EQUILIBRIA IN POPULATIONS
Basic Principles

- In a ‘stable’ population, the distribution of alleles obeys certain laws
  - Not really, and the deviations are interesting
- HW Equilibrium
  - (due to mixing in a population)
- Linkage (dis)−equilibrium
  - Due to recombination
Hardy-Weinberg principle

Given:

- Population of diploid individuals and a locus with alleles, A & a
- 3 Genotypes: AA, Aa, aa

Q: Will the frequency of alleles and genotypes remain constant from generation to generation?
To the Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making...

..........

A little mathematics of the multiplication-table type is enough to show ....the condition for this is $q^2 = pr$. And since $q_1^2 = p_1r_1$, whatever the values of $p$, $q$, and $r$ may be, the distribution will in any case continue unchanged after the second generation.
Suppose, Pr(A)=p, and Pr(a)=1-p=q.

If certain assumptions are met
- Large, diploid, population
- Discrete generations
- Random mating
- No selection, ...

Then, in every generation:

- \(Pr(A) = p\)
- \(Pr(a) = q\)
- \(Pr(AA) = p^2\)
- \(Pr(Aa) = 2pq\)
- \(Pr(aa) = q^2\)
In the next generation

- $Pr(AA) = p^2$
- $Pr(Aa) = 2pq$
- $Pr(aa) = q^2$

- $Pr(A) = p$
- $Pr(a) = q$
Hardy Weinberg: Generalization

- Multiple alleles with frequencies
  - By HW,
    \[ \theta_1, \theta_2, \ldots, \theta_H \]
    \[
    \begin{align*}
    \text{Pr[homozygous genotype } i \text{]} &= \theta_i^2 \\
    \text{Pr[heterozygous genotype } i, j \text{]} &= 2\theta_i\theta_j
    \end{align*}
    \]

- Multiple loci?
Hardy Weinberg: Implications

- The allele frequency does not change from generation to generation. True or false?
- It is observed that 1 in 10,000 caucasians have the disease phenylketonuria. The disease mutation(s) are all recessive. What fraction of the population carries the mutation?
- Males are 100 times more likely to have the ‘red’ type of color blindness than females. Why?
- Individuals homozygous for S have the sickle-cell disease. In an experiment, the ratios A/A:A/S: S/S were 9365:2993:29. Is HWE violated? Is there a reason for this violation?
- A group of individuals was chosen in NYC. Would you be surprised if HWE was violated?
- Conclusion: While the HW assumptions are rarely satisfied, the principle is still important as a baseline assumption, and significant deviations are interesting.
A modern example

- SNP-chips can give us the genotype at each site based on hybridization.
- Plot the 3 genotypes at each locus on 3 separate horizontal lines.
SNP-chips can give us the allelic value at each polymorphic site based on hybridization.

What is peculiar in the picture?

What is your conclusion?
Violation of HWE is common in nature

Non-HWE implies that some assumption is violated

Figuring out the violated assumption leads to biological insight
• HW equilibrium is the equilibrium in allele frequencies at a single locus.
• Linkage equilibrium refers to the equilibrium between allele occurrences at two loci.
  - LE is due to recombination events
  - First, we will consider the case where no recombination occurs.
Perfect phylogeny and phylogeography
Before you start,...

- SNP matrix
- The infinite sites assumption
- Recombination
- Genealogy/phylogeny
- Basic data structures and algorithms
The y-chr (mtDNA) lineage is a tree
We consider the directed (rooted) case. The root is all 0s, and all mutations are of the form 0→1
## Columns

<table>
<thead>
<tr>
<th>12345678</th>
<th>010000000</th>
<th>001101100</th>
<th>001101000</th>
<th>001100000</th>
<th>100000000</th>
<th>100010000</th>
<th>100010011</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>A: 0</td>
<td>B: 1</td>
<td>C: 1</td>
<td>D: 1</td>
<td>E: 0</td>
<td>F: 0</td>
<td>G: 0</td>
</tr>
</tbody>
</table>
Inclusion Property of Perfect Phylogeny

For any pair of columns $i,j$, one of the following holds
\[ i_1 \subseteq j_1 \]
\[ j_1 \subseteq i_1 \]
\[ i_1 \cap j_1 = \emptyset \]

For any pair of columns $i,j$
\[ i < j \text{ if and only if } i_1 \subseteq j_1 \]

Note that if $i<j$ then the edge containing $i$ is an ancestor of the edge containing $j$
Reconstruction of perfect phylogeny

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
## Sort columns

<table>
<thead>
<tr>
<th>Column</th>
<th>Binary 1</th>
<th>Binary 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>01000000</td>
<td>10000000</td>
</tr>
<tr>
<td>B</td>
<td>00110110</td>
<td>01111100</td>
</tr>
<tr>
<td>C</td>
<td>00110100</td>
<td>01110000</td>
</tr>
<tr>
<td>D</td>
<td>00110000</td>
<td>01100000</td>
</tr>
<tr>
<td>E</td>
<td>10000000</td>
<td>00000100</td>
</tr>
<tr>
<td>F</td>
<td>10001000</td>
<td>00000110</td>
</tr>
<tr>
<td>G</td>
<td>10001001</td>
<td>00000111</td>
</tr>
</tbody>
</table>
Add First column

<table>
<thead>
<tr>
<th></th>
<th>A:</th>
<th>B:</th>
<th>C:</th>
<th>D:</th>
<th>E:</th>
<th>F:</th>
<th>G:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100000000</td>
<td>011110000</td>
<td>011100000</td>
<td>011000000</td>
<td>000000100</td>
<td>000000110</td>
<td>000000111</td>
</tr>
</tbody>
</table>
Add First column

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>7</th>
<th>1</th>
<th>5</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Columns 3,4,5

2 3 4 6 7 1 5 8

A: 1 0 0 0 0 0 0 0
B: 0 1 1 1 1 0 0 0
C: 0 1 1 1 0 0 0 0
D: 0 1 1 0 0 0 0 0
E: 0 0 0 0 0 1 0 0
F: 0 0 0 0 0 1 1 0
G: 0 0 0 0 0 1 1 1
Columns

\[
\begin{array}{cccccccc}
2 & 3 & 4 & 6 & 7 & 1 & 5 & 8 \\
A: & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
B: & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
C: & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
D: & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
E: & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
F: & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
G: & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
\end{array}
\]
A perfect phylogeny

2 3 4 6 7 1 5 8
A: 1 0 0 0 0 0 0 0
B: 0 1 1 1 1 0 0 0
C: 0 1 1 1 0 0 0 0
D: 0 1 1 0 0 0 0 0
E: 0 0 0 0 0 1 0 0
F: 0 0 0 0 0 1 1 0
G: 0 0 0 0 0 1 1 1
Perfect phylogeny-- unrooted case
A perfect phylogeny

```
2 3 4 6 7 1 5 8
A: 1 0 0 0 0 0 0 0
B: 0 1 1 1 1 0 0 0
C: 0 1 1 1 0 0 0 0
D: 0 1 1 0 0 0 0 0
E: 0 0 0 0 0 1 0 0
F: 0 0 0 0 0 1 1 0
G: 0 0 0 0 0 1 1 1
```
0s and 1s can be reversed

A: 1 0 1 1 0 1 0 0
B: 0 1 0 0 1 1 0 0
C: 0 1 0 0 0 1 0 0
D: 0 1 0 1 0 1 0 0
E: 0 0 1 1 0 0 0 0
F: 0 0 1 1 0 0 1 0
G: 0 0 1 1 0 0 1 1
Unrooted case

- Switch the values in each column, so that 0 is the majority element.
- Apply the algorithm for the rooted case.
- Relabel columns and individuals to the original values.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Unrooted perfect phylogeny

- We transform matrix $M$ to a 0-major matrix $M_0$.
- If $M_0$ has a directed perfect phylogeny, $M$ has a perfect phylogeny.
- If $M$ has a perfect phylogeny, does $M_0$ have a directed perfect phylogeny?
Unrooted case

Theorem: If M has a perfect phylogeny, there exists a relabeling, and a perfect phylogeny s.t.

- Root is all 0s
- For any SNP (column), \( \#1s \leq \#0s \)
- All edges are mutated \( 0 \rightarrow 1 \)
Finding a ‘center’ of an unrooted phylogeny

- Is it possible to find a node or an edge so that none of the children have more than $n/2$ nodes?
Proof

- Consider the perfect phylogeny of M.
- Find the center:
- Root at the center, and direct all mutations from $0 \rightarrow 1$ away from the root. QED
- If the theorem is correct, then simply relabeling all columns so that the majority element is 0 is sufficient.
‘Homework’ Problems

- What if there is missing data? (An entry that can be 0 or 1)?
- What if recurrent mutations are allowed (infinite sites is violated)?
Linkage Disequilibrium
• Recall that a SNP data-set is a ‘binary’ matrix.
  - Rows are individual (chromosomes)
  - Columns are alleles at a specific locus

• Suppose you have 2 SNP datasets of a contiguous genomic region but no other information
  - One from an African population, and one from a European Population.
  - Can you tell which is which?
  - How long does the genomic region have to be?
Linkage (Dis)-equilibrium (LD)

- Consider sites A & B
- Case 1: No recombination
- Each new individual chromosome chooses a parent from the existing 'haplotype'

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Consider sites A & B

Case 2: diploidy and recombination

Each new individual chooses a parent from the existing alleles
Consider sites A & B

- **Case 1:** No recombination
  - Each new individual chooses a parent from the existing ‘haplotype’
    - $\text{Pr}[A,B=0,1] = 0.25$
      - Linkage disequilibrium
  - **Case 2:** Extensive recombination
    - Each new individual simply chooses an allele from either site
      - $\text{Pr}[A,B=(0,1)] = 0.125$
        - Linkage equilibrium

---

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
- In the absence of recombination,
  - Correlation between columns
  - The joint probability $\Pr[A=a, B=b]$ is different from $P(a)P(b)$
- With extensive recombination
  - $\Pr(a,b) = P(a)P(b)$
Measures of LD

- Consider two bi-allelic sites with alleles marked with 0 and 1
- Define
  - $P_{00} = \Pr[\text{Allele 0 in locus 1, and 0 in locus 2}]$
  - $P_{0*} = \Pr[\text{Allele 0 in locus 1}]$
- Linkage equilibrium if $P_{00} = P_{0*} P_{*0}$
- The D-measure of LD
  - $D = (P_{00} - P_{0*} P_{*0}) = -(P_{01} - P_{0*} P_{*1}) = ...$
Other measures of LD

- \( D' \) is obtained by dividing \( D \) by the largest possible value
  - Suppose \( D = (P_{00} - P_{0}P_{0}^*) > 0 \).
  - Then the maximum value of \( D_{\text{max}} = \min\{P_{0}P_{1}^*, P_{1}P_{0}^*\} \)
  - If \( D < 0 \), then maximum value is \( \max\{-P_{0}P_{0}^*, -P_{1}P_{1}^*\} \)
  - \( D' = D / D_{\text{max}} \)

```
<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>D</td>
<td>-D</td>
</tr>
<tr>
<td>1</td>
<td>-D</td>
<td>D</td>
</tr>
</tbody>
</table>
```

Site 1

Site 2

Vineet Bafna
Other measures of LD

- $D'$ is obtained by dividing $D$ by the largest possible value
  - Ex: $D' = \frac{\text{abs}(P_{11} - P_1* P_{1*})}{D_{\text{max}}}$
- $\rho = \frac{D}{(P_{1*} P_{0*} P_{1*} P_{0*})^{1/2}}$
- Let $N$ be the number of individuals
- Show that $\rho^2 N$ is the $\chi^2$ statistic between the two sites

\[
\begin{array}{cc}
0 & 1 \\
0 & P_{00N} & P_{0*N} \\
1 & & \\
\end{array}
\]
Digression: The $\chi^2$ test

The statistic

$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

behaves like a $\chi^2$ distribution (sum of squares of normal variables).

A p-value can be computed directly.
Observed and expected

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$P_{00}N$</td>
<td>$P_{01}N$</td>
</tr>
<tr>
<td>1</td>
<td>$P_{10}N$</td>
<td>$P_{11}N$</td>
</tr>
</tbody>
</table>

\[ \rho = \frac{D}{(P_{1*}P_{0*}P_{1*}P_{0*})^{1/2}} \]

- Verify that $\rho^2N$ is the $\chi^2$ statistic between the two sites
LD over time and distance

- The number of recombination events between two sites, can be assumed to be Poisson distributed.
- Let $r$ denote the recombination rate between two adjacent sites.
- $r = \#$ crossovers per bp per generation
- The recombination rate between two sites $l$ apart is $rl$
LD over time

- Decay in LD
  - Let $D^{(t)} = LD$ at time $t$ between two sites
  - $r' = lr$
  - $P^{(t)}_{00} = (1-r') P^{(t-1)}_{00} + r' P^{(t-1)}_{0*} P^{(t-1)}_{*0}$
  - $D^{(t)} = P^{(t)}_{00} - P^{(t)}_{0*} P^{(t)}_{*0} = P^{(t)}_{00} - P^{(t-1)}_{0*} P^{(t-1)}_{*0}$ (Why?)
  - $D^{(t)} = (1-r') D^{(t-1)} = (1-r')^t D^{(0)}$
LD over distance

- **Assumption**
  - Recombination rate increases linearly with distance and time
  - LD decays exponentially.
- The assumption is reasonable, but recombination rates vary from region to region, adding to complexity
- This simple fact is the basis of disease association mapping.
LD and disease mapping

- Consider a mutation that is causal for a disease.
- The goal of disease gene mapping is to discover which gene (locus) carries the mutation.
- Consider every polymorphism, and check:
  - There might be too many polymorphisms
  - Multiple mutations (even at a single locus) that lead to the same disease
- Instead, consider a dense sample of polymorphisms that span the genome
LD can be used to map disease genes

- LD decays with distance from the disease allele.
- By plotting LD, one can short list the region containing the disease gene.
- 269 individuals
  - 90 Yorubans
  - 90 Europeans (CEPH)
  - 44 Japanese
  - 45 Chinese
- ~1M SNPs
Haplotype blocks

• It was found that recombination rates vary across the genome
  - How can the recombination rate be measured?
• In regions with low recombination, you expect to see long haplotypes that are conserved. Why?
• Typically, haplotype blocks do not span recombination hot-spots
19q13

CSE280

Vineet Bafna
Long haplotypes

- Chr 2 region with high r² value (implies little/no recombination)
- History/Genealogy can be explained by a tree (a perfect phylogeny)
- Large haplotypes with high frequency
LD variation across populations

- LD is maintained upto 60kb in swedish population, 6kb in Yoruban population

Reich et al. Nature 411, 199-204 (10 May 2001)
Population specific recombination

- D’ was used as the measure between SNP pairs.
- SNP pairs were classified in one of the following
  - Strong LD
  - Strong evidence for recombination
  - Others (13% of cases)
- Plot shows fraction of pairs with strong recombination (low LD)
- This roughly favors out-of-africa. A Coalescent simulation can help give confidence values on this.

Gabriel et al., Science 2002
We described various population genetic concepts (HW, LD), and their applicability.

The values of these parameters depend critically upon the population assumptions.
- What if we do not have infinite populations
- No random mating (Ex: geographic isolation)
- Sudden growth
- Bottlenecks
- Ad-mixture

It would be nice to have a simulation of such a population to test various ideas. How would you do this simulation?