

Genetic Homogeneity in Option 12 of Mendel

Chiara Sabatti, UCLA
`csabatti@mednet.ucla.edu`

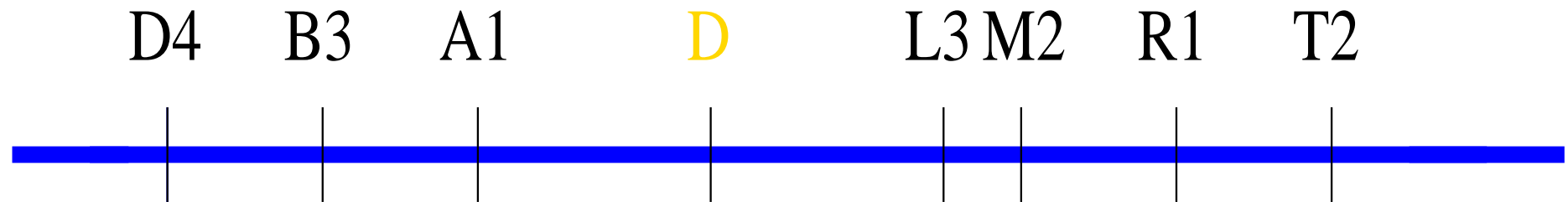
A Short Course on Statistical Genetics with Mendel
UCLA, September 17-20, 2001

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



- 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

Population	sample	Marker 1	Marker 2	Marker 3
<i>D</i>	Individual 1	<i>a</i>	<i>B</i>	<i>C</i>
<i>D</i>	Individual 2	<i>A</i>	<i>b</i>	<i>c</i>
⋮	⋮	⋮	⋮	⋮
<i>D</i>	Individual n_D	<i>a</i>	<i>B</i>	<i>c</i>
<i>C</i>	Individual 1	<i>A</i>	<i>B</i>	<i>C</i>
⋮	⋮	⋮	⋮	⋮
<i>C</i>	Individual n_C	<i>a</i>	<i>B</i>	<i>C</i>

The data table

Population	<i>abc</i>	<i>abC</i>	<i>aBc</i>	<i>Abc</i>	...	<i>ABC</i>	
<i>D</i>	n_{D1}	n_{D2}	n_{D3}	n_{D4}	...	n_{D8}	n_D
<i>C</i>	n_{C1}	n_{C2}	n_{C3}	n_{C4}	...	n_{C8}	n_C
	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$...	$n_{.8}$	$n_{.}$

⇒ 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

Population	abc	abC	aBc	Abc	\dots	ABC	
D	n_{D1}	n_{D2}	n_{D3}	n_{D4}	\dots	n_{D8}	n_D
C	n_{C1}	n_{C2}	n_{C3}	n_{C4}	\dots	n_{C8}	n_C
	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	\dots	$n_{.8}$	n

- 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

Population	sample	Marker 1	Marker 2	Marker 3
<i>D</i>	Individual 1	<i>a</i>	<i>B</i>	<i>C</i>
<i>C</i>	Individual 2	<i>A</i>	<i>b</i>	<i>c</i>
<i>D</i>	⋮	⋮	⋮	⋮
<i>D</i>	Individual n_D	<i>a</i>	<i>B</i>	<i>c</i>
<i>C</i>	Individual 1	<i>A</i>	<i>B</i>	<i>C</i>
<i>D</i>	Individual 2	<i>A</i>	<i>B</i>	<i>c</i>
	⋮	⋮	⋮	⋮
<i>C</i>	Individual n_C	<i>a</i>	<i>B</i>	<i>C</i>

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

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

⇒ it is not based on asymptotic approximations (as a χ^2 test would be) → it is good for sparse tables.

⇒ 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_{\max} using the permutations.
4. P-value of the test: frequency of permutations that lead to a Z_{\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 lose 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 → homogeneity of allele frequencies
- two or more markers → 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 multilocus 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_IN_SECOND_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_{\max} test, the most divergent haplotype/allele is also indicated

GENETIC HOMOGENEITY OPTION

POPULATION	PEOPLE	PEOPLE COUNTING REPETITIONS
1	11	52
2	28	194

THE FISHER EXACT TEST FOR GENETIC HOMOGENEITY HAS APPROXIMATE PVALUE
0.5700E-02 PLUS OR MINUS 0.1506E-02 BASED ON 10000 RESAMPLES.

THE Z_MAX EXACT TEST FOR GENETIC HOMOGENEITY HAS APPROXIMATE PVALUE
0.8000E-02 PLUS OR MINUS 0.1782E-02 BASED ON 10000 RESAMPLES.

THE HAPLOTYPES OR MULTILOCUS 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.