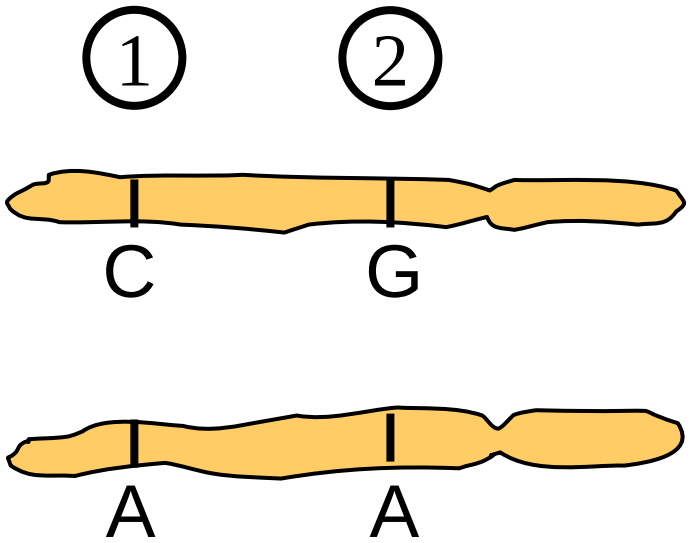
Estimating Haplotype Frequencies From Data With Incomplete Phase Information

Jurg Ott
Rockefeller University, New York

ott@rockefeller.edu

Two Marker Loci (SNPs)

- Locus 1:Alleles C and A, genotype C/A
- Locus 2: Alleles G and A, genotype G/A
- **Haplotype** = set of alleles at different loci (inherited in a gamete from one parent)



One chromosome (cytogenetics)

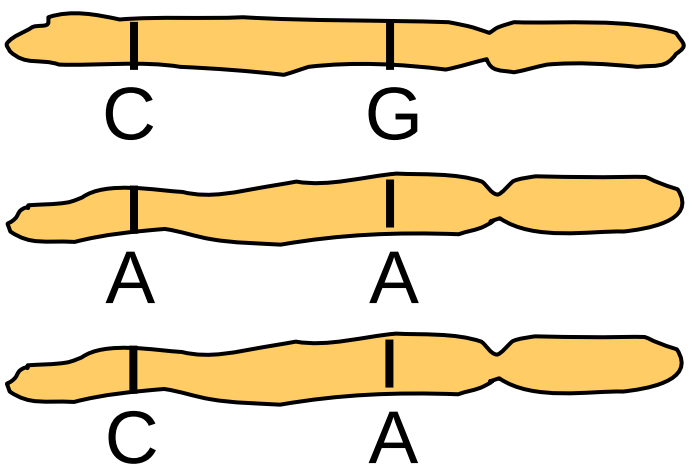
One haplotype (genetics)

Other possible haplotypes:
C-A, A-G

Note: Loci need not be on same chromosome!

Genotypes and Haplotypes

Locus 1	Locus 2		
	G/G	G/A	A/A
C/C	C-G, C-G	C-G, C-A	C-A, C-A
C/A	C-G, A-G	?	C-A, A-A
A/A	A-G, A-G	A-G, A-A	A-A, A-A



$$? = \begin{cases} C-G, A-A \text{ or} \\ C-A, A-G \end{cases}$$

Counting Haplotypes

Locus 1	Locus 2		
	G/G	G/A	A/A
C/C	0	1	2
C/A	0	1	2
A/A	1	0	1

Known haplotypes			New counts
	No.	Freq	
C-G	1	0.071	1.221
C-A	7	0.500	7.779
A-G	2	0.143	2.779
A-A	4	0.286	4.221
Total	14	1	16

*) Assumes HWE

or

Ambiguous	Frequency	Rel. freq.	New counts
1 C-G, 1 A-A	0.071 × 0.286	0.221	0.221 C-G, 0.221 A-A
1 C-A, 1 A-G	0.500 × 0.143	0.779	0.779 C-A, 0.779 A-G
Sum	0.092	1	1 1

EM Algorithm

- The iterative procedure shown on the previous slide is known to lead to maximum likelihood estimates.
- Originally called gene counting (Ceppellini, Siniscalco & Smith, 1955. The estimation of gene frequencies in a random mating population. *Ann Hum Genet* **20**, 97-115), later EM algorithm (Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Statist Soc* **39B**, 1-38).
- *Note*: In practice, start with equal phase probabilities – the two possible pairs of haplotypes for doubly heterozygous individuals are given equal weight.

Implementation

- *snphap* computer program
- Dr. David Clayton, Cambridge UK
- Estimation of haplotype frequencies by MLE using different starting values. For individuals with multiple phases, genotypes with probability < 0.01 disregarded.
- Assign (infer) haplotypes to individuals using MCMC approach (Gibbs sampling). Assumes a prior distribution (Dirichlet) of haplotype frequencies.
- *Phase* program: Perhaps better. Modify default parameter values!

Example: LEPR Gene

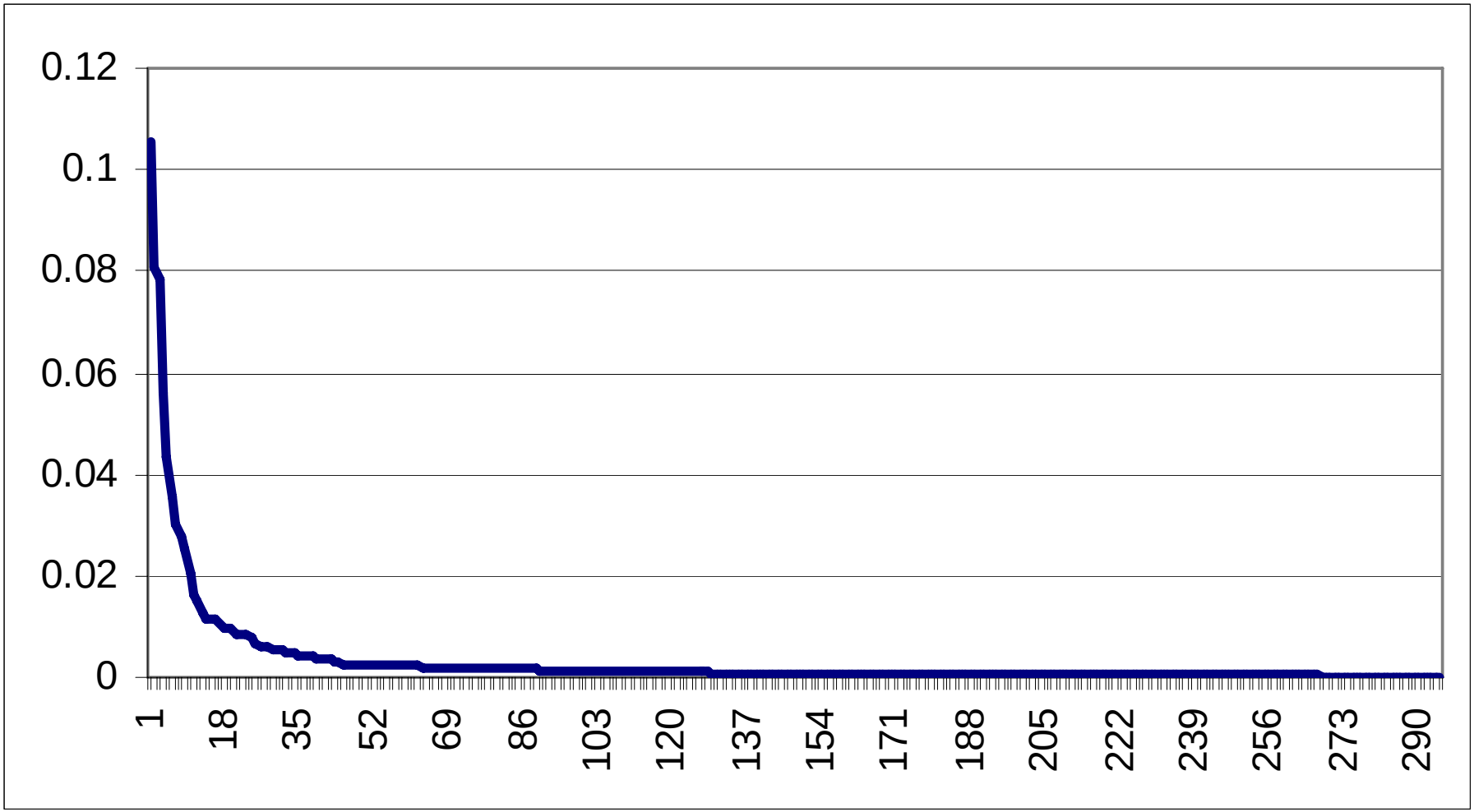
- In 564 individuals, gene fully sequenced
- Found 83 SNPs
- Potential number of haplotypes = 2^{83} $= 9.7 \times 10^{24}$.
- Most common haplotypes with estimated frequencies:

11111111111111111121111121112112211111111111111111111111111111111111111111	0.106
11111111111111111111111111111111111111111111111111111111111111111111111111	0.081
11111111111111111111111111112121111111111111111211121111121221111112111112112	0.078
11111121212121111111211121112111211111111111111112121121112111111121111121111	0.056

Estimation Results

- Total of 851 haplotypes estimated to be present.
- Of these, 295 with $f \geq 0.000{,}001$
  and 556 with $f < 0.000{,}001$
- Smallest "real" frequency:
  $n = 564 \rightarrow 2n = 1128$ haps $\rightarrow 1/1128 = 0.000{,}887$

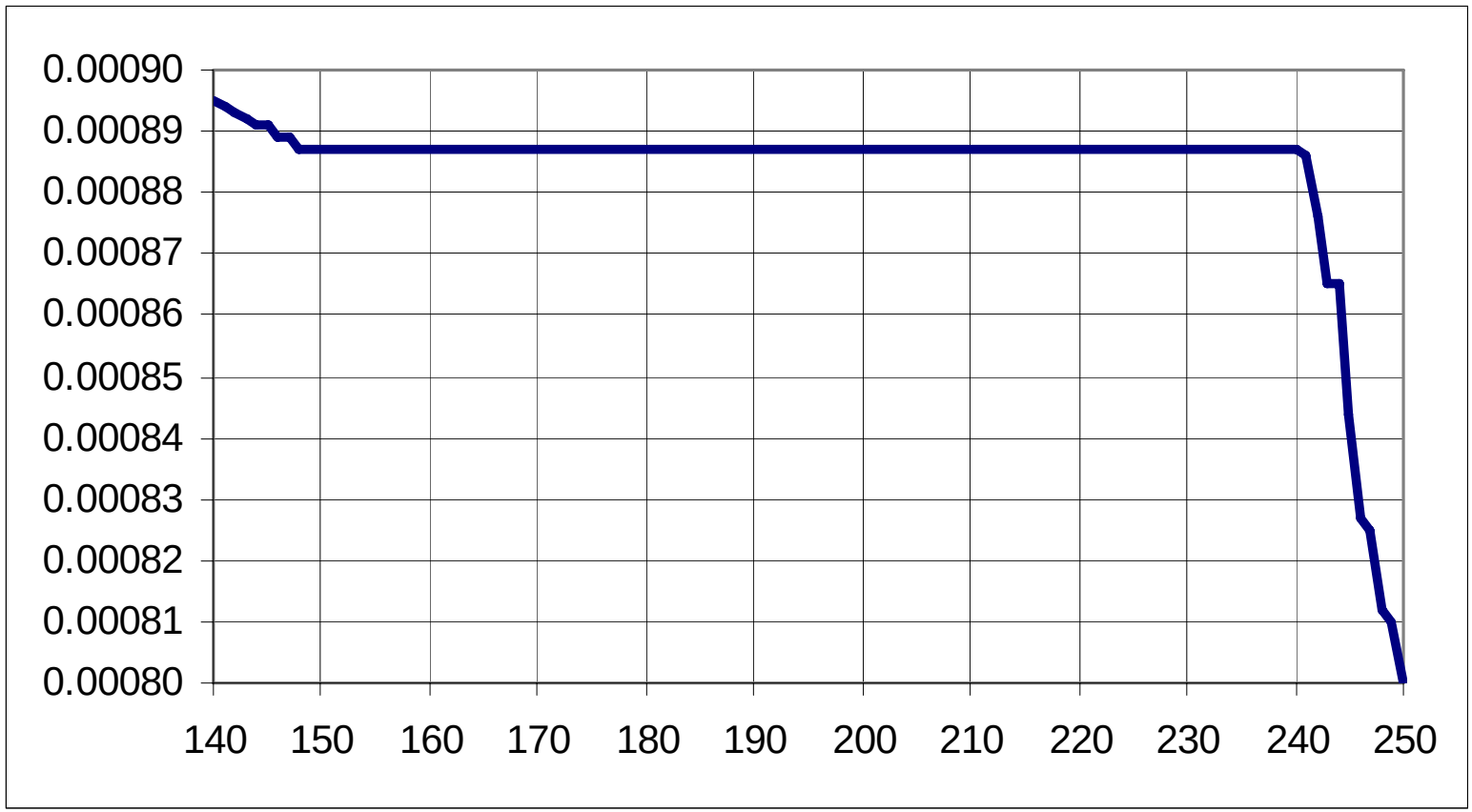
Hap Frequencies > 0.000,001



Enlargement

Horizontal line: Many values of 0.000,887

"Real" number of haplotypes ≈ 240?



Potential Solution

- Work with assigned/inferred haplotypes
- Not the same as multiplying haplotype frequencies by total number of haps!
- Of the 1128 inferred haps, only 16 have assignment probabilities < 0.50.
- Total of 265 different haplotypes inferred, compared with 240 haplotypes with frequencies $\geq$ 0.000,887
- **Problem**: Different assignment schemes are based on different priors $\rightarrow$ diff. results.

Is it important to know the number of haplotypes?

- LR test for haplotype frequency differences between case and control individuals
- Number of df = (# haps in cases – 1) + (# haps in controls – 1) – (# of haps in cases and controls – 1)

	Hap 1	Hap 2	...	Hap 20
cases	...	...	...	...
controls	...	...	...	...

Here,
#df = 19

Dataset from Beijing

Assigned haplotypes, partial table

Number	Hap	cases	prop.	controls	prop.	OR	1/OR	chisquare
1	GCCIGCA	**253**	0.4765	**608**	0.4780	0.99	1.01	0.004
2	ATADATA	**120**	0.2260	**263**	0.2068	1.12	0.89	0.828
5	ACCIGCA	**12**	0.0226	**60**	0.0472	0.47	**2.14**	**5.899**
13	ACADATA	**0**	0	**5**	0.0039	0	inf	2.093
14	GCADATT	**0**	0	**5**	0.0039	0	inf	2.093
15	GCCIGTA	**0**	0	**5**	0.0039	0	inf	2.093
26	ACCDGCA	**0**	0	**1**	0.0008	0	inf	0.418
27	ACCIATT	**0**	0	**1**	0.0008	0	inf	0.418
36	GCCDGCT	**2**	0.0038	**0**	0	inf	0	**4.796**
37	GCCIGTT	**1**	0.0019	**0**	0	inf	0	2.397
38	GTCDATT	**1**	0.0019	**0**	0	inf	0	2.397
39	GTCIATT	**1**	0.0019	**0**	0	inf	0	2.397
		531	1	1272	1			25.832

Summary of Pooling Haplotypes

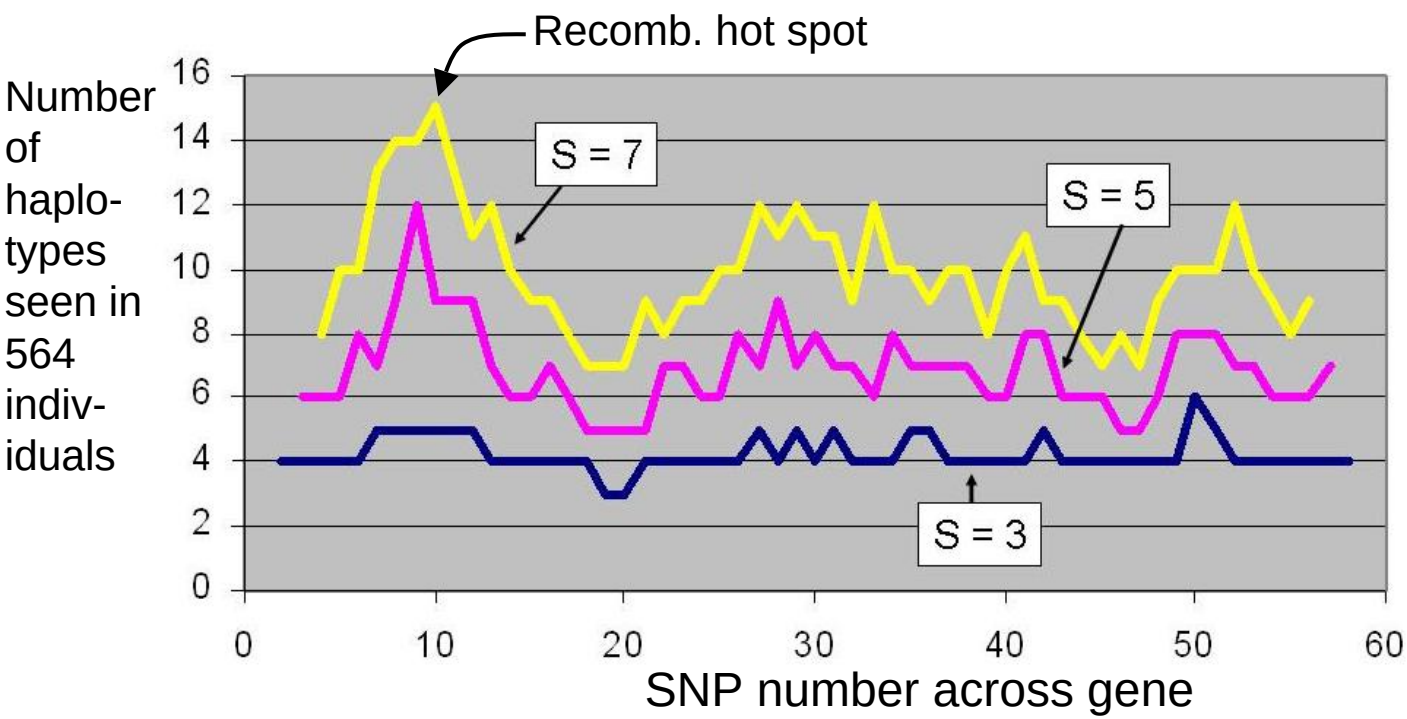
Data	Pearson chi-sq	df	table p	Fisher p	Chi-square	table p
No pooling of cells	53.75	38	0.0466	0.0217	64.24	0.0049
Cells with 0 in one group and 1 in the other group are merged	53.75	28	0.0024	0.0016	64.24	0.0001
Cells with 0 in one group are merged	53.75	21	0.0001	<0.0001	64.24	<0.0001
Cells with freq<0.01 in each group are pooled to form "rare" category	10.37	9	0.3216	0.2946	11.18	
Cells with freq<0.05 in each group are pooled to form "rare" category	2.17	4	0.7046	0.7003	2.18	
The last row reflects results obtained by the PHASE program						

Solutions

- Work with inferred haplotypes (treat them as if observed): Not reliable unless different programs give similar results.
- Find null distribution of LR statistic via permutation sampling — no need for # df! Implemented in *PHASE* program. *Note*: #repetitians (MCMC) and #permutations should be ≥ 5000 each.

LD across genome

- 4-gamete test: Pairs of adjacent SNPs. The more haplotypes, the smaller LD
- *GOLD* plot (heat map)
- LEPR gene, sliding window of *s* SNPs:

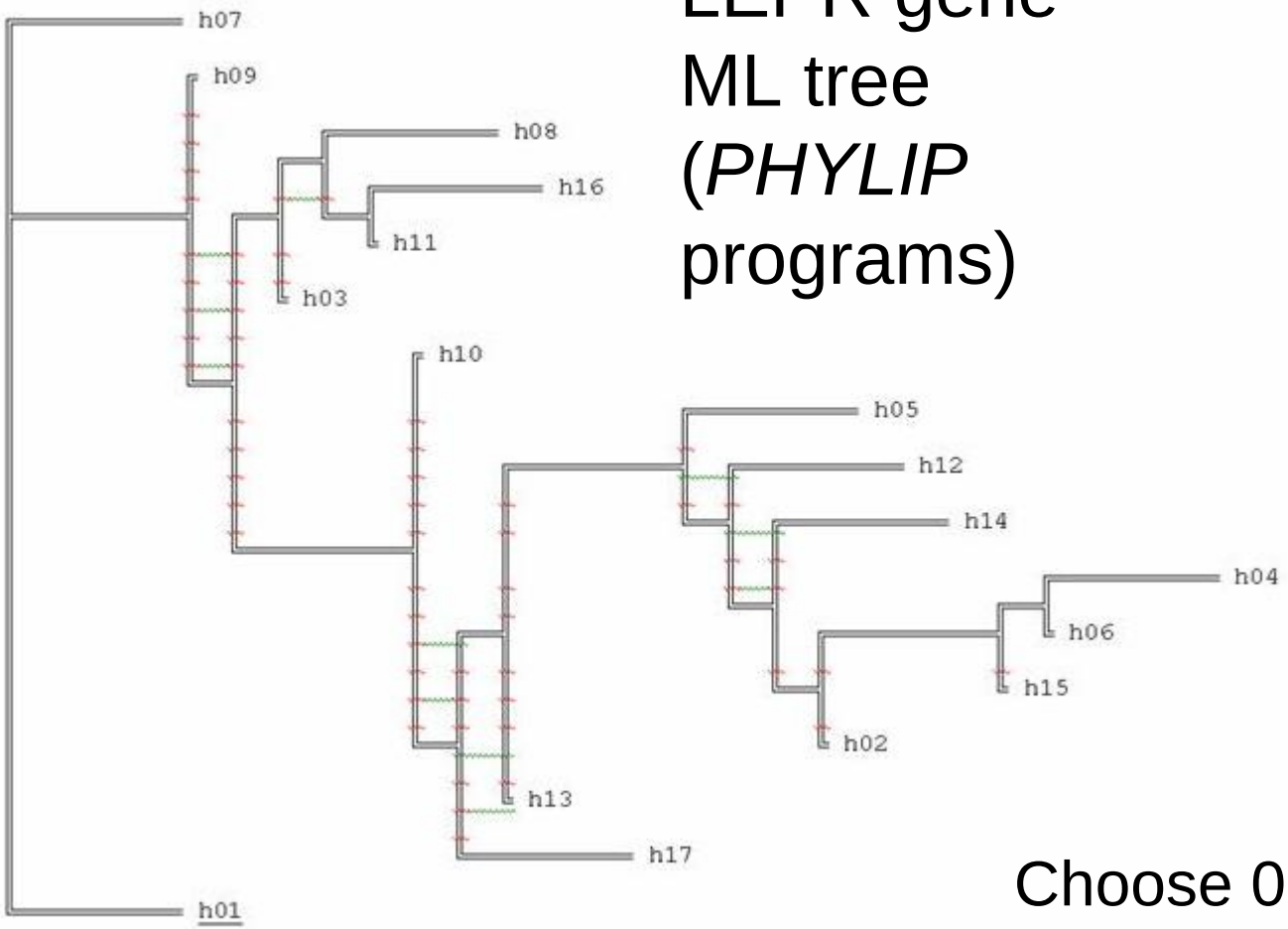


s SNPs:

Max. # haps = 2^s

s	2^s
7	128
5	32
3	8

"Evolutionary" Tree



LEPR gene ML tree (*PHYLIP* programs)

h01	0000000000000
h02	0100000110110
h03	0011110110100
h04	0001011011000
h05	1000000000000
h06	0010000110000
h07	0000000000001
h08	0100000000000
h09	0011100110100
h10	0010000110110
h11	0001110110100
h12	0001000011000
h13	0000000110110
h14	0001010011000
h15	0000000110000
h16	0011111110100
h17	0010000000000

Choose 0 = common allele (ancestral)