LD Mapping and the Coalescent

Zhaojun Zhang
zzj@cs.unc.edu

April 2, 2009
Outline

1. Linkage Mapping
   - Introduction

2. Linkage Disequilibrium Mapping
   - Introduction

3. A role for coalescent

4. Prove existence of LD on simulated data
   - Qualitative measure
   - Quantitative measure
Linkage and Linkage Mapping

Definition

Genetic Linkage occurs when particular genetic loci or alleles for genes are inherited jointly. Genetic loci on the same chromosome are physically connected and tend to segregate together during meiosis, and are thus genetically linked.

Reason

Recombination
A example of linkage
Figure 7.1  Linkage mapping. Search for cosegregation of a marker allele with disease status. In this case a marker locus $M$ is typed in an affected female (filled circle) and an unaffected male (empty square) and their children. The allele $m_1$ is seen to cosegregate with the dominant disease in this example.
Disadvantage of Linkage Mapping

- Successfully been applied for the identification of a large number of rare, single gene human diseases.
Disadvantage of Linkage Mapping

- Successfully been applied for the identification of a large number of rare, single gene human diseases.
- However, the resolution of the method is determined by the number of meiosis observed in pedigree, which is not very big number.
Disadvantage of Linkage Mapping

- Successfully been applied for the identification of a large number of rare, single gene human diseases.
- However, the resolution of the method is determined by the number of meiosis observed in pedigree, which is not very big number.
- Recombination rate will also be very low in the nearby region. (Limited meiosis)
Disadvantage of Linkage Mapping

- Successfully been applied for the identification of a large number of rare, single gene human diseases.
- However, the resolution of the method is determined by the number of meiosis observed in pedigree, which is not very big number.
- Recombination rate will also be very low in the nearby region. (Limited meiosis)
- The region is likely to cover several Mb of DNA sequence and to contain several hundred genes.
A example for complex disease

**Figure 7.2** An example genealogy for a case-control data set of a complex disease. The mutation increasing risk of the disease may not be unique (i.e. has happened more than once in the sample’s history), or several mutations in positions in tight linkage with each other may have occurred. Further the disease may not be fully penetrant. Cases have been oversampled compared to the frequency of the disease in the population. (Figure adapted from McVean, personal communication.)
Linkage Disequilibrium Mapping (LD Mapping)

From previous chapter, we know that for human populations might be 4,000 and 40,000 generations. It must contains more meiosis.
Introduction

Linkage Disequilibrium Mapping (LD Mapping)

- From previous chapter, we know that for human populations might be 4,000 and 40,000 generations. It must contains more meiosis.

- The more meiosis, and the more the recombination happens, and the shorter the relative region is.
Introduction

Linkage Disequilibrium Mapping (LD Mapping)

- From previous chapter, we know that for human populations might be 4,000 and 40,000 generations. It must contains more meiosis.
- The more meiosis, and the more the recombination happens, and the shorter the relative region is.
- A number of genetic markers are typed at known position.
Linkage Disequilibrium Mapping (LD Mapping)

- From previous chapter, we know that for human populations might be 4,000 and 40,000 generations. It must contains more meiosis.
- The more meiosis, and the more the recombination happens, and the shorter the relative region is.
- A number of genetic markers are typed at known position.
- It becomes possible, in principle at least, to map a trait of interest much more accurately.
It is hard in practise...

- A lot of factors ...
- Hard to define phenotypes ...
- Hard to define the model ...
- Computational complexity ...
It is hard in practise...

- A lot of factors ...
- Hard to define phenotypes ...
- Hard to define the model ...
- Computational complexity ...
- Currently it is unknown how complicated the relationship between phenotype and genotype can be.
Complex disease aetiology

![Diagram showing markers and factors](image)

**Figure 7.3** Some factors complicating identification of a disease-related gene: Markers $M_1$, $M_2$, and $M_3$ are used indirectly to infer the disease-related gene, which may contain different mutations conferring the disease phenotype (here $D_1$ and $D_2$). A number of other factors complicate localisation, including the effect of other genes elsewhere in the genome, population stratification, and the environment. Some of these complicating factors can be directly modelled whereas others are best treated as noise.
Phenotype of disease

- Normally, the phenotype of a disease is defined as a binary character that takes the value 1 if an individual has the disease and 0 if the individual is not affected by the disease.
Phenotype of disease

- Normally, the phenotype of a disease is defined as a binary character that takes the value 1 if an individual has the disease and 0 if the individual is not affected by the disease.
- Many human disorders are very heterogeneous and the phenotype can only superficially be classified.
- (Author’s view) Dictomising disease status might reduce the chance of locating important genes and gene variations.
General Model

\[
\phi = \phi(g_1, g_2, \ldots, g_k, e_1, e_2, \ldots, e_l), \quad (7.1)
\]

<table>
<thead>
<tr>
<th>Haplotypes</th>
<th>Cell level</th>
<th>Organism level</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCD</td>
<td>Phenotype 1</td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D G</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C EF</td>
<td>Phenotype 2</td>
<td>Disease phenotype</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Phenotype 3</td>
<td></td>
</tr>
<tr>
<td>D H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H H</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cell environment → Physical environment

**Figure 7.4** Illustration of how sets of different combinations of genetic determinants (here called A, B, \ldots, H) may lead to the same phenotype in the cell (e.g. affecting the same biochemical pathway) and how these different cell phenotypes may result in the same phenotype for the individual (e.g. the same disease diagnosis).
No explicit way ...

- Single gene with one allele gives rise to $2^3 - 2 = 6$ different mapping, without environmental effects and a binary phenotype.
No explicit way ...

- Single gene with one allele gives rise to $2^3 - 2 = 6$ different mapping, without environmental effects and a binary phenotype.
- Two genes gives rise to $2^9 - 2$
- Three genes ...
Generation of simulated data

- The whole geneology history is totally known.
Generation of simulated data

- The whole geneology history is totally known.
- Validate methods for localising genes and gene variants; in particular it is possible to investigate the effects of marker density, marker frequencies, sample sizes and so on.
Generation of simulated data

- The whole geneology history is totally known.
- Validate methods for localising genes and gene variants; in particular it is possible to investigate the effects of marker density, marker frequencies, sample sizes and so on.
- Better understand the process of how recombination and gene conversion influence these. (More details later)
Analysis for real data

\[
P(G \mid m_{ij}, \phi_j, j = 1, \ldots, n, i = 1, \ldots, m)
= \frac{P(G, m_{ij}, \phi_j, j = 1, \ldots, n, i = 1, \ldots, m)}{P(m_{ij}, \phi_j, j = 1, \ldots, n, i = 1, \ldots, m)},
\]  

(7.5)

where

\[
P(m_{ij}, \phi_j, j = 1, \ldots, n, i = 1, \ldots, m)
= \sum_G P(G, m_{ij}, \phi_j, j = 1, \ldots, n, i = 1, \ldots, m).
\]  

(7.6)
Association of thirteen SNP markers with susceptibility to early onset of Alzheimer’s disease
**Break relationship with marker**

**Figure 7.6** LD. A new mutation, $D$, contributing to a disease initially enters a population through mutation on a specific genetic background. After many generations recombination has decoupled $D$ from its initial genetic background except for a (small) region around $D$. A marker, $M$, in this region is in LD with $D$. 

Zhaojun Zhang  zzj@cs.unc.edu
LD Mapping and the Coalescent
It is not graph for each position
Qualitative measures

Figure 7.7  Venn diagram of the three qualitative measures of tree similarity.

Zhaojun Zhang  zzj@cs.unc.edu
LD Mapping and the Coalescent
Without gene conversion
Qualitative measure

With gene conversion

Zhaojun Zhang  zzj@cs.unc.edu
LD Mapping and the Coalescent
The number of segments with the same tree

Figure 7.10 The number of segments with the same tree as at the target position for two situations: with (grey columns) and without (black columns) gene conversion.
\[ M_{AB} = \frac{\sum_{i,j} I_{i=j} a_i b_j}{l_A l_B}, \quad (7.7) \]

where \( a_i \) (\( b_j \)) is the length of branch \( i \) (\( j \)) in tree \( A \) (\( B \)), \( I_{i=j} \) is an indicator function which is one if branch \( i \) in \( A \) and branch \( j \) in \( B \) give the same bipartition, and \( l_A \) (\( l_B \)) is the total branch length of tree \( A \) (\( B \)). It is reasonable to compare \( M_{AB} \) with \( M_{AA} \), the similarity of a tree to itself.
An example

Figure 7.11 Two different trees for five sequences, differing by at least one recombination event. Dashed lines correspond to bipartitions not found in the other tree.
Quantitative measure

\[ S_{AB} \]

\[ M_{AA} = \frac{\sum_{i,j} a_i^2}{l_A^2}, \quad (7.8) \]

and define the ratio by

\[ S_{AB} = \frac{M_{AB}}{M_{AA}} \quad (7.9) \]
Without gene conversion
With gene conversion
Another possible measure is the difference between two trees, $A$ and $B$, as measured by the sum of the differences of branch lengths leading to the same bipartition relative to the total branch lengths. That is,

$$N_{AB} = \sum_i \left( \frac{a_i}{l_A} - \frac{b_i}{l_B} \right) = 1 - \sum_i \frac{b_i}{l_B},$$

(7.10)
Without gene conversion
With gene conversion
Next time ...

- LD measure
- Model based LD mapping