Genetic Homogeneity in Option 12 of Mendel

Chiara Sabatti, UCLA
csabatti@mednet.ucla.edu

A Short Course on Statistical Genetics with Mendel
UCLA, September 17-20, 2001
Genetic Homogeneity

- We are looking at two populations and looking at similarity/differences between them.

- Example of populations: affected and controls; early Finnish settlement/late Finnish settlement; etc.

- We look at the differences in
  - allele frequency
  - haplotype frequency
  - multilocus genotype frequency

- We use two non-parametric exact tests based on permutations.
Association mapping

Suppose that in a population of chromosomes, 1 undergoes a mutation in a gene that causes a disease

\[ D_4 \quad B_3 \quad A_1 \quad D \quad L_3 \quad M_2 \quad R_1 \quad T_2 \]

- Initially all the chromosomes that inherit the disease inherit this haplotype;
- recombination and mutation will erode the haplotype;
- the distribution of alleles at the markers close to the disease will be different in the disease population and in the remaining individuals.
Data Type

→ A random sample of genotypes or haplotypes from two populations ($n_D$ from the first and $n_C$ from the second).

Ex. Haplotypes on three biallelic markers

<table>
<thead>
<tr>
<th>Population</th>
<th>sample</th>
<th>Marker 1</th>
<th>Marker 2</th>
<th>Marker 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>Individual 1</td>
<td>$a$</td>
<td>$B$</td>
<td>$C$</td>
</tr>
<tr>
<td>$D$</td>
<td>Individual 2</td>
<td>$A$</td>
<td>$b$</td>
<td>$c$</td>
</tr>
<tr>
<td>$D$</td>
<td>Individual $n_D$</td>
<td>$a$</td>
<td>$B$</td>
<td>$c$</td>
</tr>
<tr>
<td>$C$</td>
<td>Individual 1</td>
<td>$A$</td>
<td>$B$</td>
<td>$C$</td>
</tr>
<tr>
<td>$C$</td>
<td>Individual $n_C$</td>
<td>$a$</td>
<td>$B$</td>
<td>$C$</td>
</tr>
</tbody>
</table>
The data table

<table>
<thead>
<tr>
<th>Population</th>
<th>abc</th>
<th>abC</th>
<th>aBc</th>
<th>Abc</th>
<th>⋯</th>
<th>ABC</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>$n_{D1}$</td>
<td>$n_{D2}$</td>
<td>$n_{D3}$</td>
<td>$n_{D4}$</td>
<td>⋯</td>
<td>$n_{D8}$</td>
</tr>
<tr>
<td>C</td>
<td>$n_{C1}$</td>
<td>$n_{C2}$</td>
<td>$n_{C3}$</td>
<td>$n_{C4}$</td>
<td>⋯</td>
<td>$n_{C8}$</td>
</tr>
<tr>
<td></td>
<td>$n.1$</td>
<td>$n.2$</td>
<td>$n.3$</td>
<td>$n.4$</td>
<td>⋯</td>
<td>$n.8$</td>
</tr>
</tbody>
</table>

⇒ If we have genotypes at one marker the table would have less columns; if we had multilocus genotypes without phase information, it would have more columns.

⇒ The more markers considered, the higher number of alleles, the higher number of columns

⇒ Often the table has lots of empty entries.
### The data table probability

<table>
<thead>
<tr>
<th>Population</th>
<th>$abc$</th>
<th>$abC$</th>
<th>$aBc$</th>
<th>$Abc$</th>
<th>...</th>
<th>$ABC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$n_{D1}$</td>
<td>$n_{D2}$</td>
<td>$n_{D3}$</td>
<td>$n_{D4}$</td>
<td>...</td>
<td>$n_{D8}$</td>
</tr>
<tr>
<td>$C$</td>
<td>$n_{C1}$</td>
<td>$n_{C2}$</td>
<td>$n_{C3}$</td>
<td>$n_{C4}$</td>
<td>...</td>
<td>$n_{C8}$</td>
</tr>
<tr>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n_3$</td>
<td>$n_4$</td>
<td>...</td>
<td>$n_8$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

- Fix the haplotype (genotypes) counts
- Assume that the two populations are equal
- Each set of counts $n_{ij}$ that fills in the table with fixed marginal have a probability.
## Permutations

Permute population labels to obtain observations from the null

<table>
<thead>
<tr>
<th>Population</th>
<th>sample</th>
<th>Marker 1</th>
<th>Marker 2</th>
<th>Marker 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>Individual 1</td>
<td>( a )</td>
<td>( B )</td>
<td>( C )</td>
</tr>
<tr>
<td>( C )</td>
<td>Individual 2</td>
<td>( A )</td>
<td>( b )</td>
<td>( c )</td>
</tr>
<tr>
<td>( D )</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>( D )</td>
<td>Individual ( n_D )</td>
<td>( a )</td>
<td>( B )</td>
<td>( c )</td>
</tr>
<tr>
<td>( C )</td>
<td>Individual 1</td>
<td>( A )</td>
<td>( B )</td>
<td>( C )</td>
</tr>
<tr>
<td>( D )</td>
<td>Individual 2</td>
<td>( A )</td>
<td>( B )</td>
<td>( c )</td>
</tr>
<tr>
<td>( C )</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>( C )</td>
<td>Individual ( n_C )</td>
<td>( a )</td>
<td>( B )</td>
<td>( C )</td>
</tr>
</tbody>
</table>
Fisher exact test of independence

**P-value**: sum of the probabilities of all the tables that have a probability smaller than the one of the observed one.

**P-value via permutations**

\[
P\text{-value} = \frac{\#\text{Permutations : } Pr(\text{permut}) \leq Pr(\text{obser})}{\#\text{Permutations}}
\]

\[\rightarrow\] it is not based on asymptotic approximations (as a \(\chi^2\) test would be) \[\rightarrow\] it is good for sparse tables.

\[\rightarrow\] we can estimate the p-value with a random sample of permutations.
Z-max test

- In presence of founder effect, often there is only one haplotype or allele that has a markedly different distribution between the two compared populations.

- To concentrate on this most divergent case, Mendel uses the Z-max test:

  1. For each haplotype (allele) $i$ calculate the standardized difference from the expected frequency under homogeneity:

     $$Z_i = \frac{|n_{Di}/n_D - n_{i}/n|}{\sqrt{n_{i}/n(1 - n_{i}/n)/n_D}}$$

  2. Find the haplotype/allele, that has the biggest difference:

     $$Z_{\max} = \max_i Z_i$$
3. Evaluate the distribution of the test-statistic $Z_{\text{max}}$ using the permutations.

4. P-value of the test: frequency of permutations that lead to a $Z_{\text{max}}$ value larger or equal than the observed one.
Association mapping

• If one has population-type data (a random sample of genotype/haplotypes), one can use option 11 for this goal.

• A strategic decision regards how many markers to analyze at the same time:
  – if one marker has high mutation rate, its alleles frequencies may be similar in disease and control population even if the marker is close to the disease → looking at haplotype is more robust.
  – if we look at too large an haplotype, we will loose the signal and encounter the problem of too sparse table

• Mendel will soon offer the option of looking at a sliding window of markers of variable length.
Outline of input files for Option 11

**Locus** Standard. Frequency information is not used.

**Map** Standard. Frequency information is not used. Use it to specify on which markers to include in the analysis:
- one marker $\rightarrow$ homogeneity of allele frequencies
- two or more markers $\rightarrow$ homogeneity of haplotype (multilocus genotype) frequencies.

**Pedigree**
- They have to be one-person pedigrees.
- You can specify the number of copies of the pedigree.
- One haplotype is entered as a everywhere homozygous mulitlocus genotype.
- enter all the pedigree from one population first, followed by the pedigree of the second population.
Control There are two required keywords. One specifies the option

   OPTION=12

The other the number of pedigrees in the second population:

   PEDIGREES_INSECOND_POPULATION=28

May need to force the program to read the number of copies

   READ_PEDIGREE_COPIES=TRUE

To control the number of sampled permutations:

   SAMPLE=30000
Option 12 Output files

- The relevant output are the p-value of the two test.
- There is also a standard deviation of the p-value.
- For the $Z_{\text{max}}$ test, the most divergent haplotype/allele is also indicated.
<table>
<thead>
<tr>
<th>POPULATION</th>
<th>PEOPLE</th>
<th>PEOPLE COUNTING</th>
<th>REPETITIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>194</td>
<td></td>
</tr>
</tbody>
</table>

The Fisher exact test for genetic homogeneity has approximate p-value $0.5700E-02$ plus or minus $0.1506E-02$ based on 10000 resamples.

The Z_max exact test for genetic homogeneity has approximate p-value $0.8000E-02$ plus or minus $0.1782E-02$ based on 10000 resamples.

The haplotypes or multi-locus genotypes of Pedigree 6 diverge most from what is expected.

Time of operation was 22.570000 seconds.
Other related Mendel options

- A parametric test for homogeneity of allele frequencies can be done with Option 6.

- If you are doing association mapping with family data, use Option 13.

- Option 8 is an alternative, more general model, for association analysis.

- If you are interested in Linkage equilibrium/Disequilibrium between markers, use Option 11.
More Coming Soon

- A sliding window of variable length that identifies which markers are in the haplotype tested for homogeneity.
- An evaluation of P-value that takes into account the problem of multiple comparison.