Population Genetics

- Different copies of a gene are called alleles; for example $A$ and $a$ at gene $A$;
- These alleles form three genotypes, $AA$, $Aa$ and $aa$;
- The allele (or gene) frequency of an allele is defined as the proportion of this allele among a group of individuals;
- Accordingly, the genotype frequency is the proportion of a genotype among a group of individuals
Calculations of allele frequencies and genotype frequencies

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Counts</th>
<th>Estimates genotype frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>224</td>
<td>( P_{AA} = \frac{224}{294} = 0.762 )</td>
</tr>
<tr>
<td>Aa</td>
<td>64</td>
<td>( P_{Aa} = \frac{64}{294} = 0.218 )</td>
</tr>
<tr>
<td>aa</td>
<td>6</td>
<td>( P_{aa} = \frac{6}{294} = 0.020 )</td>
</tr>
<tr>
<td>Total</td>
<td>294</td>
<td>( P_{AA} + P_{Aa} + P_{aa} = 1 )</td>
</tr>
</tbody>
</table>

Allele frequencies

\( p_A = \frac{(2 \times 214 + 64)}{(2 \times 294)} = 0.871, \ p_a = \frac{(2 \times 6 + 64)}{(2 \times 294)} = 0.129 \),

\( p_A + p_a = 0.871 + 0.129 = 1 \)

Expected genotype frequencies

\( AA \quad p_A^2 = 0.871^2 = 0.769 \)

\( Aa \quad 2p_Ap_a = 2 \times 0.871 \times 0.129 = 0.224 \)

\( aa \quad p_a^2 = 0.129^2 = 0.017 \)
<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Counts</th>
<th>Estimates of genotype freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$n_{AA}$</td>
<td>$P_{AA} = n_{AA}/n$</td>
</tr>
<tr>
<td>Aa</td>
<td>$n_{Aa}$</td>
<td>$P_{Aa} = n_{Aa}/n$</td>
</tr>
<tr>
<td>aa</td>
<td>$n_{aa}$</td>
<td>$P_{aa} = n_{aa}/n$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
<td>$P_{AA} + P_{Aa} + P_{aa} = 1$</td>
</tr>
</tbody>
</table>

Allele frequencies

$p_A = (2n_{AA} + n_{Aa})/2n$
p_a = (2n_{aa} + n_{Aa})/2n$

Standard error of the estimate of the allele frequency

$\text{Var}(p_A) = p_A(1 - p_A)/2n$
The Hardy-Weinberg Law

- In the Hardy-Weinberg equilibrium (HWE), the relative frequencies of the genotypes will remain unchanged from generation to generation;
- As long as a population is randomly mating, the population can reach HWE from the second generation;
- The deviation from HWE, called Hardy-Weinberg disequilibrium (HWD), results from many factors, such as selection, mutation, admixture and population structure…
Mendelian inheritance at the individual level

1. Make a cross between two individual parents
2. Consider one gene (A) with two alleles A and a → AA, Aa, aa

Thus, we have a total of nine possible cross combinations:

<table>
<thead>
<tr>
<th>Cross</th>
<th>Mendelian segregation ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. AA × AA → AA</td>
<td></td>
</tr>
<tr>
<td>2. AA × Aa → ½AA + ½Aa</td>
<td></td>
</tr>
<tr>
<td>3. AA × aa → Aa</td>
<td></td>
</tr>
<tr>
<td>4. Aa × AA → ½AA + ½Aa</td>
<td></td>
</tr>
<tr>
<td>5. Aa × Aa → ¼AA + ½Aa + ¼aa</td>
<td></td>
</tr>
<tr>
<td>6. Aa × aa → ½Aa + ¼aa</td>
<td></td>
</tr>
<tr>
<td>7. aa × AA → Aa</td>
<td></td>
</tr>
<tr>
<td>8. aa × Aa → ½Aa + ½aa</td>
<td></td>
</tr>
<tr>
<td>9. aa × aa → aa</td>
<td></td>
</tr>
</tbody>
</table>
Mendelian inheritance at the population level

- A population, a group of individuals, may contain all these nine combinations, weighted by the mating frequencies.
- Genotype frequencies: $AA, P_{AA}$; $Aa, P_{Aa}$; $aa, P_{aa}$

<table>
<thead>
<tr>
<th>Cross</th>
<th>Mating freq. ($t$)</th>
<th>Mendelian segreg. ratio ($t+1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$AA$</td>
</tr>
<tr>
<td>1.</td>
<td>$AA \times AA$</td>
<td>$P_{AA}(t)P_{AA}(t)$ → 1</td>
</tr>
<tr>
<td>2.</td>
<td>$AA \times Aa$</td>
<td>$P_{AA}(t)P_{Aa}(t)$ → $\frac{1}{2}$</td>
</tr>
<tr>
<td>3.</td>
<td>$AA \times aa$</td>
<td>$P_{AA}(t)P_{aa}(t)$ → 0</td>
</tr>
<tr>
<td>4.</td>
<td>$Aa \times AA$</td>
<td>$P_{Aa}(t)P_{AA}(t)$ → $\frac{1}{2}$</td>
</tr>
<tr>
<td>5.</td>
<td>$Aa \times Aa$</td>
<td>$P_{Aa}(t)P_{Aa}(t)$ → $\frac{1}{4}$</td>
</tr>
<tr>
<td>6.</td>
<td>$Aa \times aa$</td>
<td>$P_{Aa}(t)P_{aa}(t)$ → 0</td>
</tr>
<tr>
<td>7.</td>
<td>$aa \times AA$</td>
<td>$P_{aa}(t)P_{AA}(t)$ → 0</td>
</tr>
<tr>
<td>8.</td>
<td>$aa \times Aa$</td>
<td>$P_{aa}(t)P_{Aa}(t)$ → 0</td>
</tr>
<tr>
<td>9.</td>
<td>$aa \times aa$</td>
<td>$P_{aa}(t)P_{aa}(t)$ → 0</td>
</tr>
</tbody>
</table>
\[ P_{AA}(t+1) = 1[P_{AA}(t)]^2 + \frac{1}{2} 2[P_{AA}(t)P_{Aa}(t)] + \frac{1}{4}[P_{Aa}(t)]^2 \\
= [P_{AA}(t) + \frac{1}{2}P_{Aa}(t)]^2 \]

Similarly, we have

\[ P_{aa}(t+1) = [P_{aa}(t) + \frac{1}{2}P_{Aa}(t)]^2 \]
\[ P_{Aa}(t+1) = 2[P_{AA}(t) + \frac{1}{2}P_{Aa}(t)][P_{aa}(t) + \frac{1}{2}P_{Aa}(t)] \]

Therefore, we have

\[ [P_{Aa}(t+1)]^2 = 4P_{AA}(t+1)P_{aa}(t+1) \]

Furthermore, if random mating continues, we have

\[ P_{AA}(t+2) = [P_{AA}(t+1) + \frac{1}{2}P_{Aa}(t+1)]^2 = P_{AA}(t+1) \]
\[ P_{Aa}(t+2) = 2[P_{AA}(t+1) + \frac{1}{2}P_{Aa}(t+1)][P_{aa}(t+1) + \frac{1}{2}P_{Aa}(t+1)] = P_{Aa}(t+1) \]
\[ P_{aa}(t+2) = [P_{aa}(t+1) + \frac{1}{2}P_{Aa}(t+1)]^2 = P_{aa}(t+1) \]
Concluding remarks

A population with \([P_{AA}(t+1)]^2 = 4P_{AA}(t+1)P_{aa}(t+1)\) is said to be in Hardy-Weinberg equilibrium (HWE). The HWE population has the following properties:

1. Genotype (and allele) frequencies are constant from generation to generation,
2. Genotype frequencies = the product of the allele frequencies, i.e.,
   - \(P_{AA} = p_A^2\)
   - \(P_{Aa} = 2p_Ap_a\)
   - \(P_{aa} = p_a^2\)

For a population at Hardy-Weinberg disequilibrium (HWD), we have

- \(P_{AA} = p_A^2 + D\)
- \(P_{Aa} = 2p_Ap_a - 2D\)
- \(P_{aa} = p_a^2 + D\)

The magnitude of \(D\) determines the degree of HWD.

- \(D = 0\) means that there is no HWD.
- \(D\) has a range of \(\max(-p_A^2, -p_a^2) \leq D \leq p_Ap_a\)
Chi-square test for HWE

• Whether or not the population deviates from HWE at a particular locus can be tested using a chi-square test.

• If the population deviates from HWE (i.e., Hardy-Weinberg disequilibrium, HWD), this implies that the population is not randomly mating. Many evolutionary forces, such as mutation, genetic drift and population structure, may operate.
Example 1

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>224</td>
<td>64</td>
<td>6</td>
<td>294</td>
</tr>
<tr>
<td>Exp</td>
<td>(n(p_A^2) = 222.9)</td>
<td>(n(2p_Ap_a) = 66.2)</td>
<td>(n(p_a^2) = 4.9)</td>
<td>294</td>
</tr>
</tbody>
</table>

Test statistics

\[ \chi^2 = \sum (\text{obs} - \text{exp})^2 / \text{exp} \]

\[ = \frac{(224-222.9)^2}{222.9} + \frac{(64-66.2)^2}{66.2} + \frac{(6-4.9)^2}{4.9} = 0.32 \]

is less than

\[ \chi^2_{df=2} (\alpha = 0.05) = 5.99 \]

Therefore, the population does not deviate from HWE at this locus.
### Example 2

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>234</td>
<td>36</td>
<td>6</td>
<td>276</td>
</tr>
<tr>
<td>Exp</td>
<td>n(p_A^2)</td>
<td>n(2p_Ap_a)</td>
<td>n(p_a^2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>230.1</td>
<td>43.8</td>
<td>2.1</td>
<td>276</td>
</tr>
</tbody>
</table>

Test statistics

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

\[ = (234-230.1)^2/230.1+(36-43.8)^2/43.8+(6-2.1)^2/2.1 \]

\[ = 8.8 \]

is greater than \[ \chi^2_{df=2} (\alpha = 0.05) = 5.99 \]

Therefore, the population deviates from HWE at this locus.
Linkage disequilibrium

- Consider two loci, A and B, with alleles A, a and B, b, respectively, in a population.
- Assume that the population is at HWE.
- If the population is at Hardy-Weinberg equilibrium, we have

\[
\begin{align*}
\text{Gene A} & \\
\text{AA: } P_{AA} &= p_A^2 \\
\text{Aa: } P_{Aa} &= 2p_A p_a \\
\text{aa: } P_{aa} &= p_a^2 \\
\end{align*}
\]

\[
\begin{align*}
\text{Gene B} & \\
\text{BB: } P_{BB} &= p_B^2 \\
\text{Bb: } P_{Bb} &= 2p_B p_b \\
\text{bb: } P_{bb} &= p_b^2 \\
\end{align*}
\]

\[
\begin{align*}
P_{AA} + P_{Aa} + P_{aa} &= 1 \\
p_A + p_a &= 1 \\
\end{align*}
\]

\[
\begin{align*}
P_{BB} + P_{Bb} + P_{bb} &= 1 \\
p_B + p_b &= 1 \\
\end{align*}
\]
But the population is at Linkage Disequilibrium (for a pair of loci). Then we have

- Two-gene haplotype AB: \( p_{AB} = p_A p_B + D_{AB} \)
- Two-gene haplotype Ab: \( p_{Ab} = p_A p_b + D_{Ab} \)
- Two-gene haplotype aB: \( p_{aB} = p_a p_B + D_{aB} \)
- Two-gene haplotype ab: \( p_{ab} = p_a p_b + D_{ab} \)

\[ p_{AB} + p_{Ab} + p_{aB} + p_{ab} = 1 \]

\( D_{ij} \) is the coefficient of linkage disequilibrium (LD) between the two genes in the population. The magnitude of \( D \) reflects the degree of LD. The larger \( D \), the stronger LD.
\( p_A = p_{AB} + p_{Ab} \)
\[ = p_A p_B + D_{AB} + p_A p_b + D_{Ab} \]
\[ = p_A + D_{AB} + D_{Ab} \quad \Rightarrow \quad D_{AB} = -D_{Ab} \]
\( p_B = p_{AB} + p_{aB} \)
\[ = p_B + D_{AB} + D_{aB} \quad \Rightarrow \quad D_{AB} = -D_{aB} \]
\( p_b = p_{Ab} + p_{ab} \)
\[ = p_b + D_{aB} + D_{ab} \quad \Rightarrow \quad D_{ab} = -D_{aB} \]

Finally, we have \( D_{AB} = -D_{Ab} = -D_{aB} = D_{ab} = D. \)

Re-write four two-gene haplotypes

- AB: \( p_{AB} = p_A p_B + D \)
- Ab: \( p_{Ab} = p_A p_b - D \)
- aB: \( p_{aB} = p_a p_B - D \)
- ab: \( p_{ab} = p_a p_b + D \)

\[ D = p_{AB} p_{ab} - p_{Ab} p_{aB} \]

\( D = 0 \quad \rightarrow \quad \) the population is at the linkage equilibrium
How does D transmit from one generation (1) to the next (2)?

\[ D(2) = (1-r)^1 \cdot D(1) \]

\[ \ldots \]

\[ D(t+1) = (1-r)^t \cdot D(1) \]

\[ t \uparrow, \quad D(t+1) \uparrow \Rightarrow r \downarrow \]
Proof to $D(t+1) = (1-r)^t D(t)$

• The four gametes randomly unite to form a zygote. The proportion $1-r$ of the gametes produced by this zygote are parental (or nonrecombinant) gametes and fraction $r$ are nonparental (or recombinant) gametes. A particular gamete, say $AB$, has a proportion $(1-r)$ in generation $t+1$ produced without recombination. The frequency with which this gamete is produced in this way is $(1-r)p_{AB}(t)$.

• Also this gamete is generated as a recombinant from the genotypes formed by the gametes containing allele $A$ and the gametes containing allele $B$. The frequencies of the gametes containing alleles $A$ or $B$ are $p_A(t)$ and $p_B(t)$, respectively. So the frequency with which $AB$ arises in this way is $rp_A(t)p_B(t)$.

• Therefore the frequency of $AB$ in the generation $t+1$ is
  
  $$p_{AB}(t+1) = (1-r)p_{AB}(t) + rp_A(t)p_B(t)$$

By subtracting is $p_A(t)p_B(t)$ from both sides of the above equation, we have

$$D(t+1) = (1-r)^t D(t)$$

Whence

$$D(t+1) = (1-r)^t D(1)$$
Estimate and test for LD

Two markers A and B:

Four haplotypes

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>$p_{AB}$</td>
</tr>
<tr>
<td>Ab</td>
<td>$p_{Ab}$</td>
</tr>
<tr>
<td>aB</td>
<td>$p_{aB}$</td>
</tr>
<tr>
<td>ab</td>
<td>$p_{ab}$</td>
</tr>
</tbody>
</table>
Data Structure and expected genotype frequencies (assuming a random mating)

<table>
<thead>
<tr>
<th></th>
<th>BB (P_{BB})</th>
<th>Bb (P_{Bb})</th>
<th>bb (P_{bb})</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA (P_{AA})</td>
<td>n_{22} \quad p_{AB}^2</td>
<td>n_{21} \quad 2p_{AB}p_{Ab}</td>
<td>n_{20} \quad p_{Ab}^2</td>
</tr>
<tr>
<td>Aa (P_{Aa})</td>
<td>n_{12} \quad 2p_{AB}p_{aB}</td>
<td>n_{11} \quad 2(p_{AB}p_{ab}+p_{Ab}p_{AB})</td>
<td>n_{10} \quad 2p_{Ab}p_{ab}</td>
</tr>
<tr>
<td>aa (P_{aa})</td>
<td>n_{02} \quad p_{aB}^2</td>
<td>n_{01} \quad 2p_{aB}p_{ab}</td>
<td>n_{00} \quad p_{ab}^2</td>
</tr>
</tbody>
</table>

Multinomial pdf

H_1: D \neq 0
\[ \log f(p_{ij}|n) = \log n!/(n_{22}!...n_{00}!) \\
+ n_{22} \log p_{AB}^2 + n_{21} \log (2p_{AB}p_{Ab}) + n_{20} \log p_{Ab}^2 \\
+ ... \]
Estimate \( p_{AB}, p_{Ab}, p_{aB} \) (\( p_{ab} = 1-p_{AB}-p_{Ab}-p_{aB} \)) \( \rightarrow p_A, p_B, D \)

H_0: D = 0
\[ \log f(p_{ij}|n) = \log n!/(n_{22}!...n_{00}!) \\
+ n_{22} \log (p_{AB}p_{B})^2 + n_{21} \log (2p_{A}^2p_{B}p_{b}) + n_{20} \log (p_{A}p_{b})^2 \\
+ ... \]
Estimate \( p_A \) and \( p_B \).
<table>
<thead>
<tr>
<th>H1</th>
<th>BB (P_{BB})</th>
<th>Bb (P_{Bb})</th>
<th>bb (P_{bb})</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA (P_{AA})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freq</td>
<td>$p_{AB}^2$</td>
<td></td>
<td>$p_{AB}^2$</td>
</tr>
<tr>
<td>Obs</td>
<td>$n_{22}$</td>
<td></td>
<td>$n_{21}$</td>
</tr>
<tr>
<td>$#p_{AB}$</td>
<td>2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$#p_{Ab}$</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$#p_{aB}$</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$#p_{ab}$</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Aa (P_{Aa})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freq</td>
<td>$2p_{AB}p_{aB}$</td>
<td></td>
<td>$2p_{AB}p_{ab}$</td>
</tr>
<tr>
<td>Obs</td>
<td>$n_{12}$</td>
<td></td>
<td>$n_{11}$</td>
</tr>
<tr>
<td>$#p_{AB}$</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$#p_{Ab}$</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$#p_{aB}$</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$#p_{ab}$</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>aa (P_{aa})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freq</td>
<td>$p_{aB}^2$</td>
<td></td>
<td>$p_{aB}^2$</td>
</tr>
<tr>
<td>Obs</td>
<td>$n_{02}$</td>
<td></td>
<td>$n_{01}$</td>
</tr>
<tr>
<td>$#p_{AB}$</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$#p_{Ab}$</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$#p_{aB}$</td>
<td>2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$#p_{ab}$</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
The estimator of haplotype frequencies

\[
p_{AB} = \frac{1}{2n}[2n_{22} + (n_{21} + n_{12}) + \phi n_{11}]
\]  
(1)

\[
p_{Ab} = \frac{1}{2n}[2n_{20} + (n_{21} + n_{10}) + (1-\phi)n_{11}]
\]  
(2)

\[
p_{aB} = \frac{1}{2n}[2n_{02} + (n_{01} + n_{12}) + (1-\phi)n_{11}]
\]  
(3)

\[
p_{ab} = \frac{1}{2n}[2n_{00} + (n_{10} + n_{01}) + \phi n_{11}]
\]  
(4)

EM algorithm
E step: Calculate \(\phi\)
M step: Calculate \(p_{AB}, \ldots, p_{ab}\) using equations (1) - (4)

Specific estimating steps
1. Give initiate values for \(p_{AB}^{(1)} = p_{Ab}^{(1)} = p_{aB}^{(1)} = p_{a}^{(1)} = 0.5\);
3. Calculate \(\phi^{(1)} = (p_{AB}^{(1)} p_{ab}^{(1)}) / (p_{AB}^{(1)} p_{ab}^{(1)} + p_{Ab}^{(1)} p_{aB}^{(1)})\);
3. Calculate \(p_{AB}^{(2)}, p_{Ab}^{(2)}, p_{aB}^{(2)}, p_{ab}^{(2)}\);
4. Repeat steps 2 and 3 until the estimates of haplotype frequencies converge. The values at the convergence are the MLEs.
What is the convergence?

\[ |p_{AB}^{(t+1)} - p_{AB}^{(t)}| < \text{a very small value} \]

\[ |p_{Ab}^{(t+1)} - p_{Ab}^{(t)}| < \text{a very small value} \]

\[ |p_{aB}^{(t+1)} - p_{aB}^{(t)}| < \text{a very small value} \]

\[ |p_{ab}^{(t+1)} - p_{ab}^{(t)}| < \text{a very small value} \]

For example, this very small value is $e^{-8}$
<table>
<thead>
<tr>
<th>H0</th>
<th>Freq</th>
<th>BB ($P_{BB}$)</th>
<th>Bb ($P_{Bb}$)</th>
<th>bb ($P_{bb}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA ($P_{AA}$)</td>
<td></td>
<td>$p_A^2 p_B^2$</td>
<td>$2p_A^2 p_B p_b$</td>
<td>$p_A^2 p_b^2$</td>
</tr>
<tr>
<td>Obs</td>
<td></td>
<td>$n_{22}$</td>
<td>$n_{21}$</td>
<td>$n_{20}$</td>
</tr>
<tr>
<td>$#p_A$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$#p_a$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$#p_B$</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$#p_b$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Aa ($P_{Aa}$)</td>
<td></td>
<td>$2p_A p_a p_B^2$</td>
<td>$4p_A p_a p_B p_b$</td>
<td>$2p_A p_a p_b^2$</td>
</tr>
<tr>
<td>Obs</td>
<td></td>
<td>$n_{12}$</td>
<td>$n_{11}$</td>
<td>$n_{10}$</td>
</tr>
<tr>
<td>$#p_A$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$#p_a$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$#p_B$</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$#p_b$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>aa ($P_{aa}$)</td>
<td></td>
<td>$p_a^2 p_B^2$</td>
<td>$2p_a^2 p_B p_b$</td>
<td>$p_a^2 p_b^2$</td>
</tr>
<tr>
<td>Obs</td>
<td></td>
<td>$n_{02}$</td>
<td>$n_{01}$</td>
<td>$n_{00}$</td>
</tr>
<tr>
<td>$#p_A$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$#p_a$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$#p_B$</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$#p_b$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
The estimator of allele frequencies:

\[ p_A = \frac{1}{2n} \left[ 2(n_{22} + n_{21} + n_{20}) + (n_{12} + n_{11} + n_{10}) \right] \]
\[ p_a = \frac{1}{2n} \left[ 2(n_{02} + n_{01} + n_{00}) + (n_{12} + n_{11} + n_{10}) \right] \]
\[ p_B = \frac{1}{2n} \left[ 2(n_{22} + n_{12} + n_{02}) + (n_{21} + n_{11} + n_{01}) \right] \]
\[ p_b = \frac{1}{2n} \left[ 2(n_{20} + n_{10} + n_{00}) + (n_{21} + n_{11} + n_{01}) \right] \]

No EM algorithm is needed to obtain the MLEs of these allele frequencies
Plugging in the MLEs into the likelihood functions under the two hypotheses:

\( H_1: D \neq 0 \)
\[
\log L_1 = \log \frac{n!}{(n_{22}! \ldots n_{00}!)} + n_{22} \log p_{AB}^2 + n_{21} \log (2p_{AB}p_{Ab}) + n_{20} \log p_{Ab}^2 \\
\quad + \ldots
\]

\( H_0: D = 0 \)
\[
\log L_0 = \log \frac{n!}{(n_{22}! \ldots n_{00}!)} + n_{22} \log (p_A p_B)^2 + n_{21} \log (2p_A^2 p_B p_b) \\
\quad + n_{20} \log (p_A p_b)^2 \\
\quad + \ldots
\]

\[
LR = -2(\log L_0 - \log L_1) \\
\sim \chi^2_{(df=1, 0.05)} = 3.841
\]
Other approaches for testing LD: Chi-square Test

Test statistic

\[ \chi^2 = \frac{2nD^2}{(p_A p_a p_B p_b)} \]

is compared with the critical threshold value obtained from the chi-square table \( \chi^2_{df=1} (0.05) \). \( n \) is the number of individuals in the population.

If \( \chi^2 < \chi^2_{df=1} (0.05) \), this means that \( D \) is not significantly different from zero and that the population under study is in linkage equilibrium.

If \( \chi^2 > \chi^2_{df=1} (0.05) \), this means that \( D \) is significantly different from zero and that the population under study is in linkage disequilibrium.
Example

(1) Two genes A with allele A and a, B with alleles B and b, whose population frequencies are denoted by \(p_A\), \(p_a (=1- p_A)\) and \(p_B\), \(p_b (=1- p_b)\), respectively

(2) These two genes are associated with each other, having the coefficient of linkage disequilibrium D

Four gametes are observed as follows:

<table>
<thead>
<tr>
<th>Gamete</th>
<th>AB</th>
<th>Ab</th>
<th>aB</th>
<th>ab</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>474</td>
<td>611</td>
<td>142</td>
<td>773</td>
<td>2n=2000</td>
</tr>
<tr>
<td>Gamete frequency</td>
<td>(p_{AB})</td>
<td>(p_{Ab})</td>
<td>(p_{aB})</td>
<td>(p_{ab})</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(=474/2000)</td>
<td>(=611/2000)</td>
<td>(=142/2000)</td>
<td>(=773/2000)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>=0.237</td>
<td>=0.305</td>
<td>=0.071</td>
<td>=0.386</td>
<td></td>
</tr>
</tbody>
</table>
Estimates of allele frequencies

\[ p_A = p_{AB} + p_{Ab} = 0.237 + 0.305 = 0.542 \]
\[ p_a = p_{aB} + p_{ab} = 0.071 + 0.386 = 0.458 \]
\[ p_B = p_{AB} + p_{aB} = 0.237 + 0.071 = 0.308 \]
\[ p_b = p_{Ab} + p_{ab} = 0.305 + 0.386 = 0.692 \]

The estimate of D

\[ D = p_{AB}p_{ab} - p_{Ab}p_{aB} = 0.237 \times 0.386 - 0.305 \times 0.071 = 0.0699 \]

Test statistics

\[ \chi^2 = 2nD^2/ (p_A p_a p_B p_b) = 2 \times 1000 \times 0.0699^2/ (0.542 \times 0.458 \times 0.308 \times 0.692) = 184.78 \] is greater than \[ \chi^2_{df=3} (0.05) = 7.81. \]

Therefore, the population is in linkage disequilibrium at these two genes under consideration.
$\chi^2$ can also be calculated by

<table>
<thead>
<tr>
<th>Gamete</th>
<th>AB</th>
<th>Ab</th>
<th>aB</th>
<th>ab</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>474</td>
<td>611</td>
<td>142</td>
<td>773</td>
<td>2n=2000</td>
</tr>
<tr>
<td>Exp</td>
<td>$2n(p_Ap_B)$</td>
<td>$2n(p_Ap_b)$</td>
<td>$2n(p_ap_B)$</td>
<td>$2n(p_ap_b)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>=334.2</td>
<td>=750.8</td>
<td>=281.8</td>
<td>=633.2</td>
<td>2000</td>
</tr>
</tbody>
</table>

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

$= (474-334.2)^2/334.2 + (611-750.8)^2/750.8 + (142-281.8)^2/281.8 + (773-633.2)^2/633.2$

$= 184.78$

$= 2nD^2/ (p_Ap_ap_bp_b)$
Measures of linkage disequilibrium

(1) $D$, which has a limitation that its value depends on the allele frequencies

$D = 0.02$ is considered to be

- large for two genes each with diverse allele frequencies, e.g., $p_A = p_B = 0.9$ vs. $p_a = p_b = 0.1$
- small for two genes each with similar allele frequencies, e.g., $p_A = p_B = 0.5$ vs. $p_a = p_b = 0.5$
To make a comparison between gene pairs with different allele frequencies, we need a new normalized measure.

The range of LD is

\[
\max(-p_A p_B, -p_a p_b) \leq D \leq \min(p_A p_b, p_a p_B)
\]

The normalized LD (Lewontin 1964) is defined as

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

where \(D_{\text{max}}\) is the maximum that \(D\) can have, which is

\[
D_{\text{max}} = \begin{cases} 
\max(-p_A p_B, -p_a p_b) & \text{if } D < 0, \\
\min(p_A p_b, p_a p_B) & \text{if } D > 0.
\end{cases}
\]

For the above example, we have \(D' = 0.0699/\min(p_a p_b, p_a p_B)\).
(3) Linkage disequilibrium measured as the correlation between the A and B alleles

\[
R = \frac{D}{\sqrt{p_A p_a p_B p_b}}, \quad r: [-1, 1]
\]

Note: \( \chi^2 = 2nR^2 \) follows the chi-square distribution with df = 3 under the null hypothesis of D = 0.

For the above example, we have

\[
R = 0.0699/\sqrt{p_A p_b p_a p_B} = 0.3040.
\]
Application of LD analysis

$D(t+1) = (1-r)^t D(1)$,

This means that when the population undergoes random mating, the LD decays exponentially in a proportion related to the recombination fraction.

(1) Population structure and evolution

Estimating $D$, $D'$ and $R \rightarrow$ the mating history of population

The larger the $D'$ and $R$ estimates, the more likely the population in nonrandom mating, the more likely the population to have a small size, the more likely the population to be affected by evolutionary forces.
Human origin studies based on LD analysis


LD curve for Swedish and Yoruban samples. To minimize ascertainment bias, data are only shown for marker comparisons involving the core SNP. Alleles are paired such that $D' > 0$ in the Utah population. $D' > 0$ in the other populations indicates the same direction of allelic association and $D' < 0$ indicates the opposite association. **a**, In Sweden, average $D'$ is nearly identical to the average $|D'|$ values up to 40-kb distances, and the overall curve has a similar shape to that of the Utah population (thin line in **a** and **b**). **b**, LD extends less far in the Yoruban sample, with most of the long-range LD coming from a single region, $HCF2$. Even at 5 kb, the average values of $|D'|$ and $D'$ diverge substantially. To make the comparisons between populations appropriate, the Utah LD curves are calculated solely on the basis of SNPs that had been successfully genotyped and met the minimum frequency criterion in both populations (Swedish and Yoruban) (Reich, et al. 2001)
(2) Fine mapping of disease genes

The detection of LD may imply that the recombination fraction between two genes is small and therefore closer (given the assumption that t is large).