Introduction to different measures of linkage disequilibrium (LD) and their calculation

By

Dr. M. Awais Khan
University of Illinois, Urbana-Champaign
Calculation of linkage disequilibrium

To understand the calculation of linkage disequilibrium consider following example
Suppose there are two genes on Chromosome 5 of apple, each with two alleles

```
ACTGGGATAT.................GATCAACCCAG
ACTCGGTAT.................GATCAACCCAG
ACTCGGTAT.................GATCATCCAG
```

**SNP1**

**SNP2**

Showing only alleles for both SNPs

<table>
<thead>
<tr>
<th>Alleles</th>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele 1</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>Allele 2</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>
Steps in LD calculation

For better understanding of LD calculation, it is divided into five steps

Step 1) Calculate allele frequencies

If \( p_1 \) and \( p_2 \) = frequency of the alleles at SNP1 and \( q_1 \) and \( q_2 \) = frequency of the alleles at SNP2 then in tabular form it could be written as follows

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>Frequency</td>
</tr>
<tr>
<td>G</td>
<td>( p_1 )</td>
</tr>
<tr>
<td>C</td>
<td>( p_2 )</td>
</tr>
</tbody>
</table>
Step 2) Calculate haplotype frequencies

From our example of two SNPs each with two alleles all possible haplotypes are

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>A</td>
</tr>
<tr>
<td>SNP1</td>
<td>GA</td>
</tr>
<tr>
<td>SNP1</td>
<td>CA</td>
</tr>
</tbody>
</table>

Suppose haplotype frequencies are as follows

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Frequency</th>
<th>Haplotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA</td>
<td>p11</td>
<td>GT</td>
<td>p12</td>
</tr>
<tr>
<td>CA</td>
<td>p21</td>
<td>CT</td>
<td>q22</td>
</tr>
</tbody>
</table>
Step 3) Linkage equilibrium

When haplotype frequencies are equal to the product of their corresponding allele frequencies, it means the loci are in linkage equilibrium.

<table>
<thead>
<tr>
<th>Haplotype frequency</th>
<th>Product of allelic frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>p11</td>
<td>= p1q1</td>
</tr>
<tr>
<td>p12</td>
<td>= p1q2</td>
</tr>
<tr>
<td>p21</td>
<td>= p2q1</td>
</tr>
<tr>
<td>p22</td>
<td>= p2q2</td>
</tr>
</tbody>
</table>
Step 4) Linkage disequilibrium

We can deduce linkage disequilibrium for each haplotype as the deviation of observed haplotype frequency from its corresponding allelic frequencies expected under equilibrium.

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

After solving above for D, we get as follows:

\[
D_{11} = p_1 q_1
\]
\[
D_{12} = p_1 q_2
\]
\[
D_{21} = p_2 q_1
\]
\[
D_{22} = p_2 q_2
\]
Step 5) Calculation of Linkage disequilibrium measure $D$

Commonly used measure of linkage disequilibrium, $D$ equals to $p_{11}p_{22}$ $-$ $p_{12}p_{21}$ and we can prove it by solving the four equations from previous slide

\begin{align*}
a) \quad p_{11}p_{22} &= (p_1 q_1 + D)(p_2 q_2 + D) \\ &= p_1q_1p_2q_2 + p_1q_1D + p_2q_2D + D^2 \\
b) \quad p_{12}p_{21} &= (p_1 q_2 - D)(p_2 q_1 - D) \\ &= p_1q_1p_2q_2 - p_2q_1D - p_1q_2D + D^2 \\
c) \quad \text{Subtracting } (p_1q_1p_2q_2 + p_1q_1D + p_2q_2D + D^2) - (p_1q_1p_2q_2 - p_2q_1D - p_1q_2D + D^2) \\
\quad 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) \\
\quad &= D \times (1) = D
\end{align*}
Estimate of $D$ in case of Linkage Equilibrium
If allele frequencies of $p_1$ and $q_1$ are both 0.5 and equilibrium occurs (only $Ab$ and $aB$ exist in the population)

\[
P_{11} = p_1q_1 = 0.5 \times 0.5 = 0.25
\]
\[
P_{22} = p_2q_2 = 0.5 \times 0.5 = 0.25
\]
\[
P_{12} = p_1q_2 = 0.5 \times 0.5 = 0.25
\]
\[
P_{21} = p_2q_1 = 0.5 \times 0.5 = 0.25
\]
\[
D = (P_{11})(P_{22}) - (P_{12})(P_{21})
\]
\[
D = (0.25)(0.25) - (0.25)(0.25) = 0
\]
Estimate of D in case of Linkage Disequilibrium
If allele frequency of p1 and q1 are both 0.5 and there is complete non-random association (only AB and ab exist in the population) with equal allele frequencies at all loci

\[ P_{11} = p_1q_1 + D = 0.25 + D = 0.5 \]
\[ P_{22} = p_2q_2 + D = 0.25 + D = 0.5 \]
\[ P_{12} = p_1q_2 - D = 0.25 - D = 0 \]
\[ P_{21} = p_2q_1 - D = 0.25 - D = 0 \]

\[ D = (P_{11})(P_{22}) - (P_{12})(P_{21}) \]
\[ D = (0.5)(0.5) - (0)(0) = 0.25 \]
Standardization of D

Sometimes, depending on allele frequency of two loci, the value of D can be negative, but actual gametic frequencies cannot be negative.

To overcome this issue, standardization methods have been proposed.
In a common standardization method, a relative measure of disequilibrium (D) compared to its maximum is used:

\[ D' = \frac{D}{D_{\text{max}}} \]

**When D is positive**
\[ D_{\text{max}} = \min \left[ (p_1q_2) \text{ or } (p_2q_1) \right] \]

**When D is negative**
\[ D_{\text{max}} = \min \left[ (p_1q_1) \text{ or } (p_2q_2) \right] \]

This standardization makes D-values range between 0 and 1.
Correlation coefficient as a measure of LD

Another commonly used measure to calculate LD between loci is Pearson coefficient of correlation (r)

\[ r = \frac{D}{\sqrt{p_1p_2q_1q_2}} \]

However, squared coefficient of correlation (\( r^2 \)) is often used to remove the arbitrary sign introduced
Testing significance of LD

To test if LD is statistically significant we can do a $\chi^2$ test

$$\chi^2 = \Sigma (\text{obs-exp})^2 / \text{exp}$$

expected is random associations between alleles

However, r can be conveniently used for chi-test, as

$$\chi^2 = r^2 N$$

where N is the number of chromosomes in the sample
Example

Let’s assume that we have genotypic data for the two SNPs with two alleles each (same example used to deduce the equations for different LD measures)

Genotypic data
GA = 474         GT = 611         CA = 142           CT = 773
Total = 2000

Calculation of haplotype and allele frequencies

Haplotype Frequencies
GA = 474 / 2000 = .2370
GT = 611 / 2000 = .3055
CA = 142 / 2000 = .0710
CT = 773 / 2000 = .3865

Allele frequencies
G = 0.542
C = 0.457
A = 0.308
T = 0.692
Input values in the equation for \( D \) to calculate linkage disequilibrium

\[
D = (P_{11} \, P_{22}) - (P_{12} \, P_{21})
\]

\[
D = (0.2370 \times 0.3865) - (0.3055 \times 0.0710) = 0.0699
\]

To estimate \( D_{\text{max}} \) input allelic frequencies and value for \( D \) in the following equation

\[
D_{\text{max}} = \min \left[ (p_1q_2) \text{ or } (p_2q_1) \right]
\]

\[
D_{\text{max}} = (0.5425 \times 0.692) = 0.375
\]

or \( = (0.4575 \times 0.308) = 0.141 \)
Now calculate D’ input value of D and Dmax calculated in previous step in the following equation

\[ D' = \frac{D}{D_{\text{max}}} \]

\[ D' = \frac{0.0699}{0.141} = 0.496 \approx 0.5 \]

To calculate coefficient of correlation (r), input value of D and allele frequencies calculated in previous steps in the following equation

\[ r = \frac{D}{(p_1 p_2 q_1 q_2)^{1/2}} \]

\[ r = \frac{0.0699}{(0.5425 \times 0.4575 \times 0.308 \times 0.692)^{1/2}} \]

\[ r = \frac{0.0699}{0.23} = 0.304 \]

\[ r^2 = (0.304)^2 = 0.092 \]
To check the significance of LD between loci use following equation

$$\chi^2 = r^2 N$$

$$\chi^2 = 0.092 \times 2000 = 184.8 \ (1\ df)$$

At 184.8 and df of 1, P-value is 0.0001

So, we can conclude based on our calculations that there is a significant LD between loci and it is 50% of the theoretical maximum

Also note that two SNPs are in complete LD (not separated by recombination) when $D' = 1$ or $r^2=1$
Reference: