Inferring coancestry in structured populations

Chris Glazner, Marshall Brown, Zheng Cai, and Elizabeth Thompson

University of Washington, Statistics

June 21, 2010
Detecting IBD

- Identity by descent (IBD) consists of DNA segments shared due to relatively recent common ancestry.
- Identifying IBD segments within and between individuals can increase the power of pedigree-based linkage analysis.
- The goal is to connect smaller pedigrees whose relationship to each other is not known.
- IBD segments shared by individuals in different pedigrees indicate ancestral links between the pedigrees at those loci.
Identity by descent

IBD can be expressed in terms of founder genome labels (FGLs), arbitrary labels assigned to founder haplotypes in a pedigree. This diagram shows two pairs of haplotypes, made up of recombined segments of founder haplotypes.

Color indicates FGL. Same color $\Rightarrow$ IBD.
Detecting IBD

We use a Hidden Markov Model to detect IBD segments. The IBD state is the hidden state.

Thompson (2008) describes the continuous time Markov process of the IBD states along haplotypes.

- The stationary distribution is parametrized by the marginal probability of IBD between two haplotypes
- The transition rate matrix is additionally parametrized by the average length of IBD segments
To gauge the performance of the method, we simulate a population of haplotypes.
We simulate recombination events in a population:

- 7,000 individuals
- 200 generations
- Constant population size
- Two offspring per marriage, random marriages

The resulting “haplotypes” consist of FGLs—which define the IBD structure among them—but have no alleles assigned to marker loci.
Allele model and data

- The observed state in the HMM is the shared allele pattern at each locus.
- Realistic patterns of allele dependence are hard to simulate, so real data was needed as input.
- We obtained haplotypes from the Framingham Heart Study. To avoid having to phase genotypes, we used X chromosomes from males.
- By assigning a haplotype to each FGL, we can then assign alleles and create simulated haplotypes.
Summaries of the SNPs used as founder alleles.

- Mean = 0.017

<table>
<thead>
<tr>
<th>Distance between SNPs (Mbp)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>0.15</td>
<td>0</td>
</tr>
<tr>
<td>0.20</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minor allele frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>0.5</td>
</tr>
</tbody>
</table>
Latent IBD among Framingham haplotypes

We would like our founder data to have realistic patterns of linkage disequilibrium (LD) but not show a high degree of relatedness. Recall: different colors $\Rightarrow$ no IBD.

Instead, we detect a great deal of IBD among our founders. Below, a black segment indicates IBD inferred among a random set of four founders:
We ran the BEAGLE program of Sharon and Brian Browning on our 1,900 founder haplotypes to fit a Markov model of the dependencies among markers.

Using haplotypes simulated from the BEAGLE model, we infer about as much IBD as the original founder haplotypes. This suggests that the detected IBD is not due to cryptic relationships among founders.
Simulating haplotypes

To retain realistic LD over short genetic distances but eliminate the inferred IBD among founders, we simulated from the BEAGLE model while breaking up dependence over long distances. The fitted Markov model was “reset” every few markers to break up LD. These haplotypes were used in the final analysis.

Below, the inferred IBD among four simulated founder haplotypes.
An example

- Inference using haplotypic data
- Colors indicate state (for example, \([1,1,2,3]\))
Fraction of markers in true IBD segments correctly identified as being in some IBD state, by segment length.
Fraction of markers in true IBD segments with state correctly identified, by segment length.
The model detects IBD well at lengths over 1 cM with genotypic data and 0.5 cM with haplotypic data.

Population data alone will in most cases only provide genotypic data.

When the model is used in conjunction with local pedigrees, phasing is possible, so haplotypic data may be available.

Tuning parameters: the model appears robust to different assumptions about kinship in the population, but changing the assumed rate of switching between states does change inferences.