Introduction to Linkage Disequilibrium

Sorin Istrail

September 10, 2014
Suppose we have two genes on a single chromosome

gene A and gene B

such that each gene has only two alleles

\( A\text{alleles} : A_1 \text{ and } A_2 \)

\( B\text{alleles} : B_1 \text{ and } B_2 \)
Suppose we have two genes on a single chromosome

gene A and gene B

such that each gene has only two alleles

\[ Aalleles : A_1 \text{ and } A_2 \]

\[ Balleles : B_1 \text{ and } B_2 \]

Possible allele combinations:

\[ A_1 B_1, A_1 B_2, A_2 B_1 \text{ and } A_2 B_2 \]
- $p_1 = \text{probability of seeing allele } A_1$
- $p_2 = \text{probability of seeing allele } A_2$
- $p_1 + p_2 = 1$

By Hardy-Weinberg principle, the probability of genotype $A_1A_1$ is $p_1^2$, the probability of genotype $A_1A_2$ is $2p_1p_2$, and the probability of genotype $A_2A_2$ is $p_2^2$.

HW equilibrium is about a single locus (with two alleles). How do we generalize to two loci?
Linkage Disequilibrium

- \( p_1 \) = probability of seeing allele \( A_1 \)
- \( p_2 \) = probability of seeing allele \( A_2 \)
- \( p_1 + p_2 = 1 \)

By Hardy-Weinberg principle,

- probability of genotype \( A_1A_1 \) is \( p_1^2 \)
- probability of genotype \( A_1A_2 \) is \( 2p_1p_2 \)
- probability of genotype \( A_2A_2 \) is \( p_2^2 \)
Linkage Disequilibrium

- $p_1 = \text{probability of seeing allele } A_1$
- $p_2 = \text{probability of seeing allele } A_2$
- $p_1 + p_2 = 1$

By Hardy-Weinberg principle,
- probability of genotype $A_1A_1$ is $p_1^2$
- probability of genotype $A_1A_2$ is $2p_1p_2$
- probability of genotype $A_2A_2$ is $p_2^2$

HW equilibrium is about a *single* locus (with two alleles). How do we generalize to two loci??
Linkage Disequilibrium

Linkage Equilibrium: two sites/genes each with two alleles

- Linkage Equilibrium: Random Association
- Linkage Disequilibrium: correlation between two loci
Linkage Disequilibrium: two sites/genes each with two alleles

- Linkage Equilibrium: Random Association
- Linkage Disequilibrium: correlation between two loci

\[
\begin{align*}
p_{11} &= \text{probability of seeing the } A_1B_1 \text{ haplotype} \\
p_{12} &= \text{probability of seeing the } A_1B_2 \text{ haplotype} \\
p_{21} &= \text{probability of seeing the } A_2B_1 \text{ haplotype} \\
p_{22} &= \text{probability of seeing the } A_2B_2 \text{ haplotype}
\end{align*}
\]
Linkage Equilibrium: two sites/genes each with two alleles

- Linkage Equilibrium: Random Association
- Linkage Disequilibrium: correlation between two loci

\[
p_{11} = \text{probability of seeing the } A_1 B_1 \text{ haplotype} \\
p_{12} = \text{probability of seeing the } A_1 B_2 \text{ haplotype} \\
p_{21} = \text{probability of seeing the } A_2 B_1 \text{ haplotype} \\
p_{22} = \text{probability of seeing the } A_2 B_2 \text{ haplotype}
\]

The sites are in Linkage Equilibrium if \( p_{11} = p_1 q_1, \ p_{12} = p_1 q_2, \ etc. \)
Linkage Disequilibrium is a *deviation* from this equilibrium:

\[ D = p_{11} - p_1 q_1 \]

Note that

\[ p_{11} = p_1 q_1 + D, \]
\[ p_{12} = p_1 q_2 - D, \]
\[ p_{21} = p_2 q_1 - D, \]
\[ p_{22} = p_2 q_2 + D. \]
### Lemma

\[ D = p_{11}p_{22} - p_{12}p_{21}. \]
Lemma

\[ D = p_{11}p_{22} - p_{12}p_{21}. \]

Proof:

\[ p_{11}p_{22} = (p_1 q_1 + D)(p_2 q_2 + D) = p_1 q_1 p_2 q_2 + p_1 q_1 D + p_2 q_2 D + D^2 \]
\[ p_{12}p_{21} = (p_1 q_2 - D)(p_2 q_1 - D) = p_1 q_1 p_2 q_2 - p_2 q_1 D - p_1 q_2 D + D^2 \]
Lemma

\[ D = p_{11}p_{22} - p_{12}p_{21}. \]

Proof:

\[
\begin{align*}
p_{11}p_{22} &= (p_1 q_1 + D)(p_2 q_2 + D) \\
&= p_1 q_1 p_2 q_2 + p_1 q_1 D + p_2 q_2 D + D^2 \\
p_{12}p_{21} &= (p_1 q_2 - D)(p_2 q_1 - D) \\
&= p_1 q_1 p_2 q_2 - p_2 q_1 D - p_1 q_2 D + D^2
\end{align*}
\]

And by subtracting these, we obtain

\[
p_{11}p_{22} - p_{12}p_{21} = D(p_1 q_1 + p_2 q_1 + p_2 q_2 + p_1 q_2) = D \times (1) = D
\]
What is the range of $D$?

Let

$D_{\text{min}} = \max \{ -p_1q_1, -p_2q_2 \}$

$D_{\text{max}} = \min \{ p_1q_2, p_2q_1 \}$

Now define:

$D' = \begin{cases} D & D > 0 \\ D_{\text{min}} & D < 0 \end{cases}$

Since $p_{11} = p_1q_1 + D$, and $p_1q_1 + D \geq 0$ (since $p_{11}$ is a probability), this implies $D \geq -p_1q_1$ (and similarly $D \geq -p_2q_2$).
What is the range of $D$?

Let

$$D_{\text{min}} = \max\{-p_1 q_1, -p_2 q_2\}$$

$$D_{\text{max}} = \min\{p_1 q_2, p_2 q_1\}$$
What is the range of $D$?

Let

$$D_{\text{min}} = \max\{-p_1 q_1, -p_2 q_2\}$$

$$D_{\text{max}} = \min\{p_1 q_2, p_2 q_1\}$$

Now define:

$$D' = \begin{cases} \frac{D}{D_{\text{max}}}, & D > 0 \\ \frac{D}{D_{\text{min}}}, & D < 0 \end{cases}.$$

Since $p_{11} = p_1 q_1 + D$, and $p_1 q_1 + D \geq 0$ (since $p_{11}$ is a probability), this implies

$$D \geq -p_1 q_1 \text{ (and similarly } D \geq -p_2 q_2 \text{)}$$
For both to be satisfied, it must be that

\[ D \geq \max\{-p_1 q_1, -p_2 q_2\} \]
For both to be satisfied, it must be that

\[ D \geq \max\{-p_1 q_1, -p_2 q_2\} \]

Similarly,

\[ D \leq \min\{p_1 q_2, p_2 q_1\} \]
For both to be satisfied, it must be that

\[ D \geq \max\{-p_1 q_1, -p_2 q_2\} \]

Similarly,

\[ D \leq \min\{p_1 q_2, p_2 q_1\} \]

For the two loci that we are considering, each loci *individually* is in Hardy-Weinberg equilibrium, but *together* a disequilibrium exists.
**Example:** Consider two SNPs in the coding region of glycoprotein A and glycoprotein B that change the amino acid sequence. Both of the proteins are on chromosome 4 and are found on the outside of red blood cells.

<table>
<thead>
<tr>
<th>SNP</th>
<th>AminoAcids</th>
</tr>
</thead>
<tbody>
<tr>
<td>For Protein A :</td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>Serine</td>
</tr>
<tr>
<td>$G$</td>
<td>Leucine</td>
</tr>
<tr>
<td>For Protein B :</td>
<td></td>
</tr>
<tr>
<td>$T$</td>
<td>Methianine</td>
</tr>
<tr>
<td>$C$</td>
<td>Threonine</td>
</tr>
</tbody>
</table>
We have 1000 British people in the study (which means that there are 2000 chromosomes). The genotypes for each gene are as follows:

\[
\begin{array}{ccc}
\text{Protein A :} & 298 & \text{AA} \\
489 & \text{AG} \\
213 & \text{GG} \\
\text{Protein B :} & 99 & \text{TT} \\
418 & \text{TC} \\
483 & \text{CC}
\end{array}
\]

The loci are individually in HW equilibrium.
We have 1000 British people in the study (which means that there are 2000 chromosomes). The genotypes for each gene are as follows:

<table>
<thead>
<tr>
<th>Protein A</th>
<th>Protein B</th>
</tr>
</thead>
<tbody>
<tr>
<td>298 AA</td>
<td>99 TT</td>
</tr>
<tr>
<td>489 AG</td>
<td>418 TC</td>
</tr>
<tr>
<td>213 GG</td>
<td>483 CC</td>
</tr>
</tbody>
</table>

The loci are individually in HW equilibrium.

Next, we can estimate the allele frequencies:

\[
P_A = \frac{2 \times 298 + 489}{2000} = .5425
\]

\[
q_a = \frac{489 + 2 \times 213}{2000} = .4575
\]

Similarly, you can find that

\[
P_B = .3080
\]

\[
q_b = .6920
\]
If the haplotypes are in Linkage Equilibrium, then the probability of each haplotype will be \( p_Ap_B, p_Aq_b, q_ap_B, \) and \( q_aq_b \) respectively.

<table>
<thead>
<tr>
<th>HAPLOTYPE</th>
<th>OBSERVED</th>
<th>EXPECTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>474</td>
<td>((.5425)(.3080)(2000) = 334.2)</td>
</tr>
<tr>
<td>AC</td>
<td>611</td>
<td>((.5425)(.6920)(2000) = 750.8)</td>
</tr>
<tr>
<td>GT</td>
<td>142</td>
<td>((.4575)(.3080)(2000) = 281.8)</td>
</tr>
<tr>
<td>GC</td>
<td>773</td>
<td>((.4575)(.6920)(2000) = 633.2)</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \] where the degrees of freedom = 

#categories-1-#other dependencies, in this case is 4-1-2 = 1.

\( \chi^2 = 184.7 \) with 1 d.f., which yields a P-value of \( \ll 0.0001 \), so we can safely REJECT Linkage Equilibrium between the two SNPs.
What about the Linkage Disequilibrium??

\[ \hat{P}_{AB} = \frac{474}{2000} = 0.2370 \]

\[ \hat{P}_{Ab} = \frac{611}{2000} = 0.3055 \]

\[ \hat{P}_{aB} = \frac{142}{2000} = 0.0710 \]

\[ \hat{P}_{ab} = \frac{773}{2000} = 0.3865 \]

\[ D = \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{aB}\hat{P}_{Ab} = 0.07 \]

\[ D' = \frac{D}{D_{\text{max}}} = \frac{0.07}{0.14} = 50\% \]

This means that the LD is 50% of its theoretical maximum!
What about the Linkage Disequilibrium??
How much LD exists between the two loci??

\[
\hat{P}_{AB} = \frac{474}{2000} = 0.2370 \\
\hat{P}_{Ab} = \frac{611}{2000} = 0.3055 \\
\hat{P}_{aB} = \frac{142}{2000} = 0.0710 \\
\hat{P}_{ab} = \frac{773}{2000} = 0.3865
\]

- \( D = \hat{P}_{AB} \hat{P}_{ab} - \hat{P}_{aB} \hat{P}_{Ab} = 0.07 \)
- \( D_{max} = \min\{p_A q_b, q_a p_B\} = \min\{0.38, 0.14\} = 0.14 \)
- So \( D' = \frac{D}{D_{max}} = \frac{0.07}{0.14} = 50\% \). This means that the LD is 50% of its theoretical maximum!
Summary: We reject Linkage Equilibrium by the $\chi^2$ test, so that means that LD exists. How much LD? 50% of the theoretical maximum.
Summary: We reject Linkage Equilibrium by the $\chi^2$ test, so that means that LD exists. How much LD? 50% of the theoretical maximum.

Other Measures of LD

$$D' = \begin{cases} \frac{D}{D_{\text{max}}} & D > 0 \\ \frac{D}{D_{\text{min}}} & D < 0 \end{cases}$$

$$r^2 = \frac{D}{pAp_ap_Bp_b}$$

- $r^2$ is the correlation coefficient of the frequencies. It has the convenient property that $\chi^2 = r^2 N$, where N is the number of chromosomes in the sample (see the lecture on Introduction to $r^2$ for a proof).
Summary: We reject Linkage Equilibrium by the $\chi^2$ test, so that means that LD exists. How much LD? 50% of the theoretical maximum.

Other Measures of LD

\[ D' = \begin{cases} \frac{D}{D_{\text{max}}}, & D > 0 \\ \frac{D}{D_{\text{min}}}, & D < 0 \end{cases} \]

\[ r^2 = \frac{D}{\rho_A \rho_a \rho_B \rho_b} \]

- $r^2$ is the correlation coefficient of the frequencies. It has the convenient property that $\chi^2 = r^2 N$, where $N$ is the number of chromosomes in the sample (see the lecture on Introduction to $r^2$ for a proof).
$D'$ and $r^2$ are the most important measures of LD. $r^2$ is the favorite of the HapMap project.
The case $D' = 1$ is called *Complete LD*.
Definition

The case $D' = 1$ is called *Complete LD*.

Intuition for Complete LD: two SNPs are not separated by recombination. In this case, there are **at most** 3 of the 4 possible haplotypes present in the population.
**Definition**

The case \( D' = 1 \) is called *Complete LD*.

Intuition for Complete LD: two SNPs are not separated by recombination. In this case, there are at most 3 of the 4 possible haplotypes present in the population.

**Definition**

The case \( r^2 = 1 \) is called *Perfect LD*.

The case of perfect LD happens if and only if the two SNPs have not been separated by recombination, but also have the same allele frequencies.