Multilocus Linkage Analysis of Affected Sib Pairs

Iuliana Ionita¹, Shaw-Hwa Lo²

¹Courant Institute of Mathematical Sciences, New York University, ²Department of Statistics, Columbia University, New York, N.Y., USA

1 Introduction

Traditional approaches to linkage analysis assign a score to each marker position by considering the linkage information given by that marker or a few nearby markers. These approaches have been very successfully applied to Mendelian diseases; however, they have been less fruitful in the context of complex diseases. Because complex genetic diseases are caused by the action of several genes that can interact in a complicated manner, methods that can exploit interactions among multiple disease loci are expected to be more powerful. Here we report a novel linkage method for affected sib pairs (ASPs). Our approach screens a large number of polymorphisms and selects a few that seem linked to disease. The selection is based on an importance score assigned to each marker based on both marginal information as well as information coming from possible interactions among several disease loci.

Several multilocus linkage methods have been reported in the literature. These include model-based methods and model-free methods. The model-based methods calculate the full likelihood of disease and marker data under the assumed mode of inheritance (usually two-locus models, Schork et al. [1]). The model-free methods are based on comparing the observed allele-sharing among relatives that are phenotypically alike with that expected under no linkage. Their main characteristic is that they do not assume a specific mode of inheritance. Examples include Cordell et al. [2] and Farrall [3] for ASPs. Their methods...
are based on computing a maximum likelihood statistic (MLS) and are restricted to two-locus models. More recently, Cordell et al. [4] have presented a generalization of the MLS method in Cordell et al. [2] to several disease loci and affected relative pairs. Given linkage evidence at \( m - 1 \) loci, the evidence at the \( m \)-th locus is measured by the difference in MLS between the best fitting \( m - 1 \) locus model and the best fitting \( m \) locus model. However, due to the sparseness of the data when \( m \) increases and the large number of parameters that need to be estimated in their model-fitting procedure, the method is useful in practice only for the simultaneous analysis of at most \( \sim 3 \) disease loci. Also these methods are applicable only after a primary genome screen has already been performed, when the number of loci under investigation is small.

Here we report a new screening method. Our method works on datasets with a large number of markers and makes no assumption on the disease model, including the number of disease loci and their positions in the genome. This new approach uses the interactions among several disease loci to help increase the importance of moderate effect disease loci relative to other noisy loci. The method is based on the repetition of a two-phase selection-reduction process. In the first step (‘selection’) we select a small set of markers at random from the available list of polymorphisms. In the next step (‘reduction’), we remove the unimportant markers from the current set one by one in a stepwise fashion until all the remaining markers are important or a single marker remains (we call these markers ‘returned’). At the end of this process we count how many times each marker was returned. Based on these counts we decide which markers are returned at significantly high frequency. The key technical aspect of this procedure is the definition of a statistic to measure the relative importance of a marker in the current set.

We apply this new approach to real data (inflammatory bowel disease, InfBD) as well as data simulated under several complex models. The results are very good. On the real data we confirm most of the known loci. On simulated data we show that our method is consistently more powerful than the single-locus methods currently in use.

The rest of the paper is organized as follows. In section 2 (Methods) we illustrate the theoretical aspects of our approach. In section 3 (Results) we present our findings on a real dataset for InfBD and on simulated data. We conclude in Section 4 (Discussion) with a discussion of our findings.

2 Methods

2.1 Linkage Measure

The core of our approach is the definition of a linkage measure for a set of markers. In this section we describe this measure.

Notation 2.1

Most model-free methods for ASPs work with the genotypic identical-by-descent (IBD) sharing at a locus, which can be 0, 1 or 2. In our approach we work with the allelic IBD status; in this case the IBD sharing can be 0, 1 or 2. In our approach we work with the allelic IBD status; in this case the IBD sharing can be 0, 1 or 2. In our approach we work with the allelic IBD status; in this case the IBD sharing can be 0, 1 or 2. In our approach we work with the allelic IBD status; in this case the IBD sharing can be 0, 1 or 2. In our approach we work with the allelic IBD status; in this case the IBD sharing can be 0, 1 or 2.
E\left[^{n_{1,1} - n_{0,0}}^2\right] \\
= \text{Var}\left[^{n_{1,1} - n_{0,0}^2}\right] + \text{Var}\left[^{n_{1,1} - n_{0,0}^2}\right] - 2\text{Cov}\left[^{n_{1,1} - n_{0,0}^2}\right] \\
= Np_{1,1}^2(1 - p_{1,1}^2) + Np_{0,0}^2(1 - p_{0,0}^2) + 2Np_{1,1}p_{0,0} \\
= N\left[p_{1,1}^2 + p_{0,0}^2\right] - N\left[p_{1,1}^2 + p_{0,0}^2\right]^2 = N\left[p_{1,1} - p_{0,0}^2\right] = \frac{N}{2^k}.

Considering this it is easy to see that:

\[
\frac{w_{k+1}}{w_k} = \frac{2^{k+1} - 1}{2^k - 1}
\]

hence the resulting weight \( w_k \).

It is revealing to rewrite the linkage measure as follows:

\[
H_{1...k} = \frac{w_k}{w_{k-1}} H_{2...k} + H_{13...k} + \cdots + H_{1...k-1} + w_k \left(n_{1,k-1} - n_{0,k-1}\right)^2
\]

Then for \( k \geq 4 \) we have \( w_k \approx 1 \) and \( w_k \approx w_{k-1} \). Hence we can write:

\[
H_{1...k} \approx \frac{H_{2...k} + H_{13...k} + \cdots + H_{1...k-1}}{k} + \left(n_{1,k-1} - n_{0,k-1}\right)^2
\]

Essentially, our measure is defined recursively as follows. We start with the natural NPL-like measure for one marker \( H_1 = 2(n_1 - n_0)^2 \). The measure for \( k \) markers \( H_{1...k} \) is obtained as the average of the measures for all possible \( k \) combinations of \( k - 1 \) markers: \( H_{2...k}, H_{13...k}, ..., H_{1...k-1} \) plus an additional term \((n_{1,k-1} - n_{0,k-1})^2\) that measures the interaction of all \( k \) markers together.

Notice that when none of the \( k \) markers is linked to disease we have:

\[
E\left[^{n_{1,k} - n_{0,k}^2}\right] = \frac{N}{2^{k-1}}.
\]

Thus the interaction term tends to be small in this case \( (O(N)) \). However when all \( k \) markers are linked to disease, this term will become large (due to \( E^2[n_{1,k} - n_{0,k}^2] = O(N^2) \)).

Remark 2.2

Our experiments show that under the assumption of no specific interaction model (e.g. epistasis or heterogeneity), the other possible pieces of information that we could use in the definition of the measure (e.g. \( n_{10}, n_{01}, \text{etc.} \)) may introduce noise (e.g. in the case of disease loci that interact epistatically). Certainly \( n_{10} \) and \( n_{01} \) contain information in a two-locus heterogeneity model, but the choice of a consistent statistic that would work for different scenarios forces us to disregard these terms and instead focus on \( n_{11} \) and \( n_{00} \). Notice that under both the epistatic and the heterogeneity interaction models for two disease loci \( E(n_{11} - n_{00}) > 0 \), whereas when none of the loci is linked to disease \( E(n_{11} - n_{00}) = 0 \).

2.2 Screening Algorithm

The screening procedure consists of a marker selection-reduction process described below. Suppose we have a list of many markers (hundreds in a whole-genome study). We proceed as follows:

- **Step 0:** Repeat steps 1–4 \( B \) times (\( B \geq 3000 \) is a fairly large number).
- **Step 1:** Start by choosing a set of \( k \approx 10 \) markers at random from the available list of markers.
- **Step 2:** At each step compute for each marker in the current set the resulting change in the linkage measure when that marker is removed. For marker \( i \), \( \Delta_i = H_{1...k+1} - H_{1...k} \). If \( \Delta_i > 0 \), then the linkage measure decreases when removing marker \( i \) and therefore marker \( i \) is important relative to the other markers present in the current marker set. If \( \Delta_i < 0 \), then the linkage measure increases when removing marker \( i \) and therefore marker \( i \) is not important relative to the other markers present.
- **Step 3:** Remove the marker \( i \) (if any) with the largest positive \( \Delta_i \) from the current set.
- **Step 4:** Do Steps 2–3 until either all the markers in the current set are important (all \( \Delta_i \) are negative) or only one marker remains. The returned markers are recorded.
- **Step 5:** We compute for each marker a final return count denoting the total number of times it was returned in Step 4. Based on these counts we separate the markers into two classes: the important/linked to disease markers and the unimportant/unlinked ones. The details of this statistical procedure are given in Section 2.4.

2.3 Why It Works

The behavior of the screening algorithm in section 2.2 depends heavily on the properties of the statistic \( H_{1...k} \). We formulate these properties in the lemma below. The main idea is that in expectation only markers that are linked to disease are returned in Step 4 and markers that are not linked tend to be removed in Step 3. Let \( S = \{1, ..., k\} \) be the current set. For the lemma below we make the simplifying assumption that the \( k \) markers are not linked among themselves.

**Lemma 2.3**

The following properties are true:

1. If none of the markers is linked to disease, then for any marker \( i \) in \( S \) we have:

\[
E[H_{1...k+1}] = E[H_{1...k}]
\]

2. If \( S \) contains one marker linked to disease (without loss of generality, assume this is the first marker) and the rest are unlinked, then for any unlinked marker \( u \) in \( S \) we have:

\[
E[H_{1...k}] < E[H_{1...k+u}] < E[H_{1...k+u+1}]< E[H_{1...k+u+2}] ...
\]

3. If the set \( S \) has some interacting markers, linked to disease, of similar relative importance and some unlinked markers, then for any linked marker \( l \) and any unlinked one \( u \) we have:

\[
E[H_{1...k+1}] < E[H_{1...k}] < E[H_{1...k+u+1}]
\]

4. If the current set \( S \) contains only markers linked to disease that are of similar relative importance and also have non-negligible interaction, then for any marker \( l \) in \( S \)

\[
E[H_{1...k+1}] < E[H_{1...k}]
\]

Proof: see Appendix A.
Remark 2.4
We made the assumption that the \( k \) selected markers in Step 1 of the Screening Algorithm are unlinked among themselves. Given that in the majority of cases the \( k \) (\( \approx 10 \)) markers chosen at random from a large number of markers are unlinked among themselves and also because of ease of computation, that assumption is reasonable. However, even when some of the markers in the current set are linked, the effect tends to be very small (computations not shown), and the screening algorithm behaves as desired.

To better illustrate these properties, we simulated a small dataset with 7 markers. The first two of these are each closely linked (\( \theta = 0.01 \)) to a different disease gene. The other five are unlinked to disease. The disease model is epistatic RR, i.e. two mutations at each of the two disease loci are necessary to have disease. As shown in figure 1a, the measure \( H_{12347} \) decreases significantly when removing either one of the linked markers (1 or 2) and increases significantly when removing either of the unlinked markers (3, 4 or 7). In figure 1b we see that when none of the markers in the current set is linked to disease, the values of the measure are small and not as well separated as the ones in figure 1a. In fact a random (unlinked) marker is removed. In this example (fig. 1b), marker 3 is removed (i.e. \( \Delta_3 = H_{34567} - H_{34567} \) is the largest positive \( \Delta \)).

Note that we have:

\[
H_{12347} \approx \frac{H_{1347} + H_{1247} + H_{1234} + H_{1237} + H_{1247}}{5} \quad \text{for (a)}
\]

\[
H_{34567} \approx \frac{H_{1467} + H_{3456} + H_{4357} + H_{3567} + H_{4567}}{5} \quad \text{for (b)}
\]

Therefore according to Eq (4) the interaction term \((H_{11111} - n_{00000})^2\) is small in this case due to the presence of unlinked markers in the current set.

2.4 Important vs. Unimportant Markers

The goal of our method is to separate the important/linked to disease markers from the unimportant/unlinked markers. We present two different methods to achieve this goal. Both methods yield a good balance between false positive results and true positive results. In our experience the two methods behave similarly.

1. Normal-Mixture Method

We first fit a two-component normal-mixture model to the histogram of return counts:

\[
p_1N(\mu_1, \sigma_1^2) + p_2N(\mu_2, \sigma_2^2)
\]

where \( \mu_2 > \mu_1 \) and \( p_2 = 1 - p_1 \); \( \mu_2 \) and \( \mu_1 \) are the means for the distribution of important and unimportant markers respectively. To control the false-positive rate (FPR), we select as threshold the \( 1 - \alpha \) percentile for the unimportant markers at a certain level \( \alpha \). The markers that have a return higher than this cutoff are claimed to be important (linked to disease genes).

2. Efron’s Method

Another method to achieve this separation is based on an idea of Efron [5]. He proposes a method to divide the data values into two classes, interesting and uninteresting, when a large number of tests need to be evaluated as is the case in whole-genome scans. This is in contrast to the classical significant versus non-significant categorization used when the number of tests is small.

The method first fits a natural spline to the histogram of return counts by Poisson regression. We call this curve: \( f \) (mixture density). Also an empirical null distribution is estimated, denoted by \( f_0 \) (empirical null density). Then for each marker \( M \) the local false discovery rate (LocFdr) is defined as:

\[
\text{locfdr} (M) = \frac{p_0 f_0 (M)}{f (M)}
\]

Controlling the false discovery rate suggests that the markers with \( \text{locfdr} < \alpha \) be declared interesting (for a certain level \( \alpha \)).

2.5 Choice of \( B \) and \( k \)

As explained in section 2.2, our screening algorithm repeats \( B \) times the process of random selection of \( k \) markers and then evaluation of each of the markers in that set. We want to choose \( B \) and \( k \) large enough such that we get as clear a separation between the markers linked to disease and the unlinked ones as possible. We present a heuristic derivation of a formula for \( B \) in the Appendix.
The formula predicts conservatively that for 200 markers \( B \) should be about 8,000 and for 500 markers \( B \approx 20,000 \). The size of \( k \) influences the number of times certain markers are chosen together in the random subset. It should not be too small, since we want a good probability to select markers together. On the other hand, due to the sparseness of the data in large dimensions and also due to computational issues, \( k \) should not be too large. In our experience \( k = 10 \) works well.

### 3 Results

We evaluated our method on both simulated data and real data.

#### 3.1 Application to Simulated Data

We applied our method to two complex disease models.

##### 3.1.1 First Simulated Disease Model

In the first disease model there are 9 unlinked disease loci. The disease is present when at least 5 of the 9 disease genes are mutated. The sample contains 200 ASPs genotyped at 50 markers, with 20\% of the data sporadic (diseased because of nongenetic causes). Nine markers out of the total of 50 are linked to disease genes (\( \theta = 0.05 \)), one marker for each disease gene. The rest are independent markers, not linked to disease and among themselves. The disease gene frequencies are all set to 0.05 and the marker frequencies are all 0.5. We assume we have complete data: the inference of the IBD sharing is without ambiguity. For each marker we compute two statistics:

- the single-locus statistic (the ASP mean test):

\[
\frac{(n_1 - n_0)^2}{n_1 + n_0} \sim \chi^2
\]

where \( n_1 (n_0) \) is the number of 1 (0) IBD sharing at that particular marker.

- the return count computed by the proposed method (\( B = 3000 \) and \( k = 10 \) in our screening procedure).

For each of the two methods we report the number of loci above certain significance thresholds: [1, 2, ..., 10\%] false positive rates. Since in the simulated data we know exactly which markers are linked to disease and which are unlinked, we can approximate the threshold corresponding to a specific false positive rate empirically by simulation.

Figure 2 shows an example of a simulated dataset according to the complex model outlined above. The horizontal lines represent the thresholds for the 1, 2, 5 and 10\% FPR. It illustrates the advantage of our method; because the markers linked to disease are returned together in Step 4 of the screening algorithm in section 2.2, they will separate better from the unlinked markers. Therefore the proposed method can be very powerful in increasing

![Fig. 2. Comparison between a simple single-locus method (a) and our new method (b) on a complex disease model with 9 disease loci. The figure illustrates how the multilocus approach can increase the significance of moderate effect loci.](image-url)
the importance of disease loci of moderate effect by making use of interactions among disease loci.

To investigate the power, we generated 600 independent replicates. Figure 3a depicts the average percentage of disease loci selected by each of the two methods while keeping the false positive rate at the \{1, 2, ..., 10\%\} level. Our method is more powerful than the single-locus method at all levels. At the 1\% significance level, our method discovers on average 3.1 of the 9 disease loci, while the single-locus method finds only 2.2 loci. Similarly at the 3\% level we detect on average 4.5 loci, while the single-locus method finds 3.4 loci.

Finally, we compared the increase in sample size necessary for the single-locus method to achieve similar power to that of the multilocus method. For this simpler disease model, a 10–15\% increase is necessary.

We then repeated the same simulations, but this time we introduced small linkage disequilibrium (LD) levels between some of the disease genes and the nearby linked markers. Namely, \(\delta_1 = \delta_2 = 0.5\) and \(\delta_3 = \delta_4 = 0\) where \(\delta\) is the normalized LD measure. We compared the single-locus approach to a modified version of our multilocus linkage method (see Appendix D) that can also take advantage of mild linkage disequilibrium between disease loci and nearby markers. In this case the results (fig. 5) are even better compared to the ones obtained in figure 4 where no linkage disequilibrium was present. The improvement at the 1\% FPR is 23\% and an increase in sample size of over 25\% is necessary for the single-locus linkage method to achieve similar performance as the modified multilocus linkage method.

### 3.1.2 Second Simulated Disease Model

We also simulated a similar disease model with 4 disease loci. Now the disease is present when at least 2 of the 4 disease genes are mutated.

Figure 4a depicts the average percentage of disease loci selected by each of the two methods while keeping the false positive rate at the \{1, 2, ..., 10\%\} level. As we can see, our method is more powerful than the single-locus method at all significance levels. In figure 4b we illustrate the increase in sample size necessary for the single-locus method to attain similar power to that of the multilocus method. For this simpler disease model, a 10–15\% increase is necessary.

### 3.2 Application to Real Data (Inflammatory Bowel Disease)

We also analyzed a real dataset for InfBD using our method. InfBD consists of two disorders: Crohn’s Disease (CD) and Ulcerative colitis (UC). They are both inflammatory disorders of the gastrointestinal tract with a strong genetic contribution as revealed by epidemiological stud-
ies. Genome-wide searches for InfBD susceptibility loci have identified several regions of interest, showing that InfBD is a complex genetic disease caused by the action of several genes.

The present dataset is a genome screen of 106 ASPs (including parents) from Canada, affected with CD genotyped at 457 microsatellite markers; the average marker spacing is $\sim 10$ cM. These data have been previously analyzed in Rioux et al. [7] and in Lo and Zheng [8].
In order to apply our new approach to these data, we first inferred the IBD sharing probabilities for each ASP using the program GENEHUNTER 2.0 (Daly et al. [9]). Since our method requires complete IBD sharing information, we probabilistically impute the IBD sharing values. More exactly, for each sib pair under study we generate the IBD value at each position in the sharing vector according to the corresponding sharing probabilities (as calculated by GENEHUNTER 2.0); for example if the sharing probabilities at a certain position are (0.2, 0.5, 0.3) for sharing 0, 1 and 2 alleles respectively, then we generate the IBD value 0, 1 or 2 according to this distribution. In order to minimize the bias due to these probabilistic imputations, we do it 100 times, each time generating a new dataset.

We applied our algorithm on each of the 100 generated datasets. We use $B = 20,000$ and $k = 10$ in our screening procedure. The return counts (averaged over the 100 datasets) for all markers together with the fitted two-component normal mixture are depicted in figure 6b. By controlling the false positive rate at a stringent level we obtain that markers with return count above 240 should be reported as important. In order to minimize the bias due to these probabilistic imputations, we do it 100 times, each time generating a new dataset.

We also applied Efron’s approach and by using a 1% cut-off for the LocFdr, we obtain similar results (see Appendix C).

In figure 7 we show the return counts plotted versus marker locations in the genome. The mean return count is 140 and is marked by a horizontal solid line. The threshold for declaring a marker important is 240 and is marked by a broken line.

The results we obtain are extremely significant. We validated 6 (IBD1, IBD3, IBD5, IBD6, IBD7, IBD8) of the 8 known InfBD loci. Additionally, we found several other interesting regions.

1. The region 1q21 contains a cluster of genes influencing epidermal differentiation. This region is linked to other inflammatory diseases, e.g. psoriasis; psoriasis can occur in association with InfBD (Crohn’s disease), suggesting that they may share common genetic risk factors.

2. The locus on chromosome 2p11 (D2S1790) is located ~10 cM from the gene IL1R1 (interleukin 1 receptor, type 1). There is evidence for the activation of the mucosal immune system and the production of inflammatory cytokines, i.e. interleukin (IL)-1ra and IL-1beta, in the InfBD (Heresbach et al. [10]).
The region 2q32 harbors the STAT1 and STAT4 genes (signal transducers and activators of transcription), which are candidate genes for InfBD (Barmada et al. [11]).

The locus on chromosome 3p: suggestive linkage in this region was found in Rioux et al. [7].

The locus on chromosome 7p13: gene IGFBP3 (insulin-like growth factor binding protein 3) maps to this region. Katsanos et al. [12] found that the serum IGFBP3 levels are reduced in patients with InfBD.

The locus on chromosome 21q22.2 (D21S1809) is close to the TFF1 and TFF2 (trefoil factor 1 and 2) genes. These genes, located on 21q22.3, are expressed in the gastrointestinal mucosa. Increased levels of TFF1 and TFF2 have been found in serum from InfBD patients (Vestergaard et al. [13]).

In Appendix E we give the table with all selected markers and their chromosomal positions.
4 Discussion

We presented a new model-free multilocus linkage method for affected sib pairs. Our approach selects from a large number of polymorphisms a small number that appear to be linked to disease. No assumption is made on the disease model, including number of disease loci or their positions in the genome. It uses both the marginal linkage information as well as information coming from the possible interaction among several disease loci to boost the significance of modest single-locus effects. A further advantage of our method is that it can be naturally extended to take into consideration small linkage disequilibrium levels between disease loci and nearby markers, thereby gaining even greater increases in power over single-locus linkage methods. Further details on this method will appear in a paper under preparation.

We evaluated our method on both simulated data and real data. The extensive simulations that we did show consistently that the proposed approach is more powerful than the conventional single-locus linkage methods at all significance levels (up to 40% increase in power). The improvement in power increases when the number of interacting disease loci increases. In the absence of interactions our method performs similarly to the single-locus methods. Also the results on the real data are highly significant. We validated 6 of the 8 known InfBD loci and also found a few interesting loci, some of which have been already implicated in InfBD pathogenesis.

Our method is also very general; the disease loci can be anywhere in the genome (possibly on different chromosomes) and they can interact in complex, unknown ways. We did make a simplifying assumption, namely we assumed that the selected markers in the current set are unlinked among themselves. It is clear, however, that the effect of linkage between two unimportant markers is suppressed by the presence in the current set of a marker linked to disease. This point is best illustrated on real data, where we see that markers close together do not tend to have return counts higher than expected (e.g. chromosome 4 in figure 7).

The software implementing the proposed methods is available from the authors upon request. A complete package will be available online soon.

Given the complex nature of the common diseases and the many challenges in genome-wide scans, we believe that our approach is very relevant; by using both the marginal and the interaction information, our method performs better than the traditional single-locus methods. Also due to its generality, the proposed method is applicable to a large number of situations.

Acknowledgments

We thank Tian Zheng for many useful discussions. We also thank Eric Lander and Mark Daly from the Whitehead Institute for Biomedical Research for providing the data on InfBD. This research is partially supported by the National Institutes of Health (NIH) (grant RO1 GM070789-01). We also want to thank two anonymous reviewers for their comments that helped improve the manuscript.

Appendix A

Proof of Lemma 2.3

Let $S = \{1, ..., k\}$. Assume the $k$ markers in set $S$ are unlinked among themselves.

Lemma 2.3.

The following properties are true:

1. If none of the markers in the current set $S$ is linked to disease, then for any marker $i$ in $S$ we have:

$$E[H_{1\ldots i-1\ldots i+1\ldots}] = E[H_{1\ldots i}]$$

2. If the current set $S$ contains one marker linked to disease (without loss of generality, assume this is the first marker) and the rest are unlinked, then for any unlinked marker $u$ in $S$ we have:

$$E[H_{2\ldots i\ldots}] < E[H_{1\ldots i\ldots}] < E[H_{1\ldots u-1\ldots u+1\ldots}]$$

3. If the set $S$ has some interacting markers, linked to disease, of similar relative importance and some unlinked markers, then for any linked marker $l$ and any unlinked one $u$ we have:

$$E[H_{1\ldots l\ldots}] < E[H_{1\ldots i\ldots}] < E[H_{1\ldots w-1\ldots w+1\ldots}]$$

4. If the current set $S$ contains only markers linked to disease that are of similar relative importance and also have non-negligible interaction, then for any marker $l$ in $S$

$$E[H_{1\ldots l\ldots}] < E[H_{1\ldots i\ldots}]$$

Proof

1. The first part is easy; we have chosen the weights in section 2.1 such that when no marker is linked to disease, we have:

$$E[H_{1\ldots i-1\ldots i+1\ldots}] = E[H_{1\ldots i}]$$

2. From (3) in section 2.1 we can write:

$$E[H_{1\ldots i\ldots}] = \frac{w_{i-1}E[H_{1\ldots i\ldots}] + \ldots + E[H_{1\ldots i-1\ldots i+1\ldots}]}{w_{i-1}} = \frac{k w_{i-1}E[H_{1\ldots i\ldots}] + w_{i}E[\left(\theta_{i-1} - \theta_{0\ldots i-1\ldots}\right)^2]}{w_{i-1}}$$

$$= \frac{\left[ w_{i-1}E[H_{1\ldots i\ldots}] + k w_{i}E[\left(\theta_{i-1} - \theta_{0\ldots i-1\ldots}\right)^2]\right]}{w_{i-1}} + \frac{(k-1) w_{i-1}E[H_{1\ldots u-1\ldots u+1\ldots}]}{w_{i-1}}$$
It suffices to show that
\[ E[H_{2,k}] < \frac{w_k}{w_{k-1}} E[H_{2,k-1}] + k w_k \left( n_{-i}^k - n_{-0,0}^k \right)^2 \]
\[ < \frac{w_k}{w_{k-1}} E[H_{1,w-1,u+1,k}] \quad (*) \]

(we also use \( w_k < w_{k-1} \)). We prove the first inequality, namely
\[ E[H_{2,k}] < \frac{w_k}{w_{k-1}} E[H_{1,w-1,u+1,k}] + k w_k \left( n_{-i}^k - n_{-0,0}^k \right)^2 \]
Since markers 2, ..., \( k \) are not linked to disease and among themselves, one can easily show that \( E[H_{2,k}] = 2N \) where \( N \) is twice the number of ASPs. Therefore we need to show that
\[ 2N < \frac{w_k}{w_{k-1}} \cdot 2N + k w_k \left( n_{-i}^k - n_{-0,0}^k \right)^2 \]
\[ \implies \frac{N}{2^{k-1}} < k E \left[ \frac{n_{-i}^k - n_{-0,0}^k}{2} \right] \]

Now we have:
\[ E \left[ \frac{n_{-i}^k - n_{-0,0}^k}{2} \right] = E \left[ \frac{n_{-i}^k - n_{-0,0}^k}{2} \right] + \text{Var} \left[ \frac{n_{-i}^k - n_{-0,0}^k}{2} \right] \approx \]
\[ E \left[ \frac{n_{-i}^k - n_{-0,0}^k}{2} \right] + N \left( \frac{p_{-k}^k - p_{-0,0}^k}{2} \right) \approx E \left[ \frac{n_{-i}^k - n_{-0,0}^k}{2} \right] + \frac{N}{2^{k-1}} \]
where \( p_{-k}^k \) is the probability of the IBD sharing vector \( 1...1 \) at loci 1, ..., \( k \). Therefore we showed the first inequality.
For the second inequality in (\( \ast \)),
\[ \frac{w_k}{w_{k-1}} E[H_{2,k}] + k w_k \left( n_{-i}^k - n_{-0,0}^k \right)^2 < \frac{w_k}{w_{k-1}} E[H_{1,w-1,u+1,k}] \]
we proceed as follows. By definition in section 2.1 we have:
\[ H_{2,k} = \]
\[ \frac{w_k}{w_{k-1}} \cdot \frac{\sum_{i=1}^{k-1} \left( n_i^k - n_{-0,0}^k \right)^2}{\left( k-1 \right) \left( k-2 \right)} + \sum_{i=1}^{k} \left( n_{-i}^k - n_{-0,0}^k \right)^2 \]
\[ H_{1,w-1,u+1,k} = \]
\[ \frac{w_{k-1}}{w_k} \cdot \frac{\sum_{i=1}^{k-1} \left( n_i^k - n_{-0,0}^k \right)^2}{\left( k-1 \right) \left( k-2 \right)} + \sum_{i=1}^{k} \left( n_{-i}^k - n_{-0,0}^k \right)^2 \]

It suffices to show that
\[ E \left[ n_i^k - n_{-0,0}^k \right]^2 \geq \frac{1}{2} \left( k-1 \right) \left( k-2 \right) \] for \( k \geq 2 \)
(7)
Using (5) it is easy to prove that \( E \left[ n_i^k - n_{-0,0}^k \right]^2 = N \) for any unlinked marker and \( E \left[ n_i^k - n_{-0,0}^k \right]^2 = \frac{N}{2^{k-1}} \) for a linked marker. Also
\[ E \left[ n_{-i}^k - n_{-0,0}^k \right]^2 \approx \frac{1}{2^{k-1}} E \left[ n_i^k - n_{-0,0}^k \right] + \frac{N}{2^{k-1}} \]
where \( N \) is twice the number of ASPs. Hence we need to prove that:
\[ E \left[ n_i^k - n_{-0,0}^k \right] \left( 1 - \frac{k(k-1)}{4^{k-1}} \right) > \frac{k(k-1)N}{2^{k-1}} \]
If we let \( p_1 = p_k \) with \( r > 1 \) and since \( p_1 + p_0 = 1 \) we obtain:
\[ N^2 \left[ n_i^k - n_{-0,0}^k \right] \left( 1 - \frac{k(k-1)}{4^{k-1}} \right) > \frac{k(k-1)N}{2^{k-1}} \]
\[ \implies N \left( \frac{r-1}{1} \right) > \frac{1}{k(k-1)} \]

The latter inequality is true for \( N \) (twice the number of ASPs) large enough. For example when \( r = p_1/p_0 = 1.3 \) and \( k = 2 \), a sample of 60 ASPs is sufficient. Hence we have shown that (7) is true.
Now we can complete the proof of the second inequality in (\( \ast \)). If marker 1 is linked to disease and marker \( u \) is not linked, then we have:
\[ E \left[ n_i^k - n_{-0,0}^k \right] > E \left[ n_i^k - n_{-0,0}^k \right] + k(k-1)E \left[ n_{-i}^k - n_{-0,0}^k \right] \] from (7)
\[ E \left[ n_i^k - n_{-0,0}^k \right] > E \left[ n_i^k - n_{-0,0}^k \right] \] for any \( j \neq \{1,u\} \)
\[ E \left[ n_{-i}^k - n_{-0,0}^k \right] \]
Using the definitions of \( H_{2,k} \) and \( H_{1,w-1,u+1,k} \) in (6) and together with the inequalities above we obtain the second inequality in (\( \ast \)). This completes our proof.
3. We assume \( k \geq 4 \) (\( k = 2 \) and \( k = 3 \) can be proved using case-by-case computations). We assume the first \( t \) markers are linked to disease and the rest \( k-t \) are unlinked.
Let \( E[H_{1,i-1,i+1,k}] = A \) for any \( i \leq t \) and \( E[H_{1,i-1,i+1,k}] = B \) for any \( i > t \). Clearly, \( A < B \). We now use the approximation in section 2.1:
\[ E[H_{2,k}] = \]
\[ A + \frac{k-t}{k} \left( B - A \right) + E \left[ n_{-i}^k - n_{-0,0}^k \right] \]
\[ = A + \frac{k-t}{k} B + \frac{N}{2^{k-1}} \]
From this we have:
\[ E[H_{2,k}] \approx A + \frac{k-t}{k} \left( B - A \right) + \frac{N}{2^{k-1}} \]
and since \( B > A \) we obtain \( E[H_{2,k}] > A \)
Similarly:
\[ E[H_{2,k}] \approx B + \frac{t}{k} \left( A - B \right) + E \left[ n_{-i}^k - n_{-0,0}^k \right] \]
\[ \text{We show that:} \]
\[ B - A > \frac{t}{k} E \left[ n_{-i}^k - n_{-0,0}^k \right] \]
\[ \text{and therefore } E[H_{2,k}] < B \]
From (6) we can write:

$$B - A > \sum_{k=1}^{K} \frac{E[(n^1_k - n^0_k)^2]}{k} - E[(n^1_k - n^0_k)^2] \quad \text{where} \quad u > t$$

$$\approx \frac{E[(n^1_k - n^0_k)^2 + N - N]}{k} = \frac{N^2(p^1_k - p^0_k)^2}{k}$$

(9)

With (8) and (9), we need to show:

$$\frac{N^2(p^1_k - p^0_k)^2}{k} > \frac{k}{t} \left[ E\left[(n^1_k - n^0_k)^2\right] \right]$$

(10)

Now:

$$E\left[(n^1_k - n^0_k)^2\right] = E\left[p^1_k - n^0_k\right] + \text{Var}\left[p^1_k - n^0_k\right]$$

$$\approx E\left[p^1_k - n^0_k\right] + N\left(p^1_k + p^0_k\right)$$

$$\approx N^2(p^1_k - p^0_k)^2 + N\left(p^1_k + p^0_k\right)$$

$$\approx N^2(\frac{1}{k^2} + \frac{1}{k} - \frac{1}{t^2}) + \frac{k}{t} \frac{N}{k} (p^1_k + p^0_k)$$

where we used the fact that the last $k-t$ markers are not linked to disease and among themselves (hence $p^1_k - 1/2^{k-t}p^1_k$). From this and (10) follows that we need to prove that:

$$\frac{N^2(p^1_k - p^0_k)^2}{k} > \frac{k}{t} \frac{N^2(p^1_k - p^0_k)^2}{4^{k-t}} + \frac{k}{t} \frac{N}{k} (p^1_k + p^0_k)$$

i) If $t$ is small compared to $k$ (i.e., $k-t$ is large) then since $(p^1_k - p^0_k)^2 < (p^1_k - p^0_k)^2$ and $(p^1_k + p^0_k) < 1$ it is sufficient to show that:

$$N(p^1_k - p^0_k)^2 > \frac{1}{k} \frac{1}{k} \frac{1}{t} \frac{1}{4^{k-t}}$$

This inequality is similar to inequality (**) shown at point 2. of the lemma for the case $t = 1$. For $N$ large enough and when $t$ is small compared with $k$, it is true.

ii) If $t$ is comparable to $k$ and since $k \geq 4$ (from our assumption), then $p^1_k = 1/2^{k-t}$ tend to be much smaller than $p^1_k - p^0_k$ (say conservatively, $p^1_k > 1/2^{k-t}$) and also

$$\frac{N}{k} \geq \frac{1}{k} \frac{1}{k} \frac{1}{t} \frac{1}{4^{k-t}}$$

For example, if $t = k-1$ and $p^1_k - p^0_k = 0.1$ then the inequality above is

$$\frac{N}{k} \geq \frac{1}{16} \frac{1}{t} \frac{1}{4^{k-t}}$$

which is true ($k \leq 10$). For $t$ smaller than $k-1$ the inequality is even sharper.

This concludes our proof for $t$ (the number of markers linked to disease) between 2 and $k-1$. Next we show the case $t = k$.

4. We prove the case $k = 4$. The cases $k = 2$ and $k = 3$ can be verified easily through direct case-by-case computations. We use the approximation in section 2.1 and since the interacting disease loci have similar importance we obtain:

$$E[H^{\epsilon}_k] \approx E\left[H^{\epsilon}_{k-1}\right] + \cdots + E\left[H^{\epsilon}_{k-1}\right] + E\left[(n^1_{k-1} - n^0_{k-1})^2\right]$$

$$= E\left[H^{\epsilon}_{k-1}\right] + E\left[(n^1_{k-1} - n^0_{k-1})^2\right]$$

Since all $k$ markers in the current set are assumed to be linked to disease and to interact together in a non-negligible fashion, $E[(n^1_{k-1} - n^0_{k-1})^2]$ is large enough to guarantee the inequality:

$$E[H^{\epsilon}_{k-1}l_{k-1}] < E[H^{\epsilon}_k] \text{ for } k \geq 4$$

### Appendix B

#### Choice of $B$

A heuristic approach to estimating $B$ was given in Lo and Zheng [6]. The derivation of $B$ for the proposed method is very similar and we present the main ideas below.

Suppose $M$ is a marker linked to disease in the original set of polymorphisms. Let $p_1$ be the probability that a marker linked to disease is selected and returned in any single repetition (out of $B$) of the selection-reduction process; $p_0$ is the same probability but for markers not linked to disease. Assume $p_1 = p_0$ with $r > 1$. Let $X$ be the observed return count for marker $M$. Then

$$X \sim N\left(\frac{Bp_1}{r}, \frac{Bp_1(1-p_0)}{r} \right)$$

In order to clearly separate the markers linked to disease from the ones not linked, we require:

$$P(X \geq Bp_0 + 3.1\sqrt{Bp_0(1-p_0)}) \approx 99\%$$

After some algebra, this can be written equivalently as:

$$B > \frac{3.1 + 2.33\sqrt{r}}{r-1} \left(1 - \frac{p_0}{p_0} \right)$$

We can estimate $p_0$, the probability for a marker not linked to disease to be selected and returned in a single repetition, as follows:

$$p_0 = P(\text{selected and unlinked}) \cdot P(\text{selected}) - P(\text{unlinked})$$

$$\approx \frac{1}{k} \frac{1}{k} \frac{1}{k} \frac{1}{k} \frac{1}{k} \frac{n}{k} \approx \frac{1}{n}$$

where $n$ is the total number of markers and $k$ is a small number of markers (say 10) selected to be evaluated; we assume conservatively that the probability that a selected marker, not linked to disease, is returned is 1/k.

Therefore we obtain

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

$r$ can be written as:

$$r = \frac{P_1}{p_0} \approx \frac{P(\text{selected and linked})}{P(\text{selected and unlinked})} \approx k(1-\epsilon)$$
where \( 1 - \varepsilon \) is an estimate for the probability that a linked marker, once selected, is returned. If conservatively we take \( r = 2 \) we obtain \( B = 41(n - 1) \).

**Appendix C**  
**Efron’s Approach to Separating the Important from the Unimportant Markers**  
We also applied Efron’s method [5] for separating the set of markers into two categories: important vs unimportant. In figure 8 we give the results. On the left-hand side, the histogram of the return counts together with the fitted empirical null density \( f_0(z) \) and the mixture density \( f(z) \) are depicted. On the right-hand side, the LocFdr plot is added to the histogram (scaled up by a factor of 50). A return count of 242 corresponds to a LocFdr of 1%.

**Appendix D**  
**Extension of the Multilocus Linkage Method**  
We give a natural extension of the multilocus linkage method so that mild linkage disequilibrium (LD) levels between disease loci and marker loci can be used in addition to linkage to obtain even greater increases in power over single-locus linkage methods. The extension is based on combining the multilocus linkage method with a similar association method (the BHTA algorithm, Lo and Zheng [6]). They are both based on the screening procedure in section 2.2. In what follows we denote by \( \Delta L_i \) (called \( \Delta_i \) in the main text) and \( \Delta LD_i \) (defined in Lo and Zheng [6]) the change in linkage and association information respectively when removing marker \( i \) from the current set.

- **Step 0**: Repeat steps 1–4 \( B \) times.
- **Step 1**: Start by choosing a set of \( k \approx 10 \) markers at random from the available list of markers.
- **Step 2**: At each step compute for each marker in the current set the resulting change in both the linkage and association measure respectively when that marker is removed. For marker \( i \): \( \Delta L_i = L_{1,\ldots,i-1,i+1,k} - L_{1,\ldots,k} \); \( \Delta LD_i = LD_{1,\ldots,i-1,i+1,k} - LD_{1,\ldots,k} \).
- **Step 3**: Remove the marker \( i \) (if any) such that both \( \Delta L_i > 0 \) and \( \Delta LD_i > 0 \) (i.e. both the linkage measure and the association measure deem marker \( i \) unimportant) and that has the largest \( \Delta L_i + \Delta LD_i \).
Table 1. Selected important markers

<table>
<thead>
<tr>
<th>CHR</th>
<th>Selected marker</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D1S1612</td>
<td>IBD7</td>
</tr>
<tr>
<td></td>
<td>D1S534</td>
<td>1q21</td>
</tr>
<tr>
<td></td>
<td>D1S1595</td>
<td>1q21</td>
</tr>
<tr>
<td></td>
<td>D1S1677</td>
<td>1q23</td>
</tr>
<tr>
<td>2</td>
<td>D2S1394</td>
<td>2p11</td>
</tr>
<tr>
<td></td>
<td>D2S1790</td>
<td>2p11</td>
</tr>
<tr>
<td></td>
<td>D1S1649</td>
<td>2q32</td>
</tr>
<tr>
<td>3</td>
<td>D3S1285</td>
<td>3p</td>
</tr>
<tr>
<td>4</td>
<td>D4S2394</td>
<td>4q23–28</td>
</tr>
<tr>
<td>5</td>
<td>D5S500 other 18 from the region</td>
<td>IBD5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBD5</td>
</tr>
<tr>
<td>6</td>
<td>DRB1</td>
<td>IBD3</td>
</tr>
<tr>
<td></td>
<td>DQB1</td>
<td>IBD3</td>
</tr>
<tr>
<td></td>
<td>D6S1017</td>
<td>IBD3</td>
</tr>
<tr>
<td>7</td>
<td>GATA31A10</td>
<td>7p13</td>
</tr>
<tr>
<td>12</td>
<td>D12S372</td>
<td>12p</td>
</tr>
<tr>
<td>16</td>
<td>D16S2619</td>
<td>IBD8</td>
</tr>
<tr>
<td></td>
<td>D16S753</td>
<td>IBD1</td>
</tr>
<tr>
<td>19</td>
<td>D19S591</td>
<td>IBD6</td>
</tr>
<tr>
<td></td>
<td>GATA21G05</td>
<td>IBD6</td>
</tr>
<tr>
<td></td>
<td>D19S714</td>
<td>IBD6</td>
</tr>
<tr>
<td>21</td>
<td>D21S1809</td>
<td>21q22.3</td>
</tr>
</tbody>
</table>

- Step 4: Do Steps 2–3 until either all the markers in the current set are important (for each remaining marker i not both \( \Delta L_D \) and \( \Delta L_I \) are positive) or only one marker remains. We return marker \( i \) \( R_i \) times, depending on the linkage and the association evidence as follows: \( R_i = \frac{1}{H_{LD_i}} + \frac{1}{H_{LI_i}} \) where \( 1/H_{LD_i} \) is the indicator random variable for the event \( \Delta L_D < 0 \); \( 1/H_{LI_i} \) is defined similarly.
- Step 5: We compute for each marker a final return count denoting the total number of times it was returned in Step 4. Based on these counts we separate the markers into two classes: the unimportant (unlinked) markers and the important (linked AND/OR associated) ones.

This simple procedure guarantees that when evaluating a marker we consider both the marginal information, as well as the interaction information contained in a dataset. Also it takes into account two pieces of information, usually treated separately: linkage information and linkage disequilibrium information. We are currently preparing a separate paper with the details of this combined procedure.

Appendix E

Results on Real Data

Table 1 lists the markers we claim important together with their chromosomal position (InfBD dataset).

References