CSE182-L14

Population Genetics: Basics
Population Structure

- 377 locations (loci) were sampled in 1000 people from 52 populations.
- 6 genetic clusters were obtained, which corresponded to 5 geographic regions (Rosenberg et al. Science 2003)
Population Genetics

• What is it about our genetic makeup that makes us measurably different?
• These genetic differences are correlated with phenotypic differences
• With cost reduction in sequencing and genotyping technologies, we will know the sequence for entire populations of individuals.
• Here, we will study the basics of this polymorphism data, and tools that are being developed to analyze it.
What causes variation in a population?

- Mutations (may lead to SNPs)
- Recombinations
- Other genetic events (may lead to microsatellite repeats)
Single Nucleotide Polymorphisms

Infinite Sites Assumption: Each site mutates at most once

00000101011
10001101001
01000101010
01000000011
00011110000
00101100110
Short Tandem Repeats

GCTAGATCATCATCATCATCATTGCTAG
GCTAGATCATCATCATCATCATTGCTAGTTA
GCTAGATCATCATCATCATCATCATTGC
GCTAGATCATCATCATCATCATCATTGC
GCTAGATCATCATCATCATCATCATTGC
GCTAGATCATCATCATCATCATCATTGC
GCTAGATCATCATCATCATCATCATTGC
STR can be used as a DNA fingerprint

- Consider a collection of regions with variable length repeats.
- Variable length repeats will lead to variable length DNA.
- Vector of lengths is a fingerprint.
Recombination

Synapsis: Pairing of homologous chromosomes

Paternal

Maternal

Crossing over

00000000
11111111
00011111
What if there were no recombinations?

• Life would be simpler
• Each sequence would have a single parent
• The relationship is expressed as a tree.
The Infinite Sites Assumption

- The different sites are linked. A 1 in position 8 implies 0 in position 5, and vice versa.
- Some phenotypes could be linked to the polymorphisms
- Some of the linkage is “destroyed” by recombination
Infinite sites assumption and Perfect Phylogeny

- Each site is mutated at most once in the history.
- All descendants must carry the mutated value, and all others must carry the ancestral value.
Perfect Phylogeny

• Assume an evolutionary model in which no recombination takes place, only mutation.
• The evolutionary history is explained by a tree in which every mutation is on an edge of the tree. All the species in one sub-tree contain a 0, and all species in the other contain a 1. Such a tree is called a perfect phylogeny.
• How can one reconstruct such a tree?
The 4-gamete condition

• A column $i$ partitions the set of species into two sets $i_0$, and $i_1$
• A column is homogeneous w.r.t a set of species, if it has the same value for all species. Otherwise, it is heterogenous.
• EX: $i$ is heterogenous w.r.t \{A,D,E\}
4 Gamete Condition

- 4 Gamete Condition
  - There exists a perfect phylogeny if and only if for all pair of columns $(i,j)$, either $j$ is not heterogenous w.r.t $i_0$, or $i_1$.
  - Equivalent to
  - There exists a perfect phylogeny if and only if for all pairs of columns $(i,j)$, the following 4 rows do not exist $(0,0), (0,1), (1,0), (1,1)$
4-gamete condition: proof

- Depending on which edge the mutation $j$ occurs, either $i_0$, or $i_1$ should be homogenous.
- (only if) Every perfect phylogeny satisfies the 4-gamete condition.
- (if) If the 4-gamete condition is satisfied, does a prefect phylogeny exist?
An algorithm for constructing a perfect phylogeny

• We will consider the case where 0 is the ancestral state, and 1 is the mutated state. This will be fixed later.

• In any tree, each node (except the root) has a single parent.
  – It is sufficient to construct a parent for every node.

• In each step, we add a column and refine some of the nodes containing multiple children.

• Stop if all columns have been considered.
Inclusion Property

• For any pair of columns $i, j$
  - $i < j$ if and only if $i_1 \supseteq j_1$

• Note that if $i < j$ then the edge containing $i$ is an ancestor of the edge containing $i$
Initially, there is a single clade $r$, and each node has $r$ as its parent.
Sort columns

• Sort columns according to the inclusion property (note that the columns are already sorted here).

• This can be achieved by considering the columns as binary representations of numbers (most significant bit in row 1) and sorting in decreasing order.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Add first column

• In adding column $i$
  - Check each edge and decide which side you belong.
  - Finally add a node if you can resolve a clade
Adding other columns

- Add other columns on edges using the ordering property
Unrooted case

- Switch the values in each column, so that 0 is the majority element.
- Apply the algorithm for the rooted case
Handling recombination

- A tree is not sufficient as a sequence may have 2 parents
- Recombination leads to loss of correlation between columns
Linkage (Dis)-equilibrium (LD)

• Consider sites A & B
  • Case 1: No recombination
    - \( \text{Pr}[A,B=0,1] = 0.25 \)
      * Linkage disequilibrium
  • Case 2: Extensive recombination
    - \( \text{Pr}[A,B=(0,1)=0.125 \)
      * Linkage equilibrium