LD Mapping and the coalescent

Zhaojun Zhang
zzj@cs.unc.edu

April 7, 2009
NCAA Champion!

Zhaojun Zhang  zzj@cs.unc.edu
LD Mapping and the coalescent
Outline

1. Linkage disequilibrium (LD)
   - LE and LD
   - Three measurements of LD

2. Practical Issues
   - SNP
   - Differences between human populations
   - Others ...

3. LD Mapping
   - Measuring association using single markers
   - Haplotype LD Mapping
   - Model based LD mapping
Outline

1. Linkage disequilibrium (LD)
   - LE and LD
   - Three measurements of LD

2. Practical Issues
   - SNP
   - Differences between human populations
   - Others ...

3. LD Mapping
   - Measuring association using single markers
   - Haplotype LD Mapping
   - Model based LD mapping
Tree vs markers

- No tree is observed directly.
- Rely on genetic markers to get as much information as possible.
Linkage equilibrium (LE)

Locus A

- $A = \square$
- $p_A = 0.5$
- $a = \square$
- $p_a = 0.5$

Locus B

- $B = \bullet$
- $p_B = 0.5$
- $b = 0$
- $p_b = 0.5$

$p_{AB} = 0.25$
$p_{Ab} = 0.25$
$p_{aB} = 0.25$
$p_{ab} = 0.25$
Linkage Disequilibrium (LD)

Locus A
- A = □
- p_A = 0.5
- a = □
- p_a = 0.5

Locus B
- B = ○
- p_B = 0.5

- p_{AB} = 0.40
- p_{aB} = 0.10
- p_{Ab} = 0.10
- p_{ab} = 0.40

Zhaojun Zhang zzj@cs.unc.edu
LD Mapping and the coalescent
Three common measures of LD are:

- $D$ (Lewontin & Kojima, 1960)
- $D'$ (Lewontin, 1964)
- $r^2$ (Hill & Robertson, 1968 and Franklin & Lewontin, 1970)
A simple measure of LD between two loci (say locus A and locus B) is given by $D$ (Lewontin & Kojima, 1960):

$$D = p_{AB}p_{ab} - p_{Ab}p_{aB}$$

with $p_{AB}$, $p_{ab}$, $p_{Ab}$, and $p_{aB}$ the frequencies of the gametes $AB$, $ab$, $Ab$, and $aB$ respectively.

Alternatively, $D$ can be expressed as:

$$D = p_{AB} - p_Ap_B$$

with $p_{AB}$ the frequency of gamete $AB$, and $p_A$ and $p_B$ the frequencies of alleles $A$ and $B$. 
Linkage disequilibrium (LD)

Practical Issues

LD Mapping

Three measurements of LD

$p_B=0.01$ (white), $p_B=0.10$ (blue), $p_B=0.25$ (red), $p_B=0.50$ (green), $p_B=0.75$ (violet), $p_B=0.90$ (yellow), and $p_B=0.99$ (black).

Zhaojun Zhang  zzj@cs.unc.edu

LD Mapping and the coalescent
Properties of $D$

- When the loci are in linkage equilibrium, $D=0$.
- The values of $D$ can be negative or positive depending on which are the most frequent gametes. If the most frequent gametes are $AB/ab$ $D$ takes positive values ($D \geq 0$), and when the most frequent gametes are $Ab/aB$, then $D$ takes negative values ($D \leq 0$).
A disadvantage of $D$ as a measure of LD is that the range of values that it can take depends on the allele frequencies, which makes difficult to compare $D$ values from different pairs of loci.

An alternative is to express LD by $D'$ which is a normalized value of $D$ that essentially expresses $D$ as a percentage of its maximum value when $D>0$ or its minimum value when $D<0$:

If $D > 0$:
$$D' = \frac{D}{\min \{p_A p_b, p_a p_B\}}$$

If $D < 0$:
$$D' = \frac{D}{\max \{-p_A p_B, p_a p_b\}}$$
Properties of D’

- When the loci are in linkage equilibrium $D’ = 0$
- For all pairs of loci $D’$ values range between 0 and 1 (independently of allele frequencies).
Without gene conversion
With gene conversion
Another commonly used measure of LD is the \textbf{simple correlation} $r$ between the A and B alleles (Hill & Robertson, 1968 and Franklin & Lewontin, 1970). This measure is both simple and, similarly to $D'$, it allows to compare values from different loci as its range of values is independent of the allele frequencies. This LD measure can be defined as:

$$r = \frac{D}{\sqrt{P_A P_a P_B P_b}}$$
Properties of $r^2$

- When the loci are in linkage equilibrium $r^2 = 0$.
- $r^2$ values range between 0 and 1 (0 = linkage equilibrium; 1 = complete linkage).
- The test of the null hypothesis of $D = 0$ can be easily done by comparing the value of $r^2n$ (n the sample size) with a Chisquare distribution with 1 degree of freedom.
Without gene conversion
With gene conversion
Differences

Figure 7.16  The behaviour of $r^2$ and $D'$ for the three cases with two, three, and four haplotypes in a sample of four chromosomes. $D' < 1$ only when the number of haplotypes is four.
$R^2$ With gene conversion

Zhaojun Zhang  zzj@cs.unc.edu
LD Mapping and the coalescent
Linkage disequilibrium (LD)

Three measurements of LD

D’ With gene conversion
Testing for LD

\[ X = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \quad (7.14) \]

where \( e_{ij} \) is the estimated expected number of \( n_{ij} \) under the assumption of no association, \( e_{ij} = (n_{i1} + n_{i2})(n_{1j} + n_{2j})/n \). If all \( e_{ij} \) are sufficiently large (say greater than five), then \( X \) is approximately \( \chi^2 \) distributed with one degree of freedom. Figure 7.17 shows how the relative probability of observing significant LD decays with distance.
Outline

1. Linkage disequilibrium (LD)
   - LE and LD
   - Three measurements of LD

2. Practical Issues
   - SNP
   - Differences between human populations
   - Others ...

3. LD Mapping
   - Measuring association using single markers
   - Haplotype LD Mapping
   - Model based LD mapping
The genealogical process reflected in data

SNPs are the preferred marker type of LD mapping because of their abundancy in the genome, however, large differences in the density of SNPs exist over the genome.
The genealogical process reflected in data

SNPs are the preferred marker type of LD mapping because of their abundancy in the genome, however, large differences in the density of SNPs exist over the genome.

- Variance in the genealogical process.
The genealogical process reflected in data

SNPs are the preferred marker type of LD mapping because of their abundancy in the genome, however, large differences in the density of SNPs exist over the genome.

- Variance in the genealogical process.
- Differences in the per base pair recombination and gene conversion rates over the genome.
The genealogical process reflected in data

SNPs are the preferred marker type of LD mapping because of their abundancy in the genome, however, large differences in the density of SNPs exist over the genome.

- Variance in the genealogical process.
- Differences in the per base pair recombination and gene conversion rates over the genome.
- Differences in mutation rate over the genome.
The genealogical process reflected in data

SNPs are the preferred marker type of LD mapping because of their abundancy in the genome, however, large differences in the density of SNPs exist over the genome.

- Variance in the genealogical process.
- Differences in the per base pair recombination and gene conversion rates over the genome.
- Differences in mutation rate over the genome.
- Differences in the selective regime over the genome.
The genealogical process reflected in data

SNPs are the preferred marker type of LD mapping because of their abundancy in the genome, however, large differences in the density of SNPs exist over the genome.

- Variance in the genealogical process.
- Differences in the per base pair recombination and gene conversion rates over the genome.
- Differences in mutation rate over the genome.
- Differences in the selective regime over the genome.
Definition

In genetics, a centimorgan (abbreviated cM) or map unit (m.u.) is a unit of recombinant frequency for measuring genetic linkage. It is often used to imply distance along a chromosome. The number of base-pairs it corresponds to varies widely across the genome (different regions of a chromosome have different propensities towards crossover).
The centimorgan is equal to a 1% chance that a marker at one genetic locus on a chromosome will be separated from a marker at a second locus due to crossing over in a single generation.

"...in humans 1 centimorgan on average represents a distance of about 7.5x10^5 base pairs"
Nucleotide diversity

Figure 7.18 The distribution of nucleotide diversity in 150 genes in samples from an African-American population and from the CEPH reference panel. Data are from the Seattle SNP project surveying variation in 150 candidate genes for inflammatory responses in humans (SeattleSNPs. NHLBI Program for Genomic Applications, UW-FHCRC, Seattle)
Figure 7.19  The distribution of Tajima’s $D$ values for the same data set as in Figure 7.
Others ...

- Population admixture.
- Missing marker data.
- multiple founders.
- ...
Outline

1. Linkage disequilibrium (LD)
   - LE and LD
   - Three measurements of LD

2. Practical Issues
   - SNP
   - Differences between human populations
   - Others ...

3. LD Mapping
   - Measuring association using single markers
   - Haplotype LD Mapping
   - Model based LD mapping
Principle idea

The basis for a test of association of a marker allele $m$ at the marker locus $M$ for a binary disease trait is the null hypothesis

$$P(m \mid \text{disease}) = P(m \mid \text{non-disease}). \quad (7.15)$$
Observation

- None of the markers are individually strongly associated with phenotype.
Observation

- None of the markers are individually strongly associated with phenotype.
- Considering whole haplotypes consisting of several adjacent markers rather than single markers one at a time
**Figure 7.20** Haplotypes associated with schizophrenia at the neuregulin gene. Three multilocus haplotypes are highly significantly overrepresented in affected persons ($\chi^2$ test of association). These three haplotypes share a core haplotype (shaded) of seven markers covering a region of 290 kb within which the causative mutation(s) is expected to reside (data from Stefansson et al. 2003).
Principle idea

\[
f(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters}) f(\text{parameters})}{P(\text{data})}, \quad (7.16)
\]

where \( P(\text{data} \mid \text{parameters}) \) is the likelihood of the data given a set of parameters.

If one is interested in the distribution of just a subset of the parameters, called \( x \) (e.g. the disease position), integration is then done over the remaining parameters

\[
f(x \mid \text{data}) = \int_{\text{parameters except } x} f(\text{parameters} \mid \text{data}) \, d(\text{parameters except } x). \quad (7.17)
\]
Liu’s model

Figure 5  A graphical representation of the haplotype model. There are a total of $k$ markers. Parameter $\tau$ is the “recombination distance” from the disease locus to the leftmost marker (which is equal to $-\log\{(1+e^{-2d})/2\}$, where $d$ is the genetic distance). The recombination event closest to the disease locus from the left arm occurred between markers $R_1$ and $R_1 + 1$ and that from the right arm occurred between markers $R_2$ and $R_2 + 1$. 

Zhaojun Zhang  zzj@cs.unc.edu
LD Mapping and the coalescent
Liu’s model

\[ r(i, G, j_1, j_2) \approx \begin{cases} (1 - r)^G \approx 1 - rG & \text{if } j_1 = j_2, \\ \frac{[1 - (1 - r)^G]}{(n_i - 1)} & \text{if } j_1 \neq j_2, \end{cases} \]
Liu’s model

\[
\Pr(H_t^t, C_t^t = c, R_1^t, R_2^t \mid A, G, \tau) = \alpha_c \Pr(H_t^t, R_1^t, R_2^t \mid A_{c^*}, G_c, \tau, C_t^t = c) \\
= a_c \Pr(H_{\leq L}^t, R_1^t \mid A_{c^*}, G_c, \tau, C_t^t = c) \\
\times \Pr(H_{> L}^t, R_2^t \mid A_{c^*}, G_c, \tau, C_t^t = c),
\]
Liu’s model

\[
\Pr\left( H_{\leq L}, R_1^t \mid A_c, G_c, \tau, C^t = c \right) = \left( 1 - e^{-(r_{R_1^t+1} - \tau_{R_1^t})G_c} \right)e^{-(\tau - \tau_{R_1^t+1})G_c}
\]

\[
\prod_{j \leq R_1^t} p_{jH_j^t} \prod_{R_1^t < j \leq L} r(j, G_c, A_c, j, H_j^t),
\]

and

\[
\Pr\left( H_{> L}, R_2^t \mid A_c, G_c, \tau, C^t = c \right) = \left( 1 - e^{-(r_{R_2^t+1} - \tau_{R_2^t})G_c} \right)e^{-(\tau_{R_2^t} - \tau)G_c}
\]

\[
\prod_{j > R_2^t} p_{jH_j^t} \prod_{L < j \leq R_2^t} r(j, G_c, A_c, j, H_j^t).
\]
Liu’s model

\[
\Pr(H^t, C^t = c \mid A, G, \tau) =
\begin{cases} 
    a_0 \prod_{j=1}^{m} p_{jH_j} & \text{if } c = 0 \\
    a_c \sum_{R_1 \leq L} \Pr(H^t_{\leq L}, R_1^t \mid A, G, \tau, C^t = c) \\
    \times \sum_{R_2 > L} \Pr(H^t_{> L}, R_2^t \mid A, G, \tau, C^t = c) & \text{if } c \neq 0.
\end{cases}
\]

Hence, \( \Pr(H^t \mid A, G, \tau) = \sum_{c=0}^{k} \Pr(H^t, C^t = c \mid A, G, \tau) \).
Rannala and Reeve’s model
The likelihood of the sampled disease chromosome haplotypes (and unobserved ancestral haplotypes parameters in the model, is

$$f(X,Y|\theta,\tau,Y_0,d,p) = \prod_{i=1}^{n} f(X_i|Y,Y_0,\theta,d,p,v) \times \prod_{i=1}^{n-1} f(Y_i|Y,Y_0,\theta,d,p,w)$$

where $v = \{v_1,\ldots,v_n\}$ is a vector of the lengths of the terminal branches on the gene tree and $w = \{w_1,\ldots,w_{n-1}\}$ is a vector of the internal branches. These quantities, which are determined by the tree topology.
Rannala and Reeve’s model

In the present study, a Bayesian multipoint LD mapping method is developed that is based on the following probability-density function

\[
f(\theta, Y, \tau, Y_0 | X, \Lambda, p, d, \Omega) = \frac{f(X, Y | \theta, \tau, Y_0, d, p) g(\theta | \Omega) b(Y_0 | p) r(\tau | \Lambda)}{f(X | \Lambda, p, d, \Omega)}. \tag{1}
\]
Shattered genealogy

- Multiple finding disease mutations at the same locus.
- Sporadic cases of disease, caused by environmental factors or mutations at other loci.
Morris’ model
Morris’ model

\[ z_b = \begin{cases} 
1 & \text{if node } b \text{ has a parental node in the shattered genealogical tree,} \\
0 & \text{if node } b \text{ has no parental node in the shattered genealogical tree.} 
\end{cases} \]

\[
\pi(\mathcal{T}, T, z|\rho) = \left[ \prod_k \lambda_k \exp(-\lambda_k w_k) \right] \left[ \prod_b \rho^{z_b} (1 - \rho)^{1-z_b} \right],
\]
Morris’ model
Morris’ model

\[
f(x, h, \omega, T, z, N, \rho | A, U) \sim P(A, U | I, x, h, \omega, T, z, N) \\
\times f(\omega, T, z | \rho) f(\rho)
\]

Table 7.3  The key parameters in the approach of Morris et al. to Bayesian multipoint LD mapping

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>Location of disease locus</td>
</tr>
<tr>
<td>h</td>
<td>Population marker-haplotype proportions</td>
</tr>
<tr>
<td>\omega</td>
<td>Branch lengths of genealogical tree</td>
</tr>
<tr>
<td>T</td>
<td>Topology (branching pattern)</td>
</tr>
<tr>
<td>z</td>
<td>Parental status</td>
</tr>
<tr>
<td>N</td>
<td>Effective population size</td>
</tr>
<tr>
<td>\rho</td>
<td>Shattering parameter</td>
</tr>
<tr>
<td>A</td>
<td>Cases (affected)</td>
</tr>
<tr>
<td>U</td>
<td>Controls (unaffected)</td>
</tr>
<tr>
<td>I</td>
<td>Haplotypes at internal nodes in the case tree</td>
</tr>
</tbody>
</table>
An example
Conclusion

- LD Mapping just find related markers, however it cannot find causitive markers.
- LD Mapping is more popular and popular, however it cannot guarantee that it will be more successful.