Statistical Applications in Genetics and Molecular Biology

Manuscript 1718

Improving Pedigree-based Linkage Analysis by Estimating Coancestry Among Families

Chris Glazner, *University of Washington*
Elizabeth Alison Thompson, *University of Washington*

Improving Pedigree-based Linkage Analysis by Estimating Coancestry Among Families

Chris Glazner and Elizabeth Alison Thompson

Abstract

We present a method for improving the power of linkage analysis by detecting chromosome segments shared identical by descent (IBD) by individuals not known to be related. Existing Markov chain Monte Carlo methods sample descent patterns on pedigrees conditional on observed marker data. These patterns can be stored as IBD graphs, which express shared ancestry only, rather than specific family relationships. A model for IBD between unrelated individuals allows the estimation of coancestry between individuals in different pedigrees. IBD graphs on separate pedigrees can then be combined using these estimates. We report results from analyses of three sets of simulated marker data on two different pedigrees. We show that when families share a gene for a trait due to shared ancestry on the order of tens of generations, our method can detect a linkage signal when independent analyses of the families do not.

KEYWORDS: linkage, pedigrees, gene coancestry, IBD

Author Notes: The research for this paper was supported in part by NIH grants R37 GM046255 and T32 GM081062. The authors are grateful to an anonymous referee for helpful and constructive comments.

1 Introduction

Identity by descent (IBD) is the sharing of DNA copied through successive meioses from one chromosome in a recent common ancestor. IBD is always relative. In a pedigree, it is defined relative to the founders of the pedigree. In a population, it must be defined relative to some time point or ancestral population. In this paper, we follow Browning and Browning (2010) in defining IBD on the order of tens of generations. At this time-scale, the lengths of shared inherited IBD segments are on the order of several million base-pairs, and thus an order of magnitude longer than the extent of linkage disequilibrium, which is itself a reflection of more remote coancestry. Whether in pedigrees or in populations, gene mapping approaches are based on the dependence in descent of DNA at nearby locations in the genome. Hence the coancestry of segments of DNA drives all gene mapping methods.

In this paper we extend gene mapping models on defined pedigrees to incorporate coancestry on an intermediate scale. If families are selected for a rare trait or from a small population, they are likely to share ancestors in this time span. Capturing these relationships in a pedigree is impractical, both due to the unknowability or unreliability of ancestral pedigrees with multiple generations of missing genetic data and to the computational difficulty in analyzing them. On smaller 3- or 4-generation pedigrees, with few individuals unobserved, genetic marker data validate the pedigree structures (see Sun et al., 2002, for example) However, linkage analyses using such pedigrees lack resolution, due to the limited number of meioses (Boehnke, 1994). Thus, we propose the use of a population-level model of coancestry to supplement analysis of data on smaller pedigrees.

We begin by sampling descent patterns on several pedigrees, conditional on genetic marker data, as in a Monte Carlo linkage analysis (Thompson, 2007). These pedigree-based descent patterns are summarized in the form of a version of the descent graph of Sobel and Lange (1996), the IBD graph (Thompson, 2011). These sampled descent patterns are then connected using output from a population model for IBD in pairs of individuals who are not known *a priori* to be related. We calculate likelihood scores using these augmented samples, which incorporate relationships between pedigrees as well as those within. The augmentation is possible because the IBD graph provides a common data structure for expressing inferences from both pedigree and population models

Simulation studies demonstrate that this method allows linkage signals to be recovered from pedigrees which do not show linkage when analyzed independently. Additionally, broad weak signals can be significantly narrowed, and false indications of signals can be eliminated. More generally, our method shows that the IBD graph is a flexible and natural tool for modeling coancestry.

In Section 2, we briefly describe our approach to the Monte Carlo estimation of linkage lod scores and the use of the IBD graph to summarize marker-based inferences of descent patterns. Our methods for inferring population-level IBD and for combining pedigree-based and population based IBD inferences are described in Section 3. The simulation of examples and the results of simulation studies appear in Section 4.

2 Linkage analysis with IBD graphs

We observe marker data Y_M and trait data Y_T on two or more pedigrees. Our goal is to compute the likelihood ratio between the two genetic models Γ and Γ_0, where Γ hypothesizes a trait location on the chromosome of the markers, and Γ_0 assumes that the trait and marker data are independently distributed on the pedigree. Both models incorporate the pedigree structure and assume the same marginal models for Y_M and Y_T, including the genetic marker map, allele frequencies, and other parameters. The log of the likelihood-ratio is the lod score:

$$\log_{10}\frac{\Pr(Y_T,Y_M;\Gamma)}{\Pr(Y_T,Y_M;\Gamma_0)}=\log_{10}\frac{\Pr(Y_T,Y_M;\Gamma)}{\Pr(Y_T;\Gamma_T)\Pr(Y_M;\Gamma_M)}=\log_{10}\frac{\Pr(Y_T\mid Y_M;\Gamma)}{\Pr(Y_T;\Gamma_T)},$$

where we can factor the denominator because of the independence assumed in Γ_0. On large pedigrees, the numerator cannot be calculated exactly. To approximate the likelihood under Γ, we express it as an expectation over all possible inheritance patterns S on the pedigree:

$$\Pr(Y_T\mid Y_M;\Gamma)=\sum_S \Pr(Y_T\mid S;\Gamma_{T+})\Pr(S\mid Y_M;\Gamma_M)=\mathrm{E}_{\mathrm{S}|\mathrm{Y_M};\Gamma_\mathrm{M}}\Pr(\mathrm{Y_T}\mid \mathrm{S};\Gamma_{\mathrm{T}+}).$$

Here Γ_{T+} is the marginal trait model, Γ_T, augmented by the trait location hypothesized in Γ. (A realization of S may be expressed as a matrix specifying whether the maternal or paternal allele segregated at each locus in each meiosis of the pedigree.) By simulating from the conditional distribution of S given Y_M and computing the likelihood of the trait data as a function of S, we obtain a Monte Carlo estimate of $\Pr(Y_T|Y_M;\Gamma)$ (Lange and Sobel, 1991). Once a sample from S has been realized, it can be used for different traits and trait models without resimulating. Sampling is performed via Markov chain Monte Carlo, and recent methods allow sampling on large pedigrees with 2 to 3 markers per cM (Tong and Thompson, 2008).

Equivalent patterns of descent S can be represented by descent graphs, which were introduced by Sobel and Lange (1996). Each founder allele at a single locus is assigned a founder genome label (FGL). Each node in the graph represents an FGL. Both alleles carried by an individual can be traced back to a founder, so an individual is represented by an edge connecting the two FGLs he or she carries. Figure

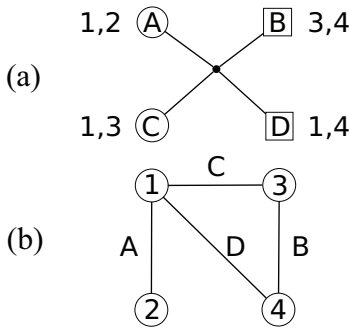


Figure 1: (a) A simple pedigree labelled with FGLs. (b) An IBD or descent graph representing one possible descent pattern on the pedigree at a single locus. The mother *A* passes her maternal allele to both children, and the father *B* gives a different allele to each child.

1b shows a simple descent graph for a nuclear family. Each parent is a founder and carries his or her two FGLs. One child receives FGLs 1 and 3, while the other receives 1 and 4.

We refer to a set of descent graphs along the chromosome as an IBD graph. IBD graphs additionally have the property that the nodes are unlabelled, so they condense distinct but equivalent descent patterns into equivalence classes. A simple example is that if the children in Figure 1 had both inherited their mother's paternal rather than maternal allele, the resulting IBD graph would be the same even though the descent patterns are different. A major advantage the IBD graph has over a matrix representation of S is that it encodes coancestry without reference to the family structure. Once the IBD graphs have been sampled, both the pedigree and the marker data can be discarded; the IBD graphs are sufficient for calculating the lod score. The IBD graphs can be stored or shared with minimal privacy concerns, and efficient likelihood computations for various trait models can be performed using only the graphs and the trait data. IBD graphs can also be stored in less space than an inheritance matrix S, because only points where the FGLs change must be recorded. More details on the useful properties of IBD graphs can be found in Thompson (2011).

Most importantly for the approach in this paper, an IBD graph, in contrast to a pedigree, can express coancestry on any desired scale. The set of ancestors which determine the IBD graph on the individuals can be as recent or as ancient as desired. This flexibility in scope allows us to augment the IBD graphs obtained from pedigrees by increasing the scope of the coancestry to encompass older and hence more distant relationships.

3 Methods

Typically, unconnected pedigrees are treated as independent and their lod scores are simply added together. We propose instead to perform joint inference by estimating the IBD between pedigrees. Our procedure involves three steps. First, we sample IBD graphs on two or more pedigrees. Next, we estimate the IBD between all possible pairs of individuals in the sample. Finally, the pairwise IBD inferences are integrated into each of the sampled graphs, creating (possibly) connected graphs which can be used as if they were the output of straightforward MCMC sampling.

3.1 Population IBD detection

The IBD between a pair of individuals can be modeled using a hidden Markov model (HMM). The pattern of IBD along the chromosome follows a continuous time Markov process as described in Thompson (2008, 2009). Four phased haplotypes can be in 15 possible IBD states, which collapse to 9 states if we have unphased genotypes. (The number of states for n haplotypes grows large very quickly with n, so this approach is only feasible for small numbers of individuals.) Two parameters determine the rate matrix of the transitions between states: the average relatedness in the population and the expected length of IBD segments. Simulation studies indicate that estimation is not sensitive to changes in either value within a reasonable range.

The symbols emitted by the model are the observed haplotypes or genotypes of the pair of individuals at genetic markers along the chromsome. The markers are assumed to be diallelic SNPs with known population frequencies. Alleles which are IBD must be observed to be of the same allelic type unless there has been a mutation or genotyping error, and each allele or set of IBD alleles is modeled as a random draw from the population. An emission probability is calculated from the hidden state and the genotype, which determine the number of independent draws, and the allele frequencies. Table 1 shows the emission probabilities for genotypes given underlying IBD states; each entry is the product of one or more binomial factors, depending on the number of independently varying sets of alleles. Mutation can be disregarded on the time scales of interest here. To incorporate genotyping error, each state has a small probability (0.01) of emitting four independent allelic types. The actual emission distribution of a state (when using genotypic data) is therefore a mixture of the corresponding row in Table 1 and the final row of table.

The forward-backward algorithm (Rabiner, 1989) gives the marginal probability at each locus that a pair of individuals is in a particular state given all of the observed marker data. When evaluating the model's performance, we declared the

	Observed genotypes								
IBD state	11,11	11,12	11,22	12,11	12,12	12,22	22,11	22,12	22,22
1 1 1	p	–	–	–	–	–	–	–	q
1 1 0	p^2	–	pq	–	–	–	pq	–	q^2
1 0 1	p^2	pq	–	–	–	–	–	pq	q^2
1 0 0	p^3	$2p^2q$	pq^2	–	–	–	p^2q	$2pq^2$	q^3
0 1 1	p^2	–	–	pq	–	pq	–	–	q^2
0 1 0	p^3	–	p^2q	$2p^2q$	–	$2pq^2$	pq^2	–	q^3
0 0 2	p^2	–	–	–	$2pq$	–	–	–	q^2
0 0 1	p^3	p^2q	–	p^2q	pq	pq^2	–	pq^2	q^3
0 0 0	p^4	$2p^3q$	p^2q^2	$2p^3q$	$4p^2q^2$	$2pq^3$	p^2q^2	$2pq^3$	q^4

Table 1: Emission probabilities for genotypes given IBD states, before allowing for genotyping error. The first two digits of the IBD state indicate whether each individual has two IBD alleles (1) or not (0), and the third gives the number of alleles shared IBD by the two individuals (see notation in Thompson, 2008). The values p and $q = (1-p)$ are the population frequencies of alleles 1 and 2, respectively.

modal probability at each locus to be the estimate state, unless the probability was less than 0.9. In that case, no estimate was made. We simulated a large population of individuals over 200 generations of descent and tested the model performance (Glazner et al., 2010) and compared the true IBD states to the estimated states. The HMM analysis detected most segments longer than 1 cM, indicating that it provides a useful framework for inferring coancestry on the scale of hundreds of years.

Our model does not account for linkage disequilibrium (LD) among markers. In principle, given a model for LD in the population, it would be possible to created a joint model for LD and IBD. Not only would the model be much more complex, but estimation of population LD requires large samples, so we have not pursued this approach. Our approach may be contrasted with that of Browning and Browning (2010), who fit an LD model using large population samples but adopt a very simple 2-state IBD model. Since LD is the result of ancestral haplotypes that have not been broken down by recombination, it can be considered ancient IBD. In practice, we find that our model does detect more IBD in populations with more LD. A detailed study of the impact of LD on inference of IBD in population samples under our HMM IBD model is given by Brown et al. (2011).

3.2 Merging IBD graphs

Once pairwise IBD between individuals has been estimated, the estimates must be combined to produce an IBD model for the entire collection of families. This task is complicated by the ambiguity in the IBD state inference when the data are unphased. We may infer that two individuals share an allele IBD, but we cannot distinguish between the two possible choices of maternal or paternal alleles for each individual.

Suppose we have two families, with a set of stored IBD graph realizations for each family. Merging is performed on a single pair of realizations, one from each family. The nodes in the two IBD graphs at a given locus compose two disjoint sets of FGLs. Combining the two graphs involves pairing nodes between them. A particular way of merging the graphs can therefore be represented as a set of Boolean variables, one for each pair of nodes. The inferred IBD states can then be interpreted as Boolean expressions on these pairing variables. For example, suppose an individual carries FGLs 1 and 2 and is inferred to share one allele IBD with an individual carrying FGLs 3 and 4. The inference can be expressed as

$$p_{1,3} \vee p_{1,4} \vee p_{2,3} \vee p_{2,4}$$

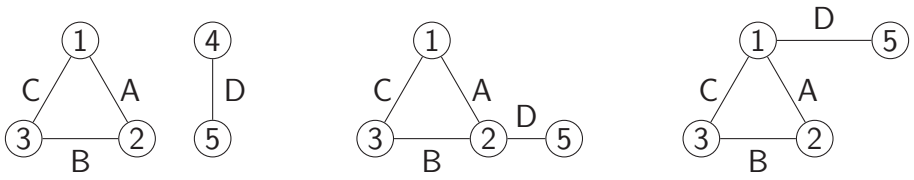
where $p_{x,y}$ denotes pairing between FGLs x and y. Additionally, further expressions must be included to ensure that exactly one of the variables is true. If the same individuals are inferred to share no alleles IBD, the resulting expression is

$$\neg p_{1,3} \wedge \neg p_{1,4} \wedge \neg p_{2,3} \wedge \neg p_{2,4}.$$

The inferences for each pair of individuals across the two families, along with the two input graphs, define a Boolean formula whose solutions are pairing arrangements compatible with the inferences. Determining the existence of solutions to a Boolean formula is known as the Boolean satisfiability problem, or SAT (Knuth, 2008). While the problem is computationally difficult in general, the instances encountered here are quickly solved by existing software.

To merge two graphs at a locus, we first run the HMM on each pair of individuals across the two graphs. We then select the modal IBD state for each pair and rank the pairs according to the modal probability. This step has no simple statistical interpretation, but it roughly orders the state inferences according to our confidence in them. Beginning with an empty Boolean formula, we successively add the expression induced by each pair and at each step test the result for consistency using the MINISAT program (Eén and Sörensson, 2003). If there is a solution to the formula, the expression is included in the formula; otherwise, the inference for that pair is discarded. We proceed until we have attempted to include each pair in the

Individuals	Alleles IBD	Modal state probability	Boolean statement	Included in solution?
Scenario 1				
A D	1	0.99	Exactly one of $p_{1,4}, p_{1,5}, p_{2,4}, p_{2,5}$	Yes
B D	1	0.95	" $p_{2,4}, p_{2,5}, p_{3,4}, p_{3,5}$	Yes
C D	1	0.94	" $p_{1,4}, p_{1,5}, p_{3,4}, p_{3,5}$	No
Scenario 2				
A D	1	0.99	Exactly one of $p_{1,4}, p_{1,5}, p_{2,4}, p_{2,5}$	Yes
C D	1	0.95	" $p_{1,4}, p_{1,5}, p_{3,4}, p_{3,5}$	Yes
B D	1	0.94	" $p_{2,4}, p_{2,5}, p_{3,4}, p_{3,5}$	No



Before merging Scenario 1 Scenario 2

Figure 2: A example demonstrating the sensitivity of the merging procedure to small differences in HMM output. Individuals B and C have nearly the same marginal probability of sharing one allele IBD with individual D, but they cannot both do so if A also shares with D. The priority in which the pairings are incorporated can affect the resulting IBD graph and the lod score at that locus.

formula, and the result is a formula with either one or many solutions. If the solution is not unique, an arbitrary solution is chosen, since all solutions are compatible with the restrictions induced by the HMM inference.

Figure 2 shows a simple example of the merging process at a single locus. Individuals A, B, and C are in one family and form the IBD graph shown in the figure; individual D is in another family. The numbered nodes represent the two FGLs carried by each individual. Each pairing of A, B, or C with D produces a set of marginal probabilities for the IBD states which the two individuals may be in. The most likely state for each pair is chosen; in this example, the selected state in each case is one allele shared IBD. None of the individuals carries two IBD alleles at this locus. As shown in the figure, the resulting merged IBD graph depends on the order in which the pairs are considered, since at most two of A, B, and C can share exactly one allele IBD with D. The two scenarios differ only slightly in the modal probabilities of pairings B-D and C-D, but since the ranking is different, different graphs are produced. (In the two scenarios, FGL 4 is merged with FGL 1

or 2; merging FGL 5 would also produce a solution, which would be topologically equivalent and therefore produce the same lod score.) Suppose that individuals A, B, and D all have high trait values, and individual C has a low value. The graph in scenario 1 would then provide evidence for trait linkage, with the high-value individuals all sharing a potential disease allele, FGL 2. The scenario 2 graph shows less of a trait association and would have a lower lod score. This example illustrates that, because IBD states are estimated pairwise rather than jointly, the algorithm is sometimes presented with incompatible inferences. Ranking the inferences and discarding some of them produces valid solutions, but the results are sensitive to the assigned ranks.

The merging algorithm imposes additional constraints on the set of solutions to ensure that no node in either graph pairs with more than one in the other graph. These constraints are included under the assumption that different nodes in the same graph are not IBD. To make this assumption reasonable, before merging two graphs we run the algorithm within each graph to condense any IBD among the founders. The procedure is similar to merging between graphs, except that more than two FGLs may be found to be IBD and merged; in the between-family setting, a node can only be merged with a single node in the other family, so at most two nodes are ever merged. Within a family, some Boolean solutions may not be transitive; for example, $p_{1,2}$ and $p_{1,3}$ may be true, while $p_{2,3}$ is false. Constraints are imposed which require solutions to be transitive. Without the constraints, $p_{1,2}$ and $p_{1,3}$ would be sufficient for FGLs 1,2, and 3 to be merged, even though the algorithm deduced that 2 and 3 are not IBD. The constraints allow a solution that includes $p_{1,2}$ or $p_{1,3}$, but not both.

This process is performed at a set of specified loci along the chromosome. One choice of merging loci is the set of of IBD graph change points that occur along the chromosome; this ensures that merging is performed on every distinct descent graph that appears along the chromosome. It may be useful to merge at more closely spaced markers, because while the two within-family IBD graphs are constant between change points, the HMM output may model changing relationships on the intervening segments.

An equal number of MCMC realizations are created for each family, and each graph is merged with a graph from the other family. The algorithm produces one IBD graph for every input pair. When more than two families are being analyzed, the merging procedure is iterated: the graphs created by merging the first two families are merged with a third set of graphs, these graphs are merged with a fourth set of graphs, and so on.

4 Examples

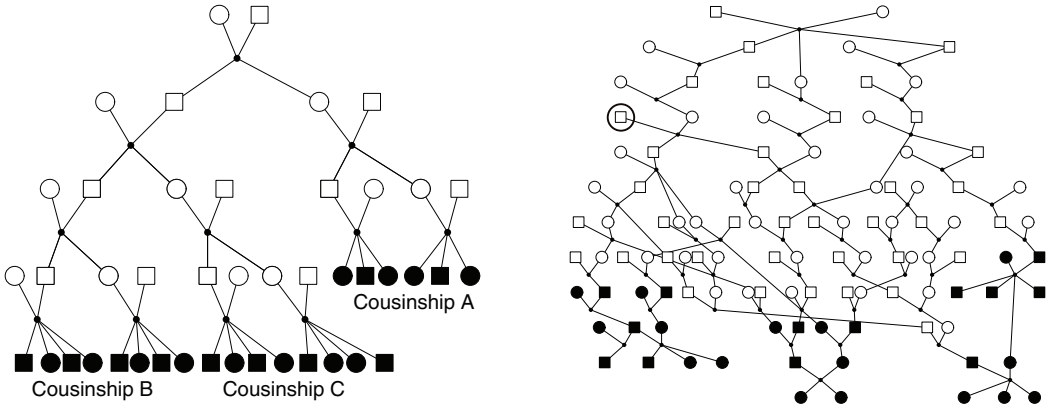
4.1 Haplotype simulation

Assessing the method required simulated marker data for which the true descent pattern was known. The data were created by simulating inheritance on extended pedigrees connecting several families. Haplotypes were then assigned to the pedigree founders. founders. For the case of dense markers, as in our second example below, we require haplotypes with realistic patterns of LD. To obtain a sufficient number of such haplotypes while avoiding the direct use of confidential data, haplotypes were simulated based on patterns of LD in real data.

The input data were X chromosome genotypes from males enrolled in the Framingham Heart Study (Cupples et al., 2009). Since males carry only a single X chromosome, the genotypes did not require a phasing step to produce haplotypes. The markers were from an Affymetrix 500K SNP chip. After filtering out markers with low minor allele frequency and a segment around the relatively uninformative centromere, 6,913 markers over 140 cM were used. The individuals were filtered to eliminate known close relationships in order to simulate an outbred founder population, leaving 1,917 chromosomes. Population allele frequencies were estimated from this sample.

The LD in the sample was modeled using the BEAGLE program (Browning and Browning, 2007). This software uses a large number of genotypes to construct a variable length Markov chain model representing the haplotype clusters in the population. At each marker, the model contains counts of the number of haplotypes moving into and out of the various clusters. Simulation of a haplotype consists in traversing the model from start to finish, with cluster transition probabilities according to the counts in each cluster. This procedure produces haplotypes with approximately the same LD pattern as the sample.

Both the Framingham sample and haplotypes simulated from the fitted BEAGLE model showed higher levels of LD than would be expected in a large outbred population, as determined by pairwise R^2 values. This result accords with the small geographic area from which the individuals were sampled. To simulate a larger population, the model traversal procedure was modified to include at each marker a possibility of jumping to a random haplotype cluster. By breaking dependence on path history, a jump simulates recombination within the population and reduces the overall level of LD. The jump probability used was 0.2, limiting the simulated LD to short distances (five markers on average). We have made available the R script, *beaglesim*, that implements our simulation approach.



(a) 44-member pedigree (b) Iceland pedigree

Figure 3: Pedigrees used in simulated examples. Inference was performed as if only the connected sets of shaded individuals were known *a priori* to be related. The circled founder in the Iceland pedigree is the origin of the disease gene.

4.2 Close relatives

The first simulated dataset examined was on a pedigree small enough that existing MCMC methods (Tong and Thompson, 2008) could be used on the entire pedigree and compared with our new approach. The 44-member pedigree shown in Figure 3 was used to simulate gene descent on a single 100 cM chromosome. The disease locus was close to the center of the chromosome, and the descent pattern and genotypes assumed at this locus were taken from a previous study (Wijsman et al., 2006) in which they provided a good linkage signal. A quantitative trait was simulated based on the genotypes at the diallelic disease locus; not all copies of the "disease allele" in observed individuals had a single origin. The genotype means and variances were chosen to allow detection of the linkage signal. Marker data on the rest of the chromosome were simulated conditional on the disease locus descent pattern; there were 201 SNP markers at 0.5 cM intervals. Each SNP had minor allele frequency 0.3. Only the 22 final members of the pedigree were assumed observed for trait and marker data.

For our current example, the pedigree was broken into three cousinships, labeled A, B, and C, as depicted in Figure 3. This breaking represents a situation in which the investigator does not know the true family relationships connecting the cousinships. IBD graph realizations on both the single large pedigree and the three cousinships were sampled; the sampler was run for 3×10^4 MCMC scans, with output every 30 scans. Thus, 1,000 IBD graphs were output for each cousinship.

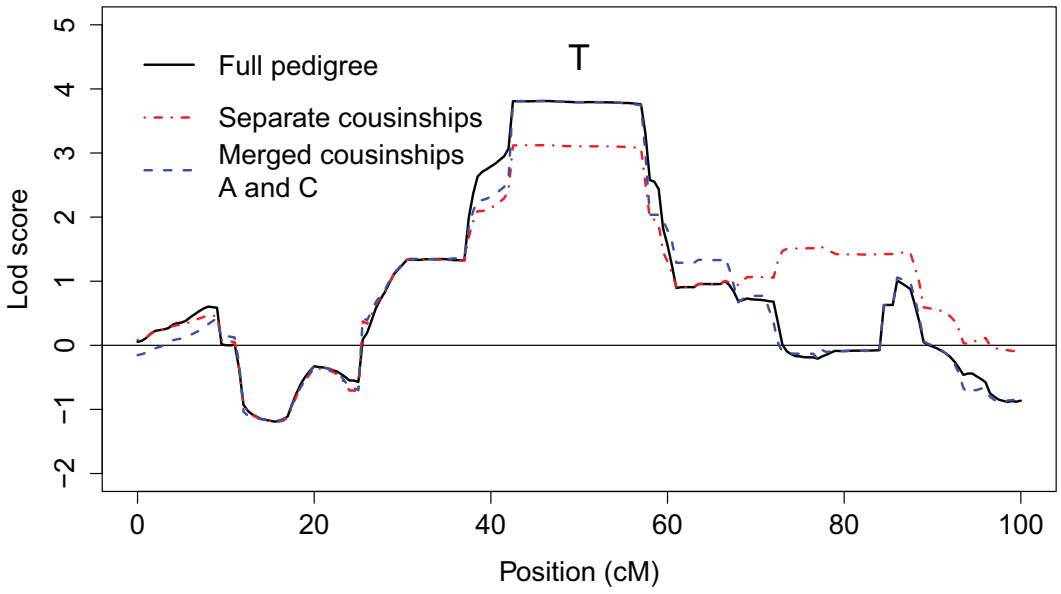


Figure 4: Lod scores on the simulated small dataset using the entire pedigree, separate cousinships, and cousinships after merging cousinships A and B. The trait locus is marked by the "T".

The merging algorithm was then used to connect cousinships A and C, since this pair of families had the most IBD sharing. Cousinship B, although more closely related to cousinship A, by chance received its "disease allele" from a different founder origin. Figure 4 shows the lod curves obtained from analysis of the full pedigree, the cousinships treated independently, and the cousinships after merging A and C. Relative to the true lod score on the full pedigree, independent analysis of the cousinships shows loss of signal in the neighborhood of the trait locus. Additionally, there was an apparent weak signal at 75 to 85 cM. The merged lod score closely tracks the score for the full pedigree, indicating that the hidden family information was largely recovered using our method. The result shows that such information can not only strengthen true signals, but can also eliminate other signals that are not supported by the joint data on the smaller pedigrees.

4.3 A large pedigree

A second, larger pedigree was used to test the ability of the method to detect shared descent among more distant relatives using denser SNP marker data. The structure

of the large pedigree in Figure 3 is based on one provided by Professor J. H. Edwards in 1995 (Thompson, 2000). It was collected in Iceland for a genetic study and spans twelves generations. On a pedigree such as this, direct analysis on the full pedigree is intractable. Due to the multiple generations of unsampled individuals, MCMC sampling of gene descent is unreliable (Thompson, 2000). Additionally, such a pedigree is likely to contain errors in the recorded ancestry, and there are ascertainment biases in limiting analysis to descent from a single founder couple 12 generations ago. In other populations, we would not normally even have a record of the entire pedigree, but have sampled only families not known to be related. This is the scenario we assume in our example; the three 3-generation families at the bottom of the full pedigree constitute our data.

The quantitative disease trait was simulated as in the small example, conditional on genotypes at the diallelic trait locus. In this case, the disease allele has a single origin, entering with a founder in the fourth generation and segregating to some members of each of the three descendant families. Again, descent across the chromosome was simulated conditional on descent at the trait locus; in this case, the trait locus was at position 105 cM in a 200 cM chromosome. Founder haplotypes simulated as described in Section 4.1 were assigned in accordance with the descent pattern. To expand our original 140 cM chromosome, we re-simulated markers from the first 60 cM, ending with 10,188 markers over the 200 cM chromosome.

The higher density of these markers required that they be thinned before the Monte Carlo analysis of the subpedigrees; the sampling process assumes markers not in LD, and MCMC is prohibitive with dense markers. To provide the most information for within-pedigree IBD, we chose markers that were heterozygous in at least half of the individuals. We then thinned these markers so that they were spaced between 0.3 and 0.7 cM apart, with preference given to markers that provided heterozygous individuals in all three of the subpedigrees. In fact, only the largest of the three pedigrees required MCMC analysis. For the two smaller components, independent realizations of IBD across the chromosome were sampled conditional on the marker data. The sampler was run using a forward-backward algorithm on the inheritance vectors (Thompson, 2000).

The IBD graphs obtained from the larger two families were merged first; this output was then merged with the graphs from the smallest family. Merging was performed at every marker in the set of breakpoints present in the two input graphs. The lod curves from the merge appear in Figure 5. Also shown is the lod score that would be obtained if the full true IBD within and among the three subpedigrees were observable across the chromosome, and the very weak signal provided by independent analysis of the three subpedigrees. With the exception of a spike at about 140 cM, the lod score from the merged pedigrees roughly captures the true curve. The location and magnitude of the peak at the trait locus are recovered.

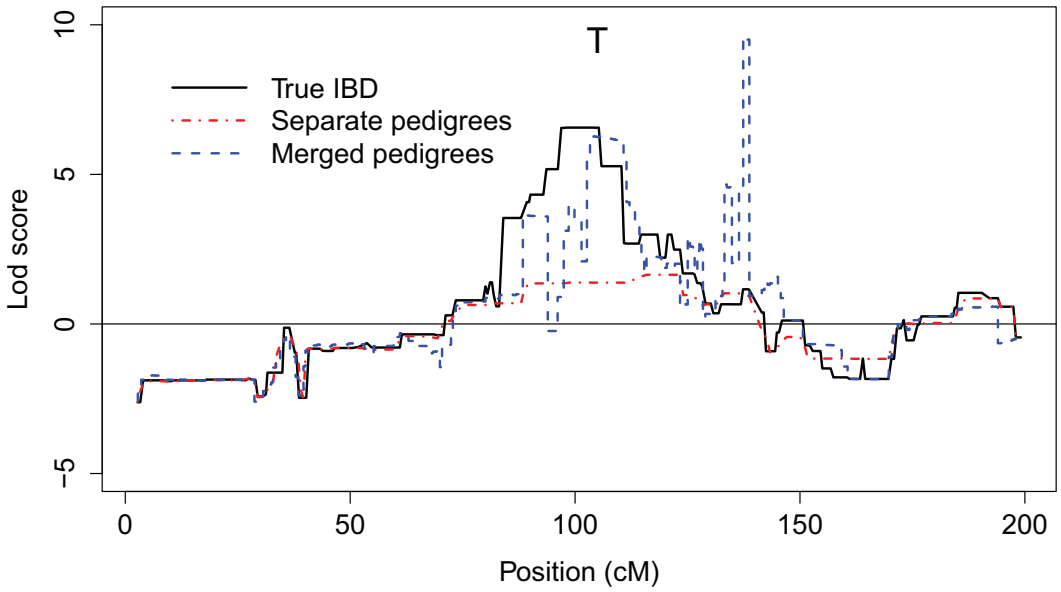


Figure 5: True lod score on the simulated Iceland dataset, and estimates using separate and merged pedigrees. The trait locus is marked by the "T".

While the maximum occurs at the location of the anomalous spike, such a spike in the lod score is improbable under normal recombination patterns on pedigrees. Closer examination of the recombination events implied by the merged IBD graphs shows the spike to be a statistical artifact. The presence of the spike highlights a weakness of our approach: since each marker is merged independently, small differences in the ordered IBD states between markers can lead to discontinuities in the resulting lod score.

A second data set was also simulated on the Iceland pedigree using the same inputs and parameters. The same trait-locus descent was assumed, but now located at 25 cM. In addition, we simulated several confounding sources of coancestry. In particular, at 175 cM, we forced descent with no trait effect from the oldest founder generation to most members of all three families. Marker data were generated and analyses performed exactly as in the first Iceland example.

Figure 6 shows that in this example the merging procedure recovered the linkage signal at the trait locus where the subpedigrees showed almost no indication of linkage. We also see that it detected the ancestral coancestry at 175 cM, which has a very high lod score but is not causally tied to the trait. This signal is correct in that the true IBD on the full pedigree shows the same result. By imposing descent to the three families, IBD among the three families at 25 cM is highly correlated with

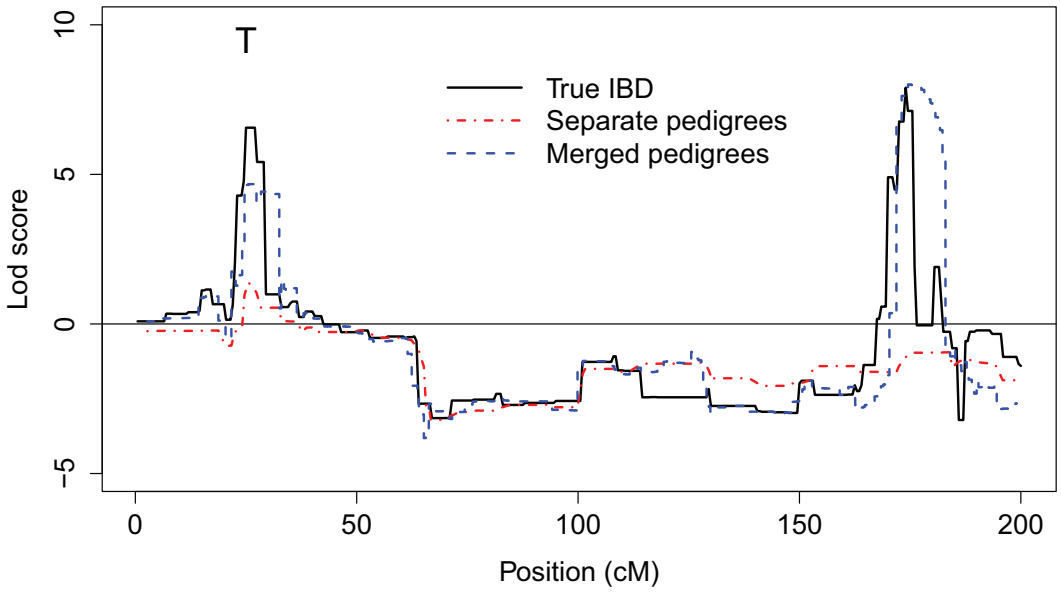


Figure 6: Lod scores for simulated data on the Iceland pedigree with confounding coancestry. The trait locus is marked by the "T".

that at 175 cM. Interestingly, this has no impact on the within-family results; the between-family correlation derives from the constrained ancestral structure. This example provides a warning that expanding the scope of the IBD used to map a trait can draw in coancestry not connected to the trait and lead to incorrect conclusions. While the example presented here is extreme, this problem could arise among families ascertained in a small or highly structured population. Long-distance correlation between causal genes and other parts of the genome poses similar problems for association mapping; see, for example, Di et al. (2011).

5 Discussion

We have presented a method for merging data from multiple families to enhance pedigree-based linkage analysis. Simulated examples indicate that when families carry a trait gene from the same recent ancestor, the shared ancestry can be detected and used to augment the information in the pedigrees. The IBD graph is an essential tool in our method, because it expresses coancestry without explicit reference to family relationships.

The merging algorithm at present is somewhat crude and satisfies only the basic requirement of producing graphs broadly consistent with the HMM output. A more sophisticated approach would treat the inference of connections between graphs as an imputation problem and account for the uncertainty introduced by this step. Doing so might remove some of the sharp discontinuities and outlying lod scores caused by choosing extreme graph configurations at some loci.

Another modification which would reduce these outliers would be to use information from adjacent loci when merging, rather than to treat loci independently. Including prior knowledge of the behavior of real lod scores, which do not jump so dramatically, could allow for some smoothing of the merged lod scores.

Computational efficiency is a concern both with the current method and potential modifications. The algorithm presented here is slow for large numbers of loci, partly because of the brute force method of testing graph configurations. Exploring the space of potential solutions in a more intelligent way would reduce the computational burden.

While it would benefit from these improvements, our current implementation is adequate for demonstrating that linkage information can be recovered by combining genetic models for pedigrees and populations. With further development, we anticipate that this approach will be a useful tool in epidemiological applications.

References

Boehnke, M. (1994): "Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes," *American Journal of Human Genetics*, 55, 379–390.

Brown, M., C. Glazner, C. Zheng, and E. Thompson (2011): "Inferring coancestry in population samples in the presence of linkage disequilibrium," *To be submitted*.

Browning, B. L. and S. R. Browning (2007): "Efficient multilocus association testing for whole genome association studies using localized haplotype clustering," *Genetic Epidemiology*, 31, 365–375.

Browning, S. R. and B. L. Browning (2010): "High-resolution detection of identity by descent in unrelated individuals," *The American Journal of Human Genetics*, 86, 526 – 539.

Cupples, L. A., N. Heard-Costa, M. Lee, and L. D. Atwood (2009): "Genetics Analysis Workshop 16 problem 2: the Framingham Heart Study data," *BMC Genetics*, 3(Suppl 7), S3.

Di, Y., G. Mi, L. Sun, R. Dong, H. Zhu, and L. Peng (2011): "Power of association tests in the presence of multiple causal variants," *BMC Proceedings*.

Eén, N. and N. Sörensson (2003): "An extensible SAT-solver," in E. Giunchiglia and A. Tacchella, eds., *SAT*, *Lecture Notes in Computer Science*, volume 2919, Springer, *Lecture Notes in Computer Science*, volume 2919, 502–518.

Glazner, C., M. D. Brown, Z. Cai, and E. A. Thompson (2010): "Inferring coancestry in structured populations," Abstract, Western North American Region of the IBS Annual Meeting.

Knuth, D. E. (2008): *The Art of Computer Programming: Volume 4, Fascicle 0. Introduction to Combinatorial Algorithms and Boolean Functions*, Upper Saddle River, NJ: Addison-Wesley.

Lange, K. and E. Sobel (1991): "A random walk method for computing genetic location scores," *American Journal of Human Genetics*, 49, 1320–1334.

Rabiner, L. R. (1989): "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, 77, 257–286.

Sobel, E. and K. Lange (1996): "Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics," *American Journal of Human Genetics*, 58, 1323–1337.

Sun, L., K. Wilder, and M. S. McPeek (2002): "Enhanced pedigree error detection," *Human Heredity*, 54, 99–110.

Thompson, E. A. (2000): *Statistical Inferences from Genetic Data on Pedigrees*, *NSF-CBMS Regional Conference Series in Probability and Statistics*, volume 6, Beachwood, OH: Institute of Mathematical Statistics.

Thompson, E. A. (2007): "Linkage analysis," in D. J. Balding, M. Bishop, and C. Cannings, eds., *Handbook of Statistical Genetics; 3rd edition*, Chichester, UK: Wiley, 1141–1167.

Thompson, E. A. (2008): "The IBD process along four chromosomes," *Theoretical Population Biology*, 73, 369–373.

Thompson, E. A. (2009): "Inferring coancestry of genome segments in populations," in *Invited Proceedings of the 57th Session of the International Statistical Institute*, Durban, South Africa, IPM13: Paper 0325.pdf.

Thompson, E. A. (2011): "The structure of genetic linkage data: from LIPED to 1M SNPs," *Human Heredity*, 71, 86–96.

Tong, L. and E. A. Thompson (2008): "Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations," *Human Heredity*, 65, 142–153.

Wijsman, E. M., J. H. Rothstein, and E. A. Thompson (2006): "Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain Monte Carlo provides practical approaches for genome scans on general pedigrees," *American Journal of Human Genetics*, 79, 846–858.