DNA Analysis Techniques
for Molecular Genealogy

Luke Hutchison (lukeh@email.byu.edu)
Project Supervisor: Scott R. Woodward
Mission: The BYU Center for Molecular Genealogy

- To establish the world’s most comprehensive genetic and genealogical database.
- To create tools for reconstruction of genealogies from DNA.
- To establish genetic links between families throughout the world.
Molecular Genealogy: Process

• 100,000 DNA samples and genealogies are being collected from 500 different populations

• Common ancestors and population structure are inferred [population and quantitative genetics]

• A searchable database is being produced for DNA-based genealogical research
Common Ancestor

(Suffolk, England, ca. 1893: 33%; Glasgow, Scotland, ca. 1905: 12%; ...)

Unknown genealogy
What is the Basis of Molecular Genealogy?

• Each individual carries within their DNA a record of who they are and how they are related to all other people.

• You received all of your DNA from your two parents (50% from each).

• Specific regions of DNA have properties that can:
  • Identify an individual
  • Link them to a family
  • Identify extended family groups (tribes or clans)
3 major types of genetic data

- **Y Chromosome**
  - Males only, paternal inheritance
  - Haploid, none or little recombination
  - 0.51% of an individual's total genetic information

- **Mitochondrial DNA**
  - Both males and females, maternal inheritance
  - Haploid, none or little recombination
  - 0.0006% of an individual's total genetic information

- **Autosomal (Nuclear)**
  - Both males and females, inherited equally from both parents
  - Diploid, undergoes recombination at each generation
  - >99% of your genetic information
### Genotypic and Genealogical data

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8549</td>
</tr>
<tr>
<td>8550</td>
</tr>
<tr>
<td>8551</td>
</tr>
<tr>
<td>8552</td>
</tr>
<tr>
<td>8553</td>
</tr>
<tr>
<td>8554</td>
</tr>
</tbody>
</table>
(a) Sequence polymorphism

```
--------AGACTAGACATT--------
--------AGATTAGGCAATT--------
```

(b) Length polymorphism

```
--------(AATG)(AATG)(AATG)--------
  3 repeats

--------(AATG)(AATG)--------
  2 repeats
```
Types of DNA Data Extracted

• Pair of alleles (numbers of repeats) for a locus (e.g., 121,123)
  • **Linked loci** (close together in chromosome)
  • **Unlinked loci** (distant enough from each other to be genetically unrelated, due to the high probability of a crossover occurring between the markers; the presence of one does not imply the presence of the other)
Linked Loci: “Haplotypes”

• The probability of a crossover event occurring in the middle of a haplotype is low, since the loci are tightly linked.

• Haplotypes are therefore likely to be passed down intact for many generations.
Haplotyping

- **Problem:** Correct order of the genetic information in a pair is unknown (which allele came from which parental chromosome?):
  - $121,123$ or $123,121$?

- The problem compounds for linked loci:
  - $121|123$ $121|123$ $123|121$ }
  - $142|144$ $144|142$ $142|144$ }... ($x 2^3=8$)
  - $115|119$ $115|119$ $115|119$ }

- Finding which alleles occur together on the same chromosome for linked loci (the *haplotypes*) is called **hapotyping**. The alignment is called the **phase**.
Properties of Haplotypes

• Populations which do not inter-breed each develop a distinctive distribution of haplotypes.

• Haplotypes may eventually appear (due to mutation and/or crossover) that do not exist in any other population.

• Haplotypes give much more discerning power than alleles alone, since there are many possible haplotypes given a set of possible alleles at each locus.
Haplotyping: A Cyclic Problem

- We could figure out the most likely phase for the alleles in a haplotype if we knew the haplotype distributions of the parent populations.

- We could figure out the haplotype distributions of the parent populations if we knew the correct phase of the alleles.
Haplotyping: A Cyclic Solution

• (1) First guess for phase probs: all equal (0.125)

• (3), (5), ... Estimate phase probabilities based on the current estimate of population haplotype probabilities

• (2), (4), ... Estimate population haplotype probabilities based on the current estimate of phase probabilities
Haplotyping: Results

• Convergence typically achieved in 3-7 iterations

• Difficult to judge accuracy since nobody knows how to get the true correct answer!

• Previous researchers’ attempts on simulated data: 50-60% accuracy

• Our algorithm on (different) simulated data: 97%

• Our algorithm on real data (accuracy measured by 'spiking' with CEPH data): 88-93%
How to Contact Us

E-mail: molecular-genealogy@email.byu.edu

Phone: 801-378-1245

Mail: 788 WIDB
      BYU
      Provo, UT
      84602

Web: http://molecular-genealogy.byu.edu