Recombination

• Recap
  – Coalescent
  – Motivation
• Linkage Disequilibrium
  – Measures
  – Factors affecting LD
  – Uses
• Coalescent with Recombination
• Estimating Recombination
RECAP: Sample variance and evolutionary stochasticity

Parameters  Population  Sample

\[ N_e \]
\[ u \]

Parameter estimation

Assuming that the population conforms to the Wright-Fisher model, what can I deduce about the underlying evolutionary parameters from the sample?

Model Choice

Is the Wright-Fisher model an accurate description of reality? Do other models (e.g. with selection, population-structure) give a better fit to the data?
RECAP: Population History

The genealogical history of a sample
RECAP: The Coalescent

Genealogy

Genealogy with mutations

Sample

Copyright: Gilean McVean, 2001
Data from a Single Locus

Genealogy with mutations

Sample 1

Sample 2

Estimate $\theta$ per kb from sequence of length L

$\text{Var}(\hat{\theta}) = A + B/L$
Effect of Recombination

Different parts of sequences may have different “parents” and hence different genealogies:

Larger sequences: more recombinations and more genealogies.
Usefulness of Recombination

- The presence of recombination means that we can “average” over the randomness in the genealogy.
- Estimating the mutation rate per kb from a recombining sequence of length $L$ gives an estimator with variance proportional to
  \[
  \log(L)/L
  \]
- Similarly, variance of $D$-statistics are reduced in the presence of recombination.
- Optimal sequencing strategy is to sequence long sequences in a few genes.
Problems of Recombination

• Current theory:
  – Single locus
  – Unlinked loci

• New theory:
  – Linked Loci
RECAP: Hardy-Weinberg equilibrium

Assumptions:

• Random-mating
• Neutrality

\[
N_t \frac{N(\bullet)}{2N} = x
\]

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>● ●</td>
<td>$x^2$</td>
</tr>
<tr>
<td>● ○</td>
<td>$2x(1-x)$</td>
</tr>
<tr>
<td>○ ●</td>
<td></td>
</tr>
<tr>
<td>○ ○</td>
<td>$(1-x)^2$</td>
</tr>
</tbody>
</table>

Copyright: Gilean McVean, 2001
Linkage Equilibrium

What about alleles at different loci?

Loci are in Linkage Equilibrium if

\[ f_{AB} = f_A \times f_B \]
\[ f_{Ab} = f_A \times f_b \]
\[ f_{aB} = f_a \times f_B \]
\[ f_{ab} = f_a \times f_b \]
Linkage Disequilibrium (LD)

- Motivates the following definition of LD:
  \[ D = f_{AB} - f_A \times f_B \]
- Absolute value does not depend on choice of alleles:
  \[ D = f_{ab} - f_a \times f_b \]
  \[ -D = f_{Ab} - f_A \times f_b \]
  \[ -D = f_{aB} - f_a \times f_B \]
- Maximum (absolute) value depends on \( f_A \) and \( f_B \). Alternative definitions:

\[ D' : D \text{ scaled to be between 0 and 1} \]

\[ \Delta^2 = \frac{D^2}{f_A \cdot f_a \cdot f_B \cdot f_b} \]
Decay of Linkage Disequilibrium

- Under HWE assumptions, if $r$ is the recombination probability per meiosis then the expected value of LD satisfies:

$$D_t = (1 - r) \cdot D_{t-1}$$

$$D_t = (1 - r)^t \cdot D_0$$
Creation of Linkage Disequilibrium

- Genetic drift:
  
  \[ \text{e.g. } (25,25,25,25) \rightarrow (28,27,19,26) \]

\[ \Delta^2 = 0.0 \rightarrow \Delta^2 = 0.007 \]
Expected Linkage Disequilibrium

- Trade-off between creation and decay:

\[ \Delta^2 \approx \frac{1}{(4N_Er)} \]
WARNING: Population Structure

Non-random mating can create linkage disequilibrium between loci, regardless of the rate of recombination between them:

- Common Ancestor: AB haplotype
- Mutation: A to a
- Mutation B to b
- Assuming Fixation
- aB
- Ab

$\Delta^2 = 1$. 
Testing for absence of recombination

- In the absence of recombination, $\Delta^2$ will be uncorrelated with physical distance.
- Otherwise, negative correlation

Example (mtDNA):

![Graph showing correlation]

Correlation is $-0.83$. 
Calculating critical region

- Permutation Test:

  **True Data:**
  - ![Graph](image)

  **Permutation:**
  - ![Graph](image)

  **Permuted Data:**
  - ![Graph](image)
  - Equally likely under the Null Hypothesis

Calculate correlation for all permuted data sets and compare actual correlation with these correlations

For mtDNA $p=0.01$; though a more recent (larger) study suggests there is no evidence for recombination in mtDNA.
Coalescent with Recombination

Represented by:
Coalescent with Recombination

Reconstructed via:
Coalescent with Recombination

- Simple to simulate recombinant genealogy; mutations; and sample.
- Simulate back in time; rates the same as in the coalescent, except recombination events occur at rate $\rho/2$ per branch. Where:

$$\rho = 4N_E r$$
Likelihood Theory

- Likelihood is the probability of the data calculated for different values of the parameters (in our case $\rho, \theta$):

- “Contains all information in the data”
- Estimate parameter at maximum.
- Approximate confidence intervals can be obtained.
Estimating the Likelihood

• Imagine we observe the recombinant genealogy. Calculating the likelihood is then simple.

• We only observe the sample; so “guess” the ancestral genealogy. Calculate the likelihood for this “guess”

• Different “guesses” give different answers. So guess many different genealogies and average the results.
Example “Guesses”
Theoretical Properties

- This approach can be formulated mathematically.
- Regardless of the number of “guesses”, you get an unbiased approximation.
- As the number of “guesses” tends to infinity the approximation becomes exact.
- The accuracy depends on the number of “guesses” and the rule for “guessing”.
Example: Beta-globin

- 3kb of sequence data surrounding the beta-globin gene. The 5’ region is believed to be part of a hot-spot.

- Likelihood surface for 5’ region:

- Estimate $\rho=33$ per kb ($r = 5 \times 10^{-4}$ per kb)

- For 3’ region; estimate of $r$ is less than genome-wide average.
Warning!

• Need to check accuracy of likelihood approximation. Can take a considerable amount of computing time.

• If not enough “guesses” are made:

  poor estimate of likelihood: poor estimate of parameter.
Summary: Estimating Recombination

- Population data enables fine-scale estimation.
- Freely available methods exist.
- Full-likelihood methods are optimal, but can take large amounts of computing time. The accuracy of the approximation to the likelihood needs to be checked before inference is made.
- Alternative methods exist with alternative approximations to the likelihood (by “throwing away data”; or analytical approximations).