CSE 291: Advanced Topics in Computational Biology

Vineet Bafna/Pavel Pevzner

www.cse.ucsd.edu/classes/sp05/cse291
Topics

- Population Genetics
- Genome Duplication Problem
- Molecular Evolution
- Student Presentations
  - Critical overview of a field
  - Research Projects
Population Genetics

- Individuals in a species (population) are phenotypically different.
- Often these differences are inherited (genetic).
- Studying these differences is important!
- Q: How predictive are these differences?
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)
Genetic differences can predict ethnicity.
Scope of these lectures

- Basic terminology
- Key principles
  - HW equilibrium
  - Sources of variation
  - Linkage
  - Coalescent theory
  - Recombination/Ancestral Recombination Graph
  - Haplotypes/Haplotype phasing
  - Population sub-structure
  - Medical genetics basis: Association mapping/pedigree analysis
Alleles

- **Genotype:** genetic makeup of an individual
- **Allele:** A specific variant at a location
  - The notion of alleles predates the concept of gene, and DNA.
  - Initially, alleles referred to variants that described a measurable phenotype (round/wrinkled seed)
  - Now, an allele might be a nucleotide on a chromosome, with no measurable phenotype.
- Humans are diploid, they have 2 copies of each chromosome.
  - They may have heterozygosity/homozygosity at a location
  - Other organisms (plants) have higher forms of ploidy.
  - Additionally, some sites might have 2 allelic forms, or even many allelic forms.
Hardy Weinberg equilibrium

- Consider a locus with 2 alleles, A, a
- $p$ (respectively, $q$) is the frequency of $A$ (resp. $a$) in the population
- 3 Genotypes: $AA$, $Aa$, $aa$
- Q: What is the frequency of each genotype

If various assumptions are satisfied, (such as random mating, no natural selection), Then

- $P_{AA} = p^2$
- $P_{Aa} = 2pq$
- $P_{aa} = q^2$
Hardy Weinberg: why?

- Assumptions:
  - Diploid
  - Sexual reproduction
  - Random mating
  - Bi-allelic sites
  - Large population size, ...

- Why? Each individual randomly picks his two chromosomes. Therefore, \( \text{Prob.} (Aa) = pq + qp = 2pq \), and so on.
Hardy Weinberg: Generalizations

- Multiple alleles with frequencies \( q_1, q_2, \ldots, q_H \)
  - By HW,
    
    \[
    \text{Pr[homozygous genotype } i ] = q_i^2 \\
    \text{Pr[heterozygous genotype } i, j ] = 2q_i q_j \]

- Multiple loci?
Hardy Weinberg: Implications

- The allele frequency does not change from generation to generation. Why?
- 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 disease?
- Males are 100 times more likely to have the "red" type of color blindness than females. Why?
- Conclusion: While the HW assumptions are rarely satisfied, the principle is still important as a baseline assumption, and significant deviations are interesting.
What causes variation in a population?

- Mutations (may lead to SNPs)
- Recombinations
- Other genetic events (gene conversion)
Single Nucleotide Polymorphisms

Infinite Sites Assumption:
Each site mutates at most once

00000101011
10001101001
01000101010
01000000011
00111100000
001011100110
Short Tandem Repeats

GCTAGATCATCATCATCATCATGCTAG
GCTAGATCATCATCATCATGCTAGTAT
GCTAGATCATCATCATCATCATCATATTGC
GCTAGATCATCATCATCATGCTAGTAT
GCTAGATCATCATCATCATGCTAGTAT
GCTAGATCATCATCATCATCATATTGC
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 finger-print.

<table>
<thead>
<tr>
<th>loci</th>
<th>individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 3</td>
<td>4 2</td>
</tr>
<tr>
<td>5 1</td>
<td>3 3</td>
</tr>
<tr>
<td>3 2</td>
<td>3 2</td>
</tr>
<tr>
<td>3 1</td>
<td>3 1</td>
</tr>
<tr>
<td>5 3</td>
<td>5 3</td>
</tr>
</tbody>
</table>
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 heterogeneous.
- EX: $i$ is heterogeneous w.r.t $\{A,D,E\}$

<table>
<thead>
<tr>
<th>Column</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
</tr>
</tbody>
</table>

$i_0$ and $i_1$
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?
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
  - $\Pr[A,B=0,1] = 0.25$
    - Linkage disequilibrium
- Case 2: Extensive recombination
  - $\Pr[A,B=(0,1)=0.125$
    - Linkage equilibrium

```
A  B
0  1
0  1
0  0
0  0
1  0
1  0
1  0
1  0
```
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 $P_{00} = P_{0*} P_{*0}$
- $D = \text{abs}(P_{00} - P_{0*} P_{*0}) = \text{abs}(P_{01} - P_{0*} P_{*1}) = \ldots$
LD over time

- With random mating, and fixed recombination rate $r$ between the sites, Linkage Disequilibrium will disappear
  - Let $D^{(t)} = LD at time t$
  - $P^{(t)}_{00} = (1-r) P^{(t-1)}_{00} + r P^{(t-1)}_{0*} P^{(t-1)}_{*0}$
  - $D^{(t)} = P^{(t)}_{00} - P^{(t)}_{0*} P^{(t)}_{*0} = P^{(t)}_{00} - P^{(t-1)}_{0*} P^{(t-1)}_{*0} \ (HW)$
  - $D^{(t)} = (1-r) D^{(t-1)} = (1-r)^t D^{(0)}$
LD over distance

- Assumption
  - Recombination rate increases linearly with distance
  - LD decays exponentially with distance.
- 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.
LD and disease gene mapping problems

- Marker density?
- Complex diseases
- Population sub-structure