a_r)) \sum_{l=1}^{n_p} (\delta_l^i(b_r) - \eta_l^i(b_r)) \right]$$

are equal to zero for any haplotype h_i not observed in the family data.

To simplify the formulation, without losing much generality, we assume first that the disease model is multiplicative. Second, each family contains one diseased child in the test². Following the assumptions, the sum in (20) may be broken into two parts. The first part is a sum of independent products, conditioning on the known parental genotypes. The assumption of independence is based on two perspectives of the data. First, the transmissions from different families are independent. Second, while conditioning on the disease status of the child, the transmission from one parent is independent from the other under the multiplicative model. The second part is the sum of cross-products within a pair. That is,

$$\begin{aligned} \frac{1}{2} \Delta \text{HTD}^r(k) &= \sum_{i=1}^{2^{(k-1)}} \sum_{l=1}^{n_p} \sum_{j \neq l}^{n_p} (\delta_l^i(a_r) - \eta_l^i(a_r)) (\delta_j^i(b_r) - \eta_j^i(b_r)) \\ &\quad + \sum_{i=1}^{2^{(k-1)}} \sum_{l=1}^{n_p} (\delta_l^i(a_r) - \eta_l^i(a_r)) (\delta_l^i(b_r) - \eta_l^i(b_r)) \\ &= \sum_{i=1}^{2^{(k-1)}} \sum_{l=1}^{n_p} \sum_{j \neq l}^{n_p} (\delta_l^i(a_r) - \eta_l^i(a_r)) (\delta_j^i(b_r) - \eta_j^i(b_r)) \end{aligned} \quad (21)$$

$$- \sum_{i=1}^{2^{(k-1)}} n \left(\begin{matrix} h_i \\ a_r \parallel b_r \end{matrix} \right) \quad (22)$$

The product $(\delta_l^i(a_r) - \eta_l^i(a_r))(\delta_l^i(b_r) - \eta_l^i(b_r))$ is non-zero if and only if the parent has genotype $\left(\begin{matrix} h_i \\ a_r \parallel b_r \end{matrix} \right)$. In this case, only two possible scenarios can occur, either a_r transmitted or b_r transmitted. Using simple algebra, we can see under either scenario, the product always

²Without the assumptions of multiplicative risks and one child per family, the dependence between two parents' transmission to one diseased child and the dependence of the transmissions from one parents to several siblings may introduce bias to the HTA statistic. The bias is on the order of n_p (the number of parent-patient transmission pairs), which is ignorable comparing to the n_p^2 order of the HTA scores. Therefore, the results discussed here still hold true for general disease models and family data.

equals -1 . Thus, the second part is simply the sum of total negative counts for genotype $(\frac{h_i}{a_r} \parallel \frac{h_i}{b_r})$. This term contains neither transmission nor linkage information. For simplicity, we adjust $\frac{1}{2} \Delta \text{HTD}^r(k)$ by adding the additional term $\sum_{i=1}^{2^{(k-1)}} n(\frac{h_i}{a_r} \parallel \frac{h_i}{b_r})$ as the statistic score $\text{HTA}^r(k)$ used in the BHTA algorithm. In the following inference, we will focus on this Haplotype Transmission Association (HTA) score,

$$\text{HTA}^r(k) = \sum_{i=1}^{2^{(k-1)}} \sum_{l=1}^{n_p} \sum_{j \neq l} (\delta_l^i(a_r) - \eta_l^i(a_r))(\delta_j^i(b_r) - \eta_j^i(b_r)). \quad (23)$$

Recall the fact that the terms in the products are independent, so we can study the expected value of the statistic $\text{HTA}^r(k)$ by summing up the products of expected differences,

$$\begin{aligned} & E(\text{HTA}^r(k) \mid \text{parental genotypes}, D) \\ &= \sum_{i=1}^{2^{(k-1)}} \sum_{\frac{h_i}{a_r} \text{ in pair } l} \sum_{\frac{h_i}{b_r} \text{ in pair } j, j \neq l} \Delta P_i(a_r) \Delta P_i(b_r) \\ &= \sum_{i=1}^{2^{(k-1)}} n_i(a_r) n_i(b_r) \Delta P_i(a_r) \Delta P_i(b_r), \end{aligned} \quad (24)$$

where $n_i(a_r)$ and $n_i(b_r)$ are counts of parent-child pairs with haplotype $\frac{h_i}{a_r}$ and $\frac{h_i}{b_r}$ respectively. Moreover,

$$\Delta P_i(a_r) = P(\frac{h_i}{a_r} \text{ trans.} \mid (\frac{h_i}{a_r}, h_?), D) - P(\frac{h_i}{a_r} \text{ untrans.} \mid (\frac{h_i}{a_r}, h_?), D) \quad (25)$$

and

$$\Delta P_i(b_r) = P(\frac{h_i}{b_r} \text{ trans.} \mid (\frac{h_i}{b_r}, h_?), D) - P(\frac{h_i}{b_r} \text{ untrans.} \mid (\frac{h_i}{b_r}, h_?), D) \quad (26)$$

are the differences of conditional transmission probabilities given the child's disease status and the parental genotype, $(\frac{h_i}{a_r}, h_?)$ or $(\frac{h_i}{b_r}, h_?)$.

In the formulas above, the term “trans.” stands for “transmission to the child”, and “untrans.” means the haplotype was not transmitted. The letter D stands for the fact that the child is known to be affected. The haplotype $h_?$ means any haplotype.

To study the values of the expectation shown in (24), a disease inheritance model needs

to be defined in order to evaluate the relative risk. Define

$$\begin{aligned} R(h_i, h_j) &= \text{Relative Risk}(D \mid (h_i, h_j) \text{ is the child's genotype}) \\ &= \frac{\text{Prob}(D \mid (h_i, h_j) \text{ is the child's genotype})}{\text{Prob}(D)}, \end{aligned} \quad (27)$$

the relative risk of a child being diseased given his or her genotype at the studied loci, which is transmitted from his or her parents at the time of birth. Let p_d denote the population prevalence of the disease in the children's generation, i.e. $\text{Prob}(D)$.

Let $R(h_i) = \frac{P(D \mid h_i \text{ trans. })}{p_d}$ be the marginal haplotype relative risk function. Simple calculations yield

$$\begin{aligned} P(D \mid h_i \text{ trans. }) &= \frac{1}{P(h_i \text{ trans. })} \sum_{h_j}^{2^{(k-1)}} P(D, (h_i, h_j) \text{ is the child's genotype}) \\ &= \frac{p_d}{P(h_i \text{ trans. })} \sum_{h_j}^{2^{(k-1)}} R(h_i, h_j) P(h_i \text{ trans. }) P(h_j \text{ trans. }) \\ &= p_d \sum_{h_j}^{2^{(k-1)}} R(h_i, h_j) P(h_j \text{ trans. }). \end{aligned} \quad (28)$$

Thus,

$$R(h_i) = \sum_{h_j}^{2^{(k-1)}} R(h_i, h_j) P(h_j \text{ trans. }), \quad (29)$$

and

$$\sum_{h_i}^{2^{(k-1)}} R(h_i) P(h_i \text{ trans. }) = 1. \quad (30)$$

Assuming that the haplotype frequencies have achieved Hardy-Weinberg equilibrium from the parental generation to the children generation, the unconditional probability that a child inherits an h_i haplotype from one of his or her parents generally equals the population probability of h_i among the parents, that is, $P(h_i \text{ trans. }) = P(h_i)$.

With the notations defined above, we can formulate the probabilities of a haplotype's transmission or nontransmission to the affected child as

$$\begin{aligned}
 P(h_i \text{ trans.} \mid (h_i, h?), D) &= \frac{P(D \mid h_i \text{ trans.}, (h_i, h?))}{P(D \mid h_i \text{ trans.}, (h_i, h?)) + P(D \mid h? \text{ trans.}, (h_i, h?))} \\
 &= \frac{R(h_i)}{R(h_i) + 1},
 \end{aligned} \tag{31}$$

$$\begin{aligned}
 P(h_i \text{ untrans.} \mid (h_i, h?), D) &= \frac{P(D \mid h? \text{ trans.}, (h_i, h?))}{P(D \mid h_i \text{ trans.}, (h_i, h?)) + P(D \mid h? \text{ trans.}, (h_i, h?))} \\
 &= \frac{1}{R(h_i) + 1},
 \end{aligned} \tag{32}$$

respectively.

Thus, the transmission probability difference is

$$\Delta P_i(a_r) = (31) - (32) = \frac{R(h_i) - 1}{R(h_i) + 1}. \tag{33}$$

In fact, the denominator is proportional to the probability of h_i presented in the genotype of a diseased child's parent:

$$\begin{aligned}
 P_c(h_i) &\stackrel{\text{def}}{=} \text{Prob}((h_i, h?) \mid \text{parent of a patient}) \\
 &= \frac{\sum_{h?} P(\text{affected child} \mid (h_i, h?)) 2P(h_i)P(h?) }{P(\text{affected child})} \\
 &= \sum_{h?} \left(\frac{1}{2} R(h_i) + \frac{1}{2} R(h?) \right) P(h_i)P(h?) \\
 &= P(h_i) \frac{(R(h_i) + 1)}{2}.
 \end{aligned}$$

We rewrite (33) into:

$$\Delta P_i(a_r) = \frac{1}{2} \frac{P(\frac{h_i}{a_r})(R(\frac{h_i}{a_r}) - 1)}{P_c(\frac{h_i}{a_r})}. \quad (34)$$

Similarly, the other probability in the product can be rewritten as

$$\Delta P_i(b_r) = \frac{1}{2} \frac{P(\frac{h_i}{b_r})(R(\frac{h_i}{b_r}) - 1)}{P_c(\frac{h_i}{b_r})}. \quad (35)$$

If we put the simplified probability differences (34) and (35) into (24), we obtain³

$$\begin{aligned} E(\text{HTA}^r(k)) &= E(E(\text{HTA}^r(k) \mid \text{parental genotype})) \\ &= E\left(\sum_{i=1}^{2^{(k-1)}} n_i(a_r)n_i(b_r)\Delta P_i(a_r)\Delta P_i(b_r)\right) \\ &= \sum_{i=1}^{2^{(k-1)}} E(n_i(a_r)n_i(b_r)) \frac{\frac{1}{4} P(\frac{h_i}{a_r})P(\frac{h_i}{b_r})(R(\frac{h_i}{a_r}) - 1)(R(\frac{h_i}{b_r}) - 1)}{P_c(\frac{h_i}{a_r})P_c(\frac{h_i}{b_r})} \end{aligned}$$

Note:

$$\begin{aligned} E(n_i(a_r)n_i(b_r)) &= E(E(n_i(a_r)n_i(b_r) \mid n_i)) \\ &= E(n_i(n_i - 1)) \frac{P_c(\frac{h_i}{a_r})P_c(\frac{h_i}{b_r})}{(P_c(\frac{h_i}{a_r}) + P_c(\frac{h_i}{b_r}))^2} \\ &= 2n_p(2n_p - 1)P_c(\frac{h_i}{a_r})P_c(\frac{h_i}{b_r}) \end{aligned}$$

$$\begin{aligned} &= 2n_p(2n_p - 1) \sum_{i=1}^{2^{(k-1)}} \frac{1}{4} P(\frac{h_i}{a_r})P(\frac{h_i}{b_r})(R(\frac{h_i}{a_r}) - 1)(R(\frac{h_i}{b_r}) - 1) \\ &= n(4n - 1) \sum_{i=1}^{2^{(k-1)}} P(\frac{h_i}{a_r})P(\frac{h_i}{b_r})(R(\frac{h_i}{a_r}) - 1)(R(\frac{h_i}{b_r}) - 1). \end{aligned} \quad (36)$$

In the previous section, we generally introduced the properties of the HTA statistic in three scenarios. Here by using (36), we are ready to prove what we claimed in terms of the haplotype risk.

³Note that n_p is the number of transmission pairs in the data. Thus $2n_p$ haplotypes are observed in the family data.

1. Suppose that the marker being tested M_r is in transmission disequilibrium (due to true linkage/linkage disequilibrium) with the disease while the other markers' transmissions are independent of the disease status.

That is,

$$\begin{aligned} R(\overset{h_i}{a_r}) &= R(a_r), \\ R(\overset{h_i}{b_r}) &= R(b_r). \end{aligned} \tag{37}$$

The allelic risks should also satisfy the equation

$$R(a_r)P(a_r) + R(b_r)P(b_r) = 1.$$

In this case, the sum of products will be less than zero by expectation, that is

$$E(\text{HTA}^r(k)) = n(4n-1) \sum_{i=1}^{2^{(k-1)}} P(\overset{h_i}{a_r})P(\overset{h_i}{b_r}) \frac{P(a_r)}{1-P(a_r)} (R(a_r)-1)(1-R(a_r)) \leq 0. \tag{38}$$

2. Suppose that M_r has no linkage/LD to the disease.

In terms of haplotypic relative risk, we have,

$$\begin{aligned} R(\overset{h_i}{a_r}) &= R(h_i), \\ R(\overset{h_i}{b_r}) &= R(h_i). \end{aligned} \tag{39}$$

Then, the sum becomes

$$E(\text{HTA}^r(k)) = n(4n-1) \sum_{i=1}^{2^{(k-1)}} P(\overset{h_i}{a_r})P(\overset{h_i}{b_r}) (R(h_i)-1)^2 \geq 0 \tag{40}$$

and will be strictly greater than 0 if and only if there is transmission disequilibrium between r^{th} -deleted marker set S_M^r and the disease.

3. *Suppose that the marker M_r is not only linked to the disease, but also shows interaction with other markers in the remaining set of markers. This case is more complicated.*

Without losing generality, we assume there is only one marker having haplotype interactive risk with M_r and that marker is denoted by M_s with alleles c and d . We also assume that the linkage disequilibrium among the markers are negligible. This assumption is reasonable since LD fades away very quickly when the distance between two markers increases. According to [27], the LD is almost zero, for two markers greater than 1 centimorgan apart, which is about the average marker density for current genome scans.

These two diallelic markers produce four possible haplotypes $\begin{smallmatrix} a_r \\ c \end{smallmatrix} |, \begin{smallmatrix} b_r \\ c \end{smallmatrix} |, \begin{smallmatrix} a_r \\ d \end{smallmatrix} |, \begin{smallmatrix} b_r \\ d \end{smallmatrix} |$. Let p_1, p_2, p_3, p_4 be the population frequencies of these four haplotypes and r_1, r_2, r_3, r_4 the corresponding relative risks. From (30), we have $p_1 r_1 + p_2 r_2 + p_3 r_3 + p_4 r_4 = 1$. For each r_i ⁴,

$$\begin{cases} r_i > 1 & \text{higher risk} \\ r_i = 1 & \text{same risk as population} \\ r_i < 1 & \text{lower risk} \end{cases} \quad (41)$$

Now define $r_i = 1 + \Delta_i$ for $i = 1, \dots, 4$, then Δ_i should satisfy,

$$p_1 \Delta_1 + p_2 \Delta_2 + p_3 \Delta_3 + p_4 \Delta_4 = 0. \quad (42)$$

⁴Scenario $r_i = 1$ exists only when the markers are not associated with the disease.

Following this notation, one can write,

$$\begin{aligned}
E(\text{HTA}^r(k)) &= n(4n-1) \sum_{i=1}^{2^{(k-1)}} P\left(\begin{smallmatrix} h_i \\ a_r \end{smallmatrix}\right) P\left(\begin{smallmatrix} h_i \\ b_r \end{smallmatrix}\right) (R\left(\begin{smallmatrix} h_i \\ a_r \end{smallmatrix}\right) - 1)(R\left(\begin{smallmatrix} h_i \\ b_r \end{smallmatrix}\right) - 1) \\
&= n(4n-1) \sum_{\substack{h_i \text{ has } c}}^{2^{(k-1)}} P\left(\begin{smallmatrix} h_i \\ a_r \end{smallmatrix}\right) P\left(\begin{smallmatrix} h_i \\ b_r \end{smallmatrix}\right) (R\left(\begin{smallmatrix} a_r \\ c \end{smallmatrix}\right) - 1)(R\left(\begin{smallmatrix} b_r \\ c \end{smallmatrix}\right) - 1) \\
&\quad + n(4n-1) \sum_{\substack{h_i \text{ has } d}}^{2^{(k-1)}} P\left(\begin{smallmatrix} h_i \\ a_r \end{smallmatrix}\right) P\left(\begin{smallmatrix} h_i \\ b_r \end{smallmatrix}\right) (R\left(\begin{smallmatrix} a_r \\ d \end{smallmatrix}\right) - 1)(R\left(\begin{smallmatrix} b_r \\ d \end{smallmatrix}\right) - 1) \\
&= C(n_p, S_M^r)(p_1 p_2 (r_1 - 1)(r_2 - 1) + p_3 p_4 (r_3 - 1)(r_4 - 1)) \\
&= C(n_p, S_M^r)(p_1 p_2 \Delta_1 \Delta_2 + p_3 p_4 \Delta_3 \Delta_4), \tag{43}
\end{aligned}$$

where $C(n_p, S_M^r) = n(4n-1) \sum_{h'_i}^{2^{(k-2)}} P(h'_i)^2$, and h'_i is defined by marker set S_M excluding (M_r, M_s) , i.e., $\{M_1, \dots, \check{M}_r, \dots, \check{M}_s, \dots, M_k\}$. The sign of the expected value of the haplotype disequilibrium association score depends on the sign of Δ_i 's. We present different possibilities in Table 4.

Table 4 goes here.

Table 4 lists all informative scenarios of possible combinations of Δ_i 's. The last two columns give the signs of the expected scores and the different levels of “importance” in the effect of deleting marker M_r . The table shows strong agreement with the algorithm in that if the marker M_r is important (high risk) the expected score will be negative and the marker will be retained. On the contrary, if M_r is in relatively low risk, the corresponding average score will result in a positive value which may lead to its deletion from the current marker set.

C A heuristic guideline for selecting B in the two-step marker selection procedure

Given that M_r is a candidate marker in the original marker set, let p_1 be the probability that M_r will be returned if it is associated with the disease, and p_2 the corresponding probability if M_r is unassociated with the disease. Assume that

$$p_1 = rp_2, \quad r > 1. \quad (44)$$

Let B be the number of repetitions as defined previously, and let X be the observed frequency that M_r is returned by BHTA screening. To achieve a clear separation of the associated markers from the unassociated markers, B should be large enough to satisfy the following inequality,

$$P(X \geq Bp_2 + 3.1\sqrt{Bp_2(1-p_2)}) \geq 99\%, \quad (45)$$

where $X \sim \text{Norm}(Bp_1, \sqrt{Bp_1(1-p_1)})$ and 3.1 is the critical value that corresponds to the significance level 10^{-3} from a standard normal table. We choose to use this stringent level 10^{-3} to adjust for the possible effect due to multiple comparisons. The quantities $(Bp_1, \sqrt{Bp_1(1-p_1)})$ and $(Bp_2, \sqrt{Bp_2(1-p_2)})$ in (45) are means and standard deviations of the observed frequencies for the associated and unassociated markers, respectively. (Here, we assume that B is large enough so that the distribution of the observed frequencies can be approximated by a normal distribution.) We also choose 99% as a required power level in order to claim a clear separation.

Condition (45) is equivalent to

$$B(p_1 - p_2) \geq 3.1\sqrt{Bp_2(1-p_2)} + 2.33\sqrt{Bp_1(1-p_1)} \quad (46)$$

Let K be the size of the original marker set. The simulations we have conducted suggest

that the average number of markers returned (step 1) is around $1 \sim 2$. Therefore, we have chosen to use a conservative $p_2 \approx \frac{1}{K}$. Inequality (46) can then be rewritten as follows,

$$B(rp_2 - p_2) \geq 3.1\sqrt{Bp_2(1 - p_2)} + 2.33\sqrt{Brp_2(1 - rp_2)} \quad (47)$$

$$B(r - 1)p_2 \geq 3.1\sqrt{Bp_2(1 - p_2)} + 2.33\sqrt{r}\sqrt{Bp_2(1 - p_2) + N(1 - r)p_2^2} \quad (48)$$

(p_2^2 is negligible)

$$\sqrt{B} > \frac{3.1 + 2.33\sqrt{r}}{r - 1} \sqrt{\frac{(1 - p_2)}{p_2}} \quad (49)$$

$$B > \left(\frac{3.1 + 2.33\sqrt{r}}{r - 1}\right)^2 \frac{1 - p_2}{p_2} \approx \left(\frac{3.1 + 2.33\sqrt{r}}{r - 1}\right)^2 (K - 1) \quad (50)$$

If we assume that the association to the trait doubles an associated marker's probability to be selected by the BHTA algorithm, that is $r = 2$, it follows that the suggested value for B is at least $B_s \approx 41(K - 1)$. The calculation above is fairly heuristic and conservative. This explains why we have chosen $B = 5000$ in all examples and project that $B = 20000$ will be needed for 500 markers and $B = 200000$ for 5000 markers.

References

- [1] Falk CT, Rubinstein P (1987) Haplotype relative risk: an easy way to construct a proper control sample for risk calculations. *Ann Hum Genet* 51:227-233
- [2] Ott J (1989) Statistical properties of the haplotype relative risk. *Genet Epidemiol* 6:127-130
- [3] Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *Am J Hum Genet* 52:506-516
- [4] Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and mixture. *Am J Hum Genet* 57:455-464
- [5] Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-Allele marker loci. *Am J Hum Genet* 59:323-336
- [6] Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 59:983-989
- [7] Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 13:423-449
- [8] Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* 48:67-81
- [9] Ott J, Hoh J (2000) Statistical approaches to gene mapping. *Am J Hum Genet* 67:289-294
- [10] Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature* 405(6788):847-856
- [11] Cardon LR, Bell JI (2001) Association study designs for complex diseases. *Nat Rev Genet* 2(2):91-99

- [12] Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11(3):241-247
- [13] Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 69:936-950
- [14] Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science* 265:2037-2048
- [15] Weeks DE, Lathrop GM (1995) Polygenic disease: methods for mapping complex disease traits. *Trends Genet* 11(12):513-519
- [16] Nadeau JH (2001) Modifier genes in mice and humans. *Nat Rev Genet* 2(3):165-174
- [17] Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 70(2):461-471
- [18] Ott J, Hoh J (2001) Statistical multilocus methods for disequilibrium analysis in complex traits. *Hum Mutat* 17(4):285-288
- [19] Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J (2000) Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Ann Hum Genet* 64(Pt 5):413-417
- [20] Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* 11(12):2115-2119
- [21] Clayton D, Jones H (1999) Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 65(4):1161-1169
- [22] Zhao H, Zhang SL, Merikangas KR, Trixler M, Wildenauer DB, Sun FZ and Kidd KK (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 67:936-946

- [23] Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69:1-14
- [24] Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, Scambler PJ, Davidson WS, Beales PL, Lupski JR (2001) Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science* 293(5538):2256-2259
- [25] Burghes HM, Vacassin HEF, de la Chapelle A (2001) The land between Mendelian and multifactorial inheritance. *Science* 293: 2213-2214
- [26] Gurling HMD, Kalsi G, Brynjolfson J, Sigmundsson T, Sherrington R, Mankoo BS, Read T, Murphy P, Blaveri E, McQuillin A, Petursson H, and Curtis D (2001) Genomewide genetic linkage analysis confirms the presence of susceptibility loci for Schizophrenia, on chromosomes 1q32.2, 5q33.2, chromosomes 11q23.3-24 and 20q12.1-11.23. *Am J Hum Genet* 68:661-673
- [27] Kruglyak L (1999) Prospects for whole-Genome linkage disequilibrium mapping of common disease genes. *Nature Genet* 22:139-144
- [28] Simonic I, Nyholt DR, Gericke GS, Gordon D, Matsumoto N, Ledbetter DH, Ott J, Weber JL (2001) Further evidence for linkage of Gilles de la Tourette syndrome (GTS) susceptibility loci on chromosomes 2p11, 8q22 and 11q23-24 in South African Afrikaners. *Am J Med Genet* 105(2):163-167
- [29] Merette C, Brassard A, Potvin A, Bouvier H, Rousseau F, Emond C, Bissonnette L, Roy MA, Maziade M, Ott J, Caron C (2000) Significant linkage for Tourette syndrome in a large French Canadian family. *Am J Hum Genet* 67(4):1008-1013
- [30] The Tourette Syndrome Association International Consortium for Genetics (1999) A complete genome screen in sib pairs affected by Gilles de la Tourette syndrome. *Am J Hum Genet* 65(5):1428-1436

- [31] Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet* 56(3):777-787.

Figure 1: Flowchart of BHTA screening

BHTA removes the least important marker from the marker set and continue to the next iteration until all the remaining markers present evidence of important information.

Figure 2: Information (HTD) flow during BHTA screening.

Assume the trait of interest has three susceptibility loci, with interacting effects (see example 1 in next section). Markers indicated by filled symbols are simulated to be in linkage/disequilibrium with one of the disease genes, while shape differentiates the target genes (2 markers at each disease locus). Markers indicated by empty circles are not in linkage/disequilibrium with the disease. As shown in the left plot: at first the information regarding the disease is contaminated by the noise due to the unassociated markers; as BHTA screens out more and more irrelevant markers, the information indicated by HTD begins to grow and the algorithm will stop at the peak (due to a significant drop after this deletion), thereby returning 3 important markers that cover all three susceptibility loci of the trait. Right plot: no marker has linkage/association with the trait. Thus the information stays low throughout the screening and the algorithm will return no markers.

Simulation specifications:

1. Disease loci are generated to be physically independent, in other words, on different chromosomes.
2. Disease genes population frequencies are set to 0.5.
3. Linkage/LD between associated marker and disease loci: $\theta = 0.01$ (recombination fraction) and $\Delta = 0.9$ (standardized LD).
4. Marker allele population frequencies are randomly generated with equal probabilities.
5. Inter-marker distance and association: randomly generated from $\theta \sim U(0.05, 0.1)$ and $\Delta \sim U(0.1, 0.2)$.

Figure 3: Relative probability of deletion: associated and unassociated markers.

Left panel: same as in Figure 2, six markers are simulated to be in linkage/association with 3 disease genes. $B = 5,000$ repetitions of step 1 are run on a data set of 200 patients. The solid line represents the average relative probability of deletion of unassociated markers (24 of them), and the dotted line represents the average relative probability of being deleted for the 6 associated markers.

Right panel: Markers are simulated with no linkage/association to the disease, while using the same marker labeling as the alternative. Under the null, the two relative probability of deletion lines not only stay constant around 1 (indicating random deletions), but also have no significant difference between one another as well

Simulation specifications:

1. Disease loci are generated to be physically independent, in other words, on different chromosomes.
2. Disease genes population frequencies are set to 0.5.
3. Linkage/LD between associated marker and disease loci: $\theta = 0.01$ (recombination fraction) and $\Delta = 0.9$ (standardized LD).
4. Marker allele population frequencies are randomly generated with equal probabilities.
5. Inter-marker distance and association: randomly generated from $\theta \sim U(0.05, 0.1)$ and $\Delta \sim U(0.1, 0.2)$.

Figure 4: Results from BHTA two-step marker selection (example 1).

For this example, $B = 5,000$ repetitions of step 1 were run in stage 2. The frequencies of the markers being selected by the BHTA screening are plotted. Dotted line in the plot stands for the selection threshold using the IQR-based criterion. Markers that are returned more frequently than this threshold (circled) will be selected in the resulting set. All the markers selected in this example are in true linkage/association with the trait. It is worth noting that one associated marker fails the calculated threshold. Actually it doesn't hurt the performance of the algorithm since two markers are simulated at each disease locus and returning either one of them will capture the information regarding this disease gene in the selected marker set

Simulation specifications:

1. Disease loci are generated to be physically independent, in other words, on different chromosomes.
2. Disease genes population frequencies are set to 0.5.
3. Linkage/LD between associated marker and disease loci: $\theta = 0.01$ (recombination fraction) and $\Delta = 0.9$ (standardized LD).
4. Marker allele population frequencies are randomly generated with equal probabilities.
5. Inter-marker distance and association: randomly generated from $\theta \sim U(0.05, 0.1)$ and $\Delta \sim U(0.1, 0.2)$.

Figure 5: Comparison between BHTA and TDT—modifier gene model.

- [a] Probability for an associated marker being detected using one data set of 200 patients (9 associated markers, 3 markers per gene).
- [b] Detecting a disease locus is defined as the return of at least one of the associated markers at this locus. Average the detecting probability for individual disease loci over three disease loci as specified in the example.
- [c] Probability for detecting all three disease loci for one data set of 200 patients using the given method (TDT/BHTA).
- [d] Average number of loci detected using a data set of 200 patients.

Probability or power is estimated using tests run on 1,000 random samples of 200 patients. In order to give a fair comparison between the two selection procedures, same type I error rate 0.001 is deployed in both methods.

Simulation specifications:

1. Disease loci are generated to be physically independent, in other words, on different chromosomes.
2. Disease genes population frequencies are set to 0.15.
3. Linkage/LD between associated marker and disease loci: $\theta = 0.01$ (recombination fraction) and $\Delta = 0.8$ (standardized LD).
4. Marker allele population frequencies are randomly generated with equal probabilities.
5. Inter-marker distance and association: randomly generated from $\theta \sim U(0.05, 0.1)$ and $\Delta \sim U(0.1, 0.2)$.

Figure 6: Comparison between BHTA and TDT—polygenic admixture model.

- [a] Probability for an associated marker being detected using one data set of 150 patients (3 associated markers, 1 marker per gene).
- [b] Probability for detecting all three disease loci for one data set of 150 patients using the given method (TDT/BHTA).
- [c] Average number of loci detected using a data set of 150 patients.

Probability or power is estimated using tests run on 1,000 random samples of 150 patients. In order to give a fair comparison between the two selection procedures, same type I error rate 0.001 is deployed in both methods.

Simulation specifications:

1. Disease loci are generated to be physically independent, in other words, on different chromosomes.
2. Disease genes population frequencies are set to 0.05.
3. Linkage/LD between associated marker and disease loci: $\theta = 0.01$ (recombination fraction) and $\Delta = 0.8$ (standardized LD).
4. Marker allele population frequencies are randomly generated with equal probabilities.
5. Inter-marker distance and association: randomly generated from $\theta \sim U(0.05, 0.1)$ and $\Delta \sim U(0.1, 0.2)$.

Haplotype	HRR	
abe	<i>high</i>	2
abE	<i>low</i>	0
aBe	<i>low</i>	0
aBE	<i>high</i>	2
Abe	<i>low</i>	0
AbE	<i>high</i>	2
ABe	<i>high</i>	2
ABE	<i>low</i>	0

Table 1: Hypothetical haplotype relative risk model (example 1).

Assume that the hypothetical complex trait has three responsible genes (alleles A/a, B/b, and E/e). The risk of this disease is determined by the haplotypes inherited from the parents. For simplicity, we assume that the genotypic relative risk (GRR) is the product of two haplotype relative risks (HRR), which is known as the multiplicative model.

Simulation specifications:

1. Disease loci are generated to be physically independent, in other words, on different chromosomes.
2. Disease genes population frequencies are set to 0.5.
3. Linkage/LD between associated marker and disease loci: $\theta = 0.01$ (recombination fraction) and $\Delta = 0.9$ (standardized LD).
4. Marker allele population frequencies are randomly generated with equal probabilities.
5. Inter-marker distance and association: randomly generated from $\theta \sim U(0.05, 0.1)$ and $\Delta \sim U(0.1, 0.2)$.

	Markers						Notes
group 1	26	18	20	10			
	Haplotypes				Trans.***	Untrans.	
P_{mhet} =0.003**	0*	0	0	0	13	6	
	0	0	0	1	5	6	Marker 26
	0	0	1	0	24	19	Chr 8 (D8S559)
	0	0	1	1	8	10	Marker 18
	0	1	0	0	5	2	Chr 11 (D11S1377)
	0	1	0	1	3	1	Marker 20
	0	1	1	0	5	3	Chr 11 (D11S4464)
	0	1	1	1	7	2	Marker 10
	1	0	0	0	14	7	Chr 20 (D20S1085)
	1	0	0	1	14	18	
	1	0	1	0	20	48	
	1	0	1	1	35	21	
	1	1	0	0	5	9	
	1	1	0	1	10	6	
	1	1	1	0	18	14	
	1	1	1	1	7	21	
group 2	4	10	12				
	Haplotypes				Trans.	Untrans.	
P_{mhet} = 0.0017	0	0	0		17	5	
	0	0	1		7	6	Marker 4
	0	1	0		25	28	Chr 8 (GATA8B01)
	0	1	1		35	28	Marker 10
	1	0	0		19	19	Chr 20 (D20S1085)
	1	0	1		24	16	Marker 12
	1	1	0		39	31	Chr 20 (D20S469)
	1	1	1		17	50	
group 3	14	6	26				
	Haplotypes				Trans.	Untrans.	
P_{mhet} = 0.004	0	0	0		20	31	
	0	0	1		8	5	Marker 14
	0	1	0		54	32	Chr 2 (D2S440)
	0	1	1		22	40	Marker 6
	1	0	0		19	10	Chr 8 (D8S270)
	1	0	1		9	9	Marker 26
	1	1	0		20	36	Chr 8 (D8S559)
	1	1	1		27	16	

Table 2: Results on data from Simonic *et al* [28] using BHTA methods.

* Markers used in this study are mostly short tandem repeat polymorphisms, typically with more than 2 alleles. To implement BHTA methods, all markers are dichotomized by combining alleles. The alleles are lumped according to their frequencies of tandem repeats so that the two combined “alleles” will have approximately equal size.

** Extended TDT for multi-allelic markers proposed in [6] is used to evaluate the evidence of linkage/LD contained in the haplotypes, that is $TDT_{mhet} = \frac{H-1}{H} \sum_i \frac{(n_i^t - n_i^u)^2}{n_i^t + n_i^u}$. P-values listed in the table are obtained from a χ^2 distribution with $H-1$ degrees of freedom.

*** Transmission counts in this table are from haplotype-wise heterozygous parents only.

markers	Returning Frequency: mean (std. dev)	
	Phases Known	Random Assignments
Associated	709 (285)	720 (281)
Unassociated	87 (127)	85 (129)

Table 3: Impacts of random phase assignments

Under the polygenic admixture model, screenings are run using the exact simulated haplotype phases and haplotypes with possible random phase assignments, respectively. To assess the impacts of the random phase assignments, the mean and standard deviation of the returning frequencies (out of 5000 screenings) for the associated and unassociated markers under these two scenarios are listed in the table for comparison.

Simulation specifications:

1. Disease loci are generated to be physically independent, in other words, on different chromosomes.
2. Disease genes population frequencies are set to 0.05.
3. Linkage/LD between associated marker and disease loci: $\theta = 0.01$ (recombination fraction) and $\Delta = 0.8$ (standardized LD).
4. Marker allele population frequencies are randomly generated with equal probabilities.
5. Inter-marker distance and association: randomly generated from $\theta \sim U(0.05, 0.1)$ and $\Delta \sim U(0.1, 0.2)$.

$\Delta_1, \left(\begin{smallmatrix} a_r \\ c \end{smallmatrix} \middle \right)$	$\Delta_2, \left(\begin{smallmatrix} b_r \\ c \end{smallmatrix} \middle \right)$	$\Delta_3, \left(\begin{smallmatrix} a_r \\ d \end{smallmatrix} \middle \right)$	$\Delta_4, \left(\begin{smallmatrix} b_r \\ d \end{smallmatrix} \middle \right)$	$E(\text{HTA}^r(k))$	M_r importance
$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	negative	high
$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	negative	high
$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	negative	high
$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	negative	high
$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	<i>positive</i>	<i>low</i> **
$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	negative	high
$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} - \\ (+) \end{smallmatrix}$	$\begin{smallmatrix} + \\ (-) \end{smallmatrix}$	negative	high

Table 4: Sign of expected values of HTA scores.

Assume two markers are in haplotypic transmission disequilibrium with the disease, one with alleles a_r and b_r , the other with c and d . Each of the first 4 columns stands for one possible haplotype defined by these two markers. Signs of “+” or “−” mean either higher risk (> 1) or lower risk (< 1) (see text). Thus, each row specifies one possible haplotype relative risk pattern. The expected value of HTA score is studied under each scenario and the importance of the marker with allele a_r and b_r is evaluated according to the sign of $E(\text{HTA}^r(k))$ according to the properties of HTA. (Inference is based on $p_1 \approx p_2 \approx p_3 \approx p_4$, which is usually true for carefully selected markers.) Signs in parentheses in each row indicates another combination that leads to the same conclusion.

** : note that the alleles of M_r contribute minor effects comparing to the alleles of M_s . By changing M_r from a_r to b_r on any background haplotype, the risk won’t change sign, while the changes of alleles on the locus of M_s will “flip” the risks from high to low.