Topics in Statistical Genetics

Karl W Broman

Department of Biostatistics
Johns Hopkins School of Public Health

Past

- QTLs in experimental crosses
- CEPH data { 8 families (~ 130 people) >8000 STRPs
  - Genetic maps
  - Recombinational variation
  - Autozygosity
  - Crossover interference
- Genotyping/pedigree errors
- Dog genetic maps
- Various disease projects

Traditional genetics: **binary** traits

e.g. disease or no disease

Quantitative traits

e.g. yield of tomato crop

bristles on a fly's abdomen

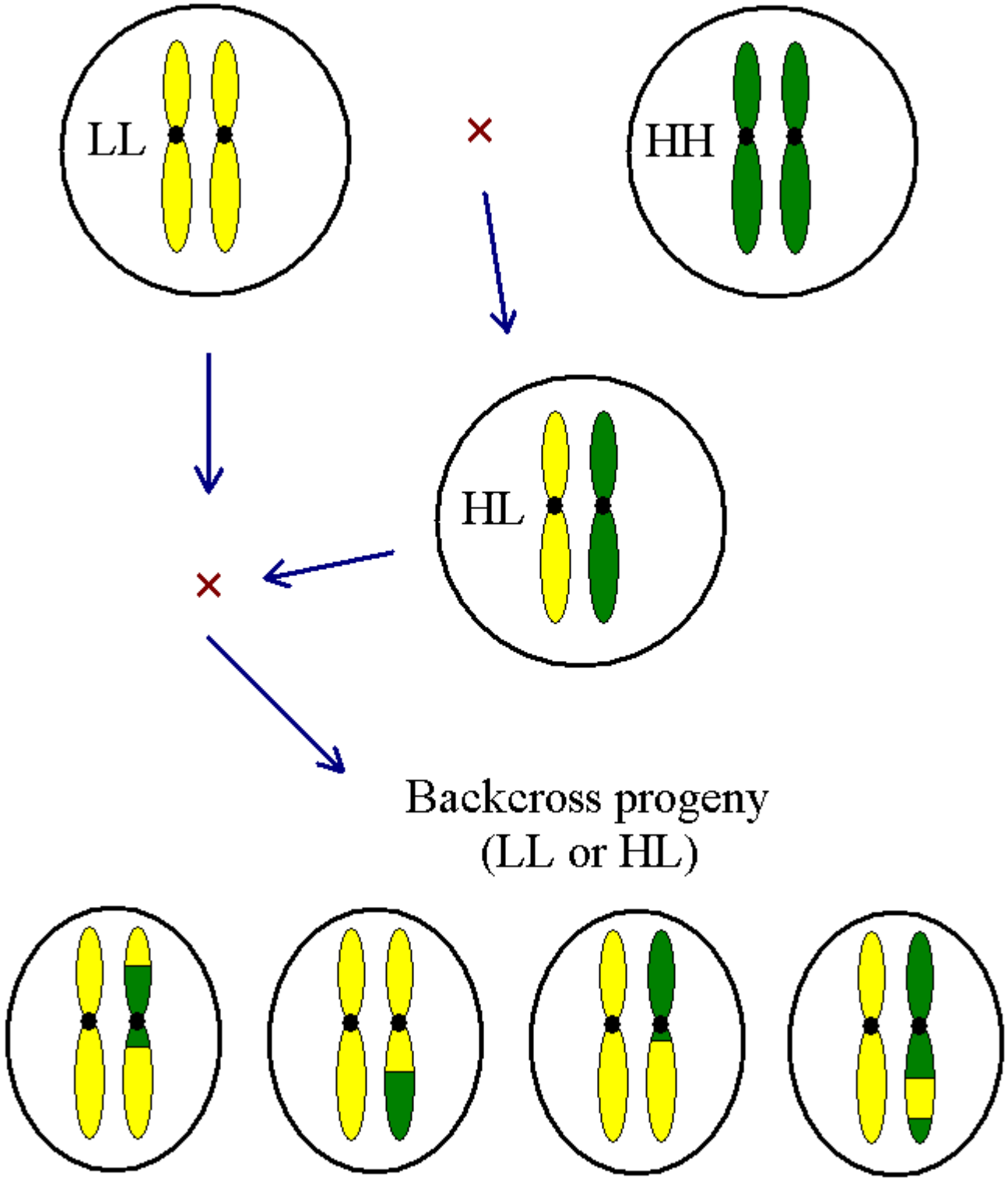
– Many genes
– Environmental variation

Why?

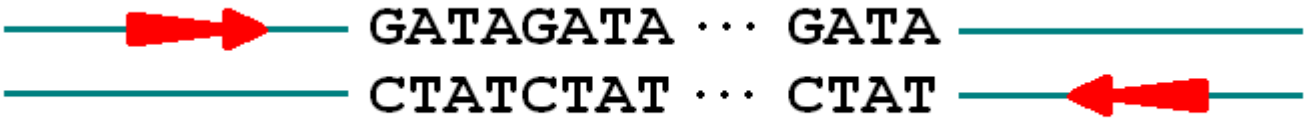
– Biochemical basis of trait
– Selection experiments
– Evolution

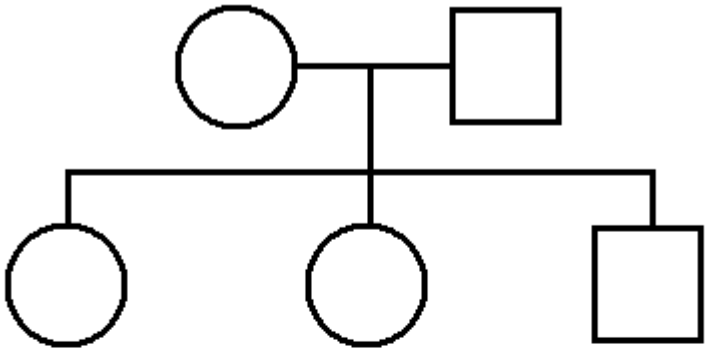
Goal: find (some of) the genes

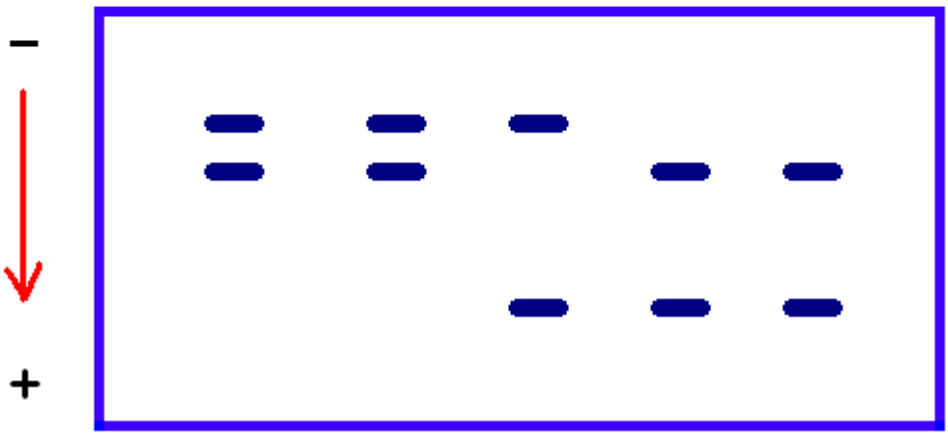
Backcross experiment



Genetic markers:
STRPs or microsatellites







Data

Phenotypes (trait values)

y_i = phenotype for individual i

Marker genotypes

x_{ij} = 1/0 if i is HL/LL at marker j

Genetic map

Locations of markers

Models

Recombination: No interference

Phenotype/genotype connection

$$y = \mu + \sum \beta_j z_j + \varepsilon$$

$$\varepsilon \sim \text{Normal}(0, \sigma^2)$$

Problem

100 to 1000 backcross progeny

100 to 400 markers

$y = \mu + \sum \beta_j x_j + \varepsilon$

Find the x's with $\beta_j \neq 0$

Errors:

- Miss important loci
- Include extraneous loci

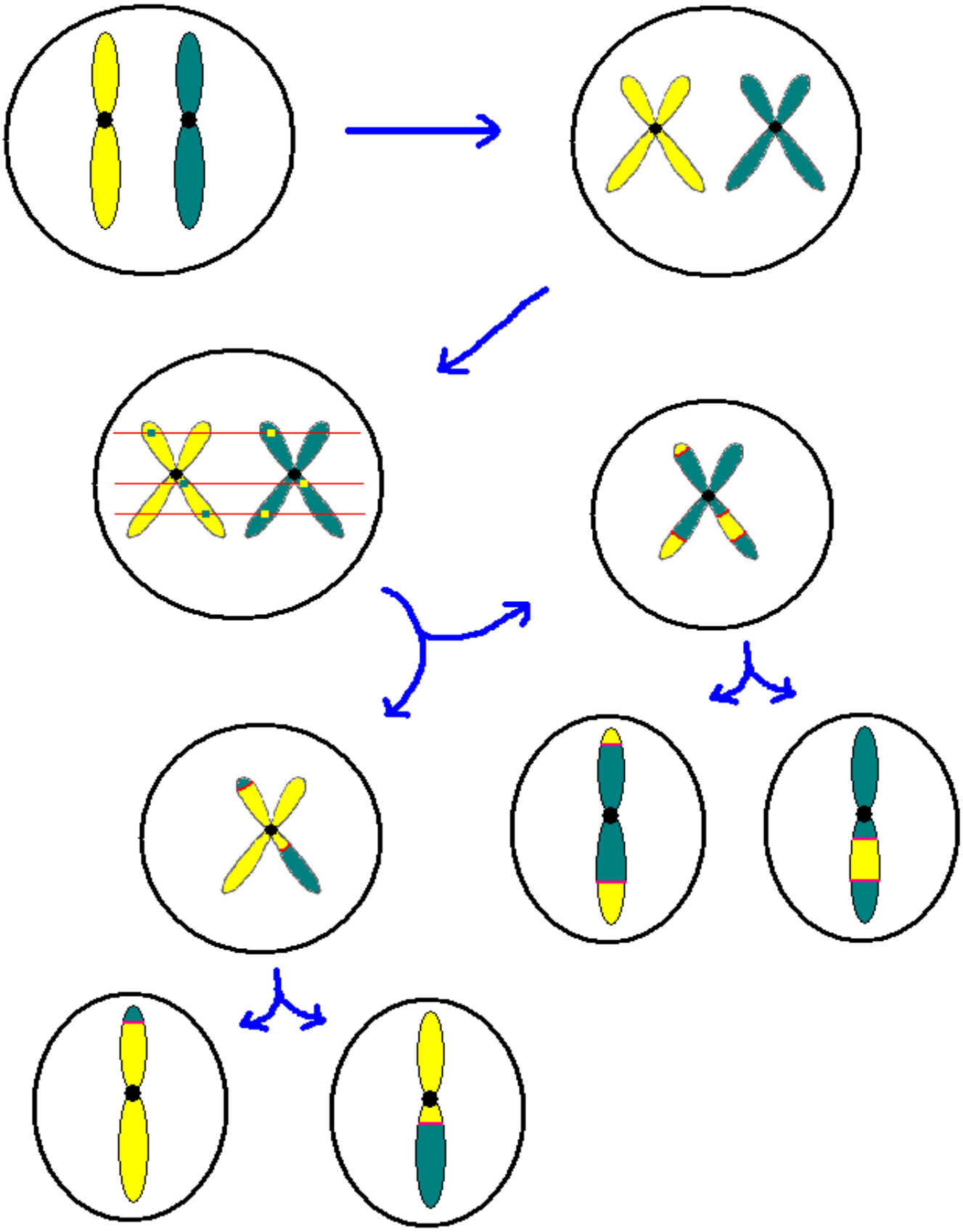
Discussion

- Identifying QTLs is model selection
- Simulation studies are necessary
  - compare procedures
  - understand a procedure's performance
- Different situations will require different procedures

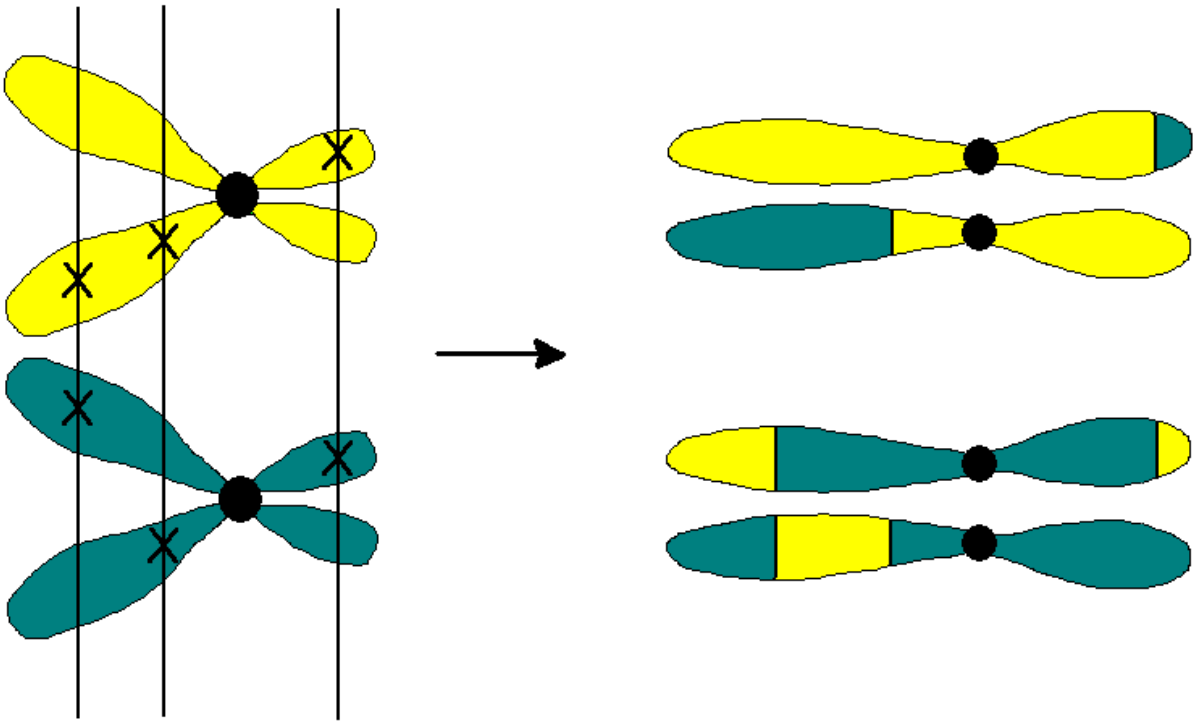
Human data

- www.marshmed.org/genetics
- 8 CEPH families
  - three generations
  - 11 to 15 progeny
  - 92 meioses
- ~8,000 STRP markers
  - 90 ± 7 % typed
- Average spacing
  - female: 0.6 ± 1.2 cM
  - male: 0.4 ± 1.0 cM
  - sex-ave: 0.5 ± 0.9 cM
- Data cleaning
  - Removed 764/964,425 (~0.08%) genotypes resulting in tight double recombinants

Meiosis



Interference



- Strand choice

  → Chromatid interference

- Spacing

  → Chiasma (crossover) interference

Genetic distance

distance (cM) = average # crossovers in 100 meiotic products

per Morgan { 2 chiasmata on 4-strand bundle
1 crossover on meiotic product

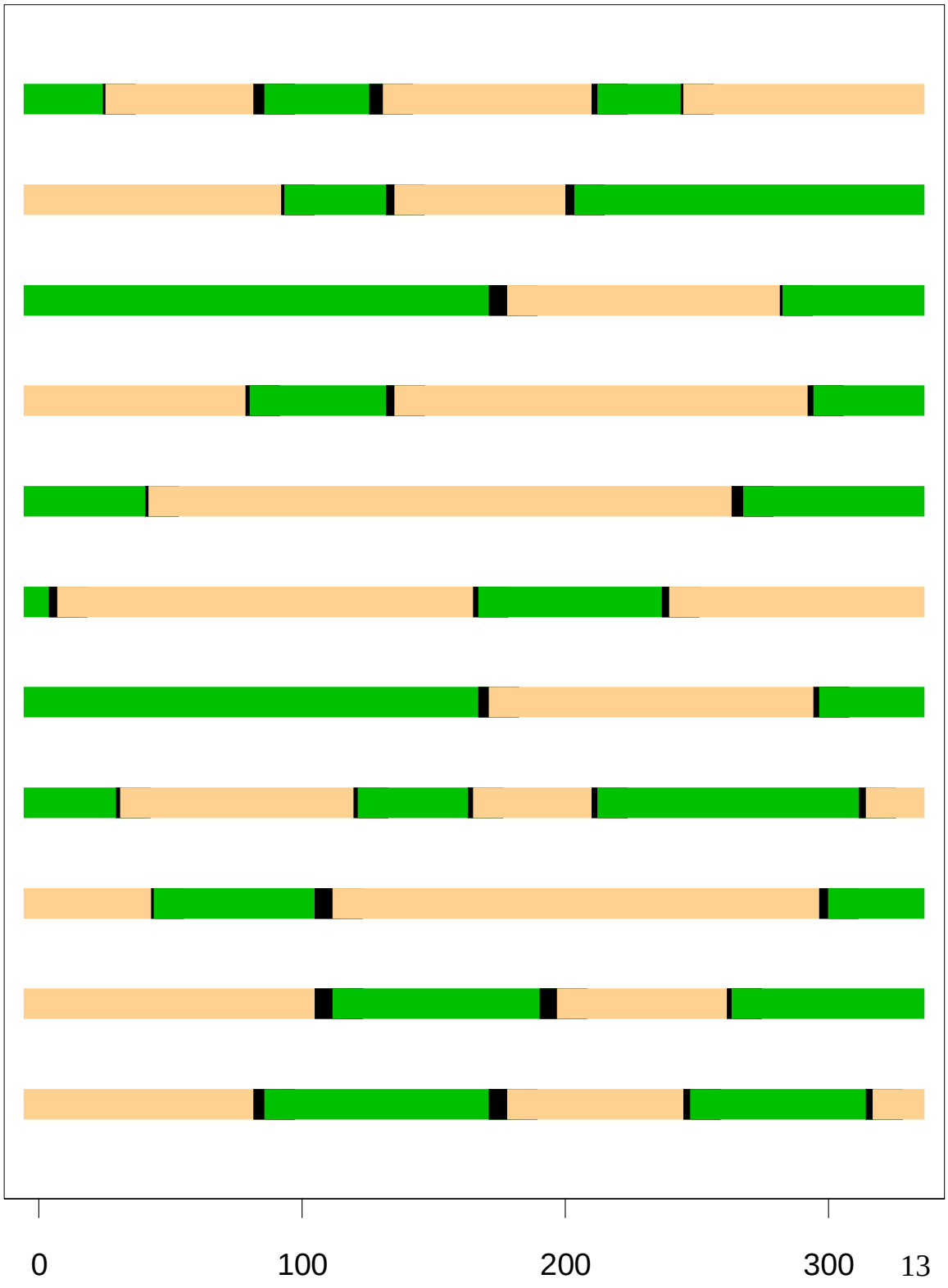
Map function

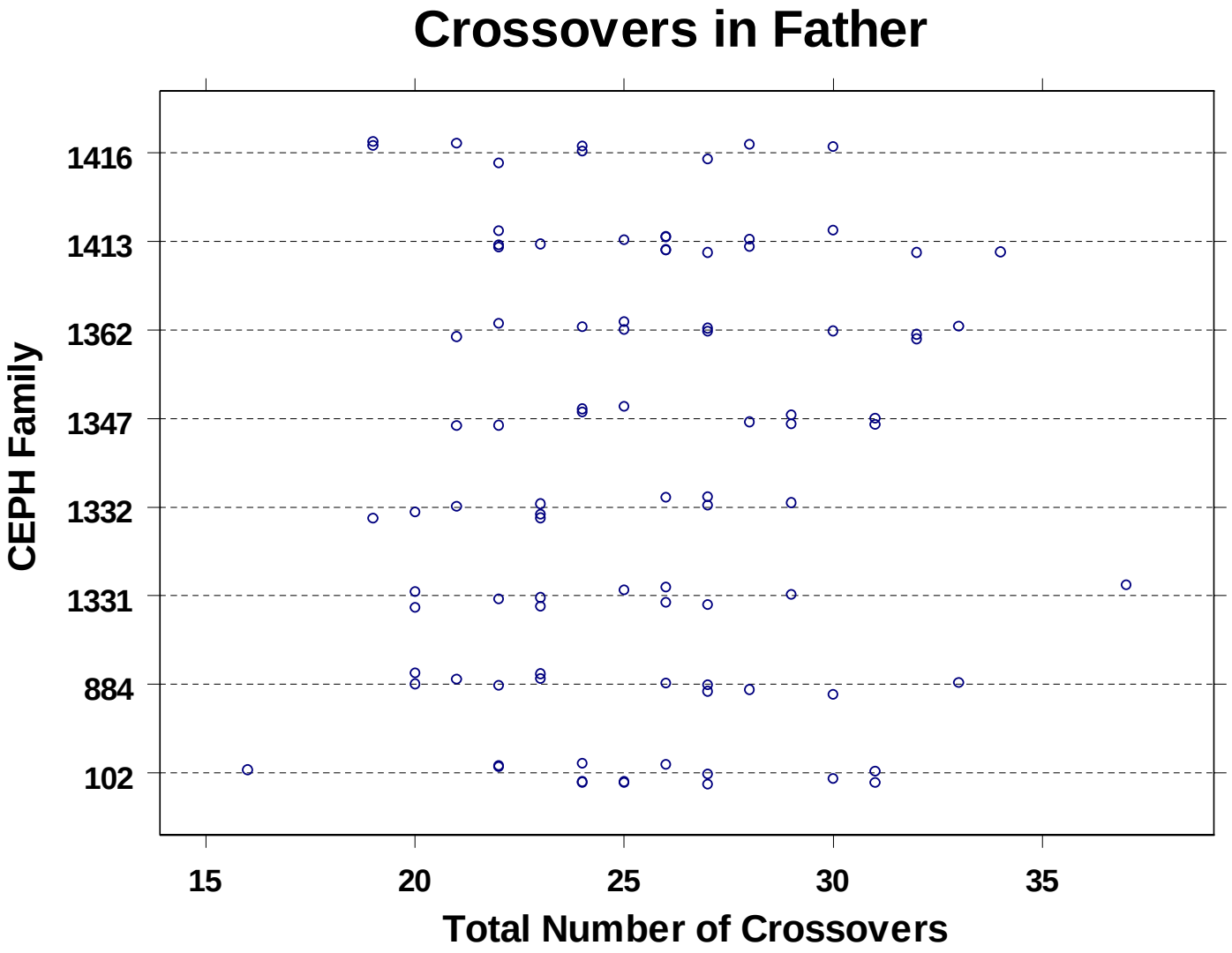
recombination fraction as a function of genetic distance

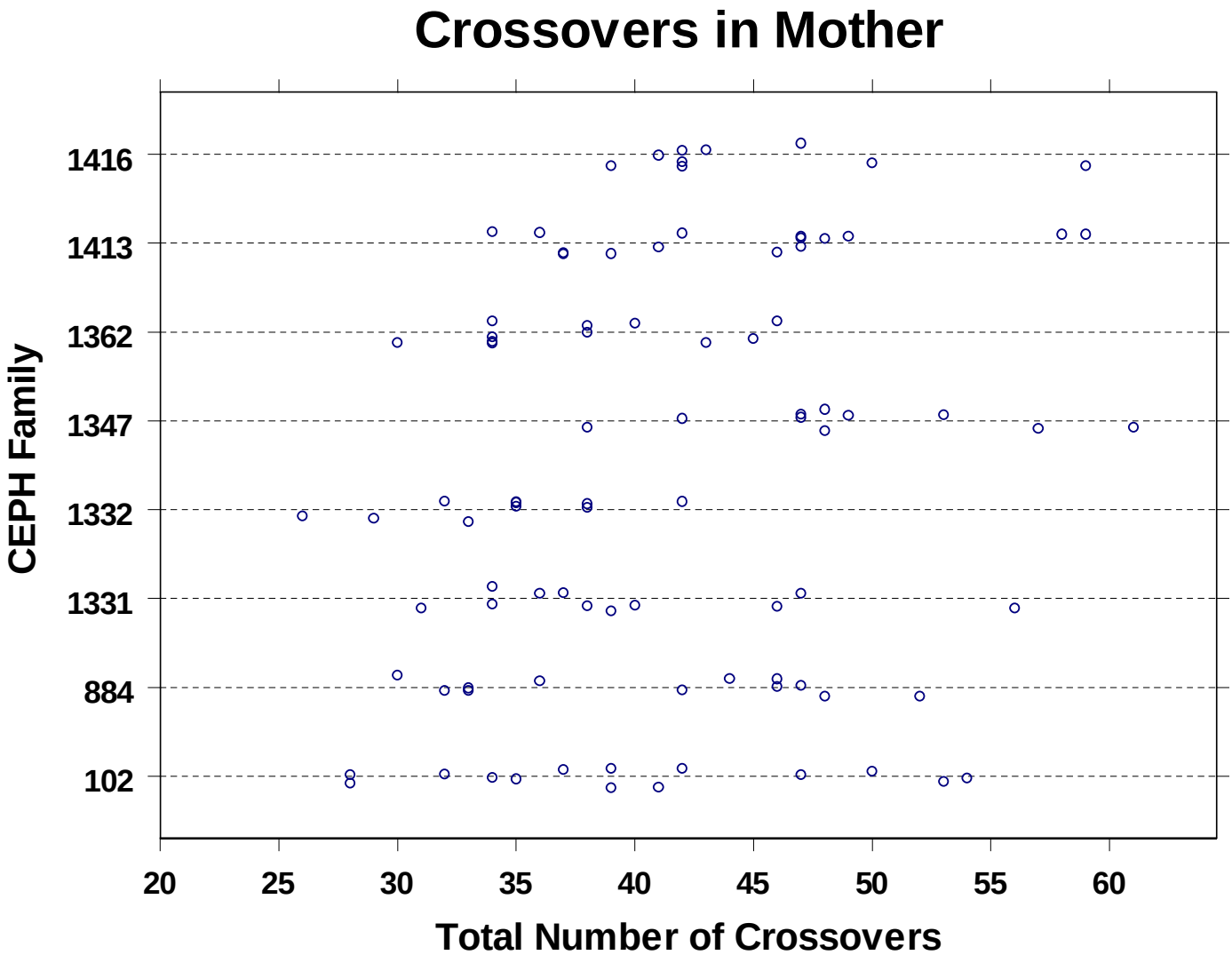
Haldane $r(d) = \frac{1}{2} [1 - \exp(-2d)]$

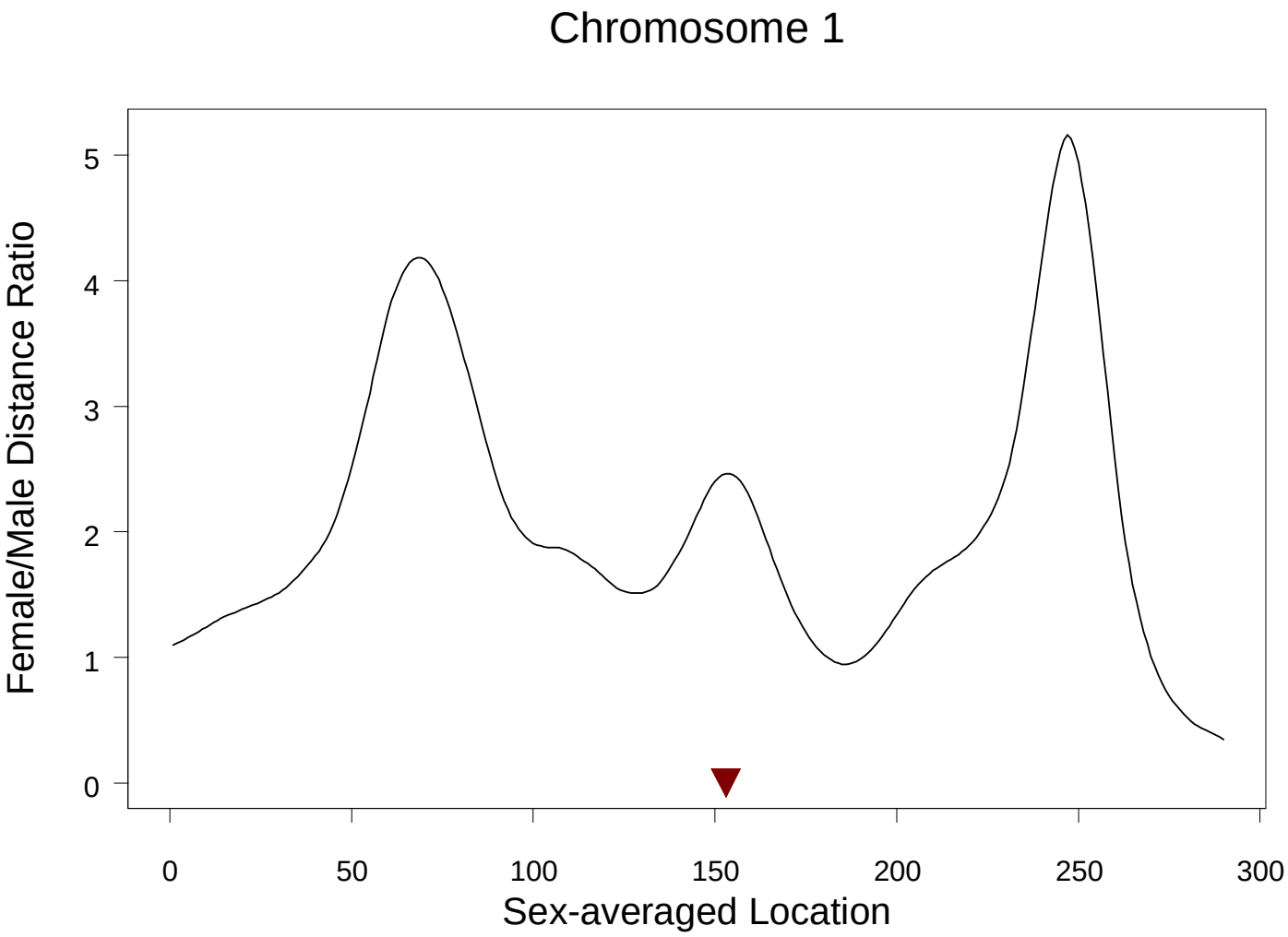
Kosambi $r(d) = \frac{1}{2} \tanh(2d)$

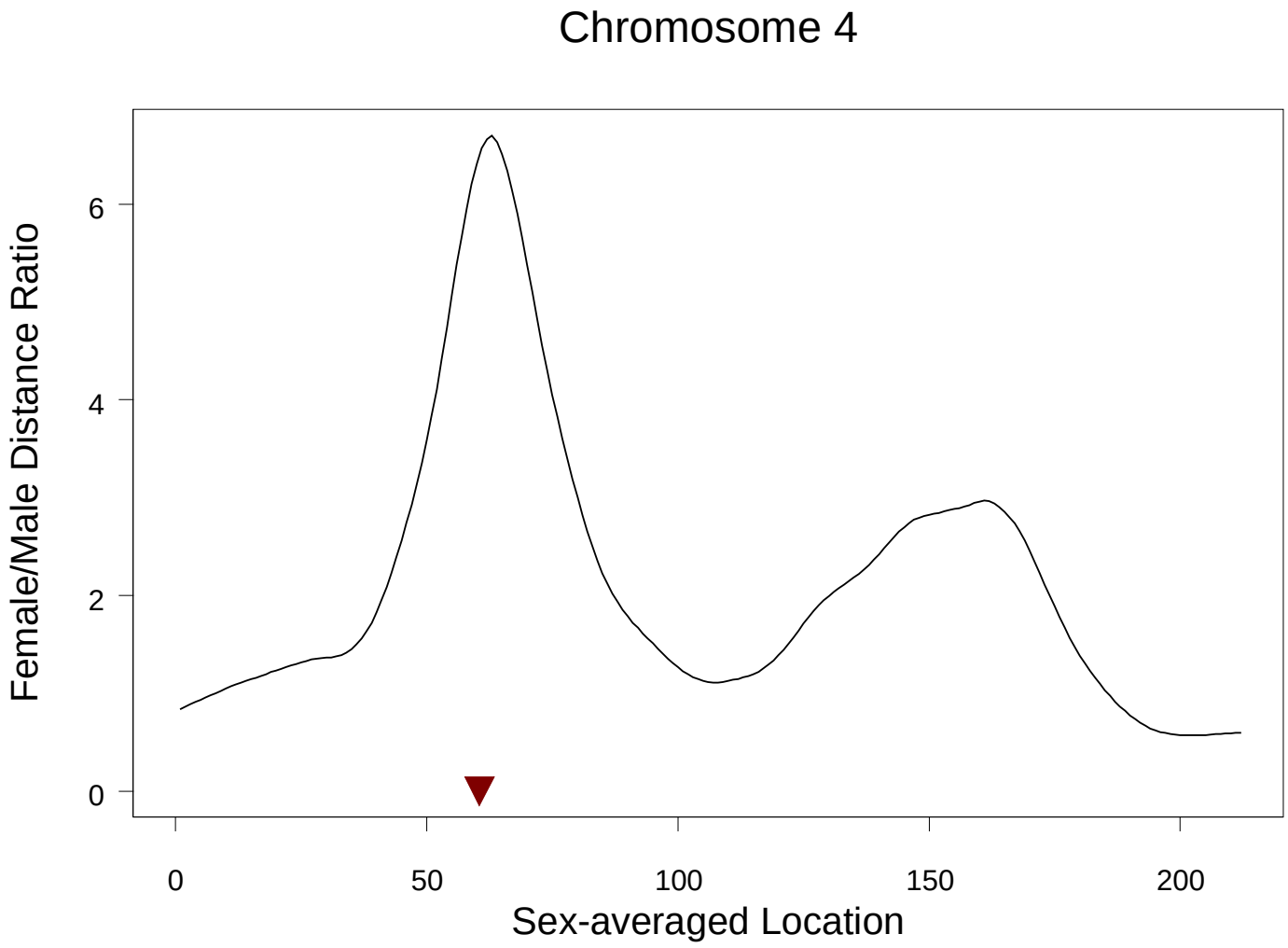
The data

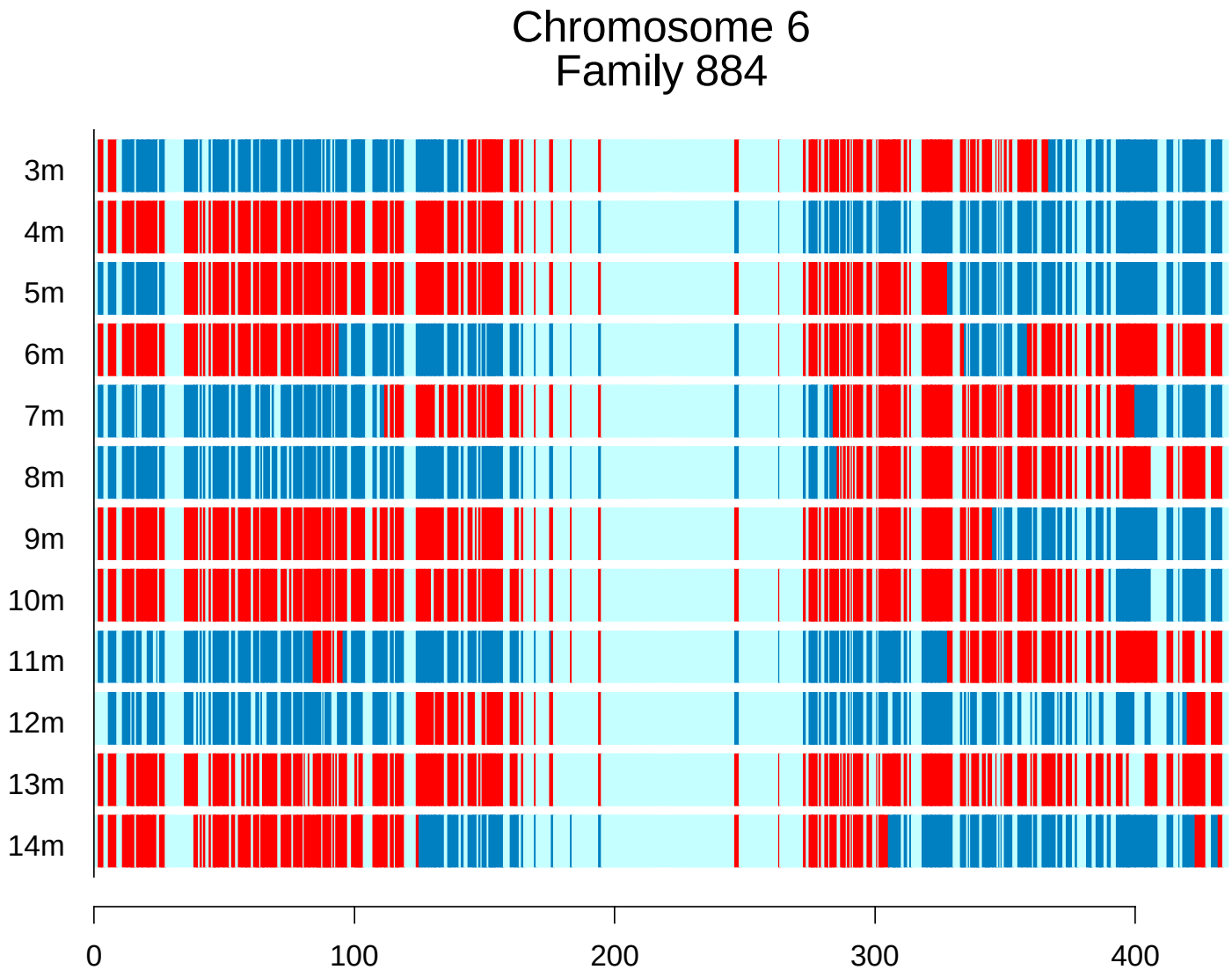



Crossovers in Father
CEPH Family
1416
1413
1362
1347
1332
1331
884
102
15
20
25
30
35
Total Number of Crossovers


Crossovers in Mother
CEPH Family
1416
1413
1362
1347
1332
1331
884
102
20
25
30
35
40
45
50
55
60
Total Number of Crossovers


Chromosome 1
Female/Male Distance Ratio
0
1
2
3
4
5
0
50
100
150
200
250
300
Sex-averaged Location


Chromosome 4
Female/Male Distance Ratio
0
2
4
6
0
50
100
150
200
Sex-averaged Location

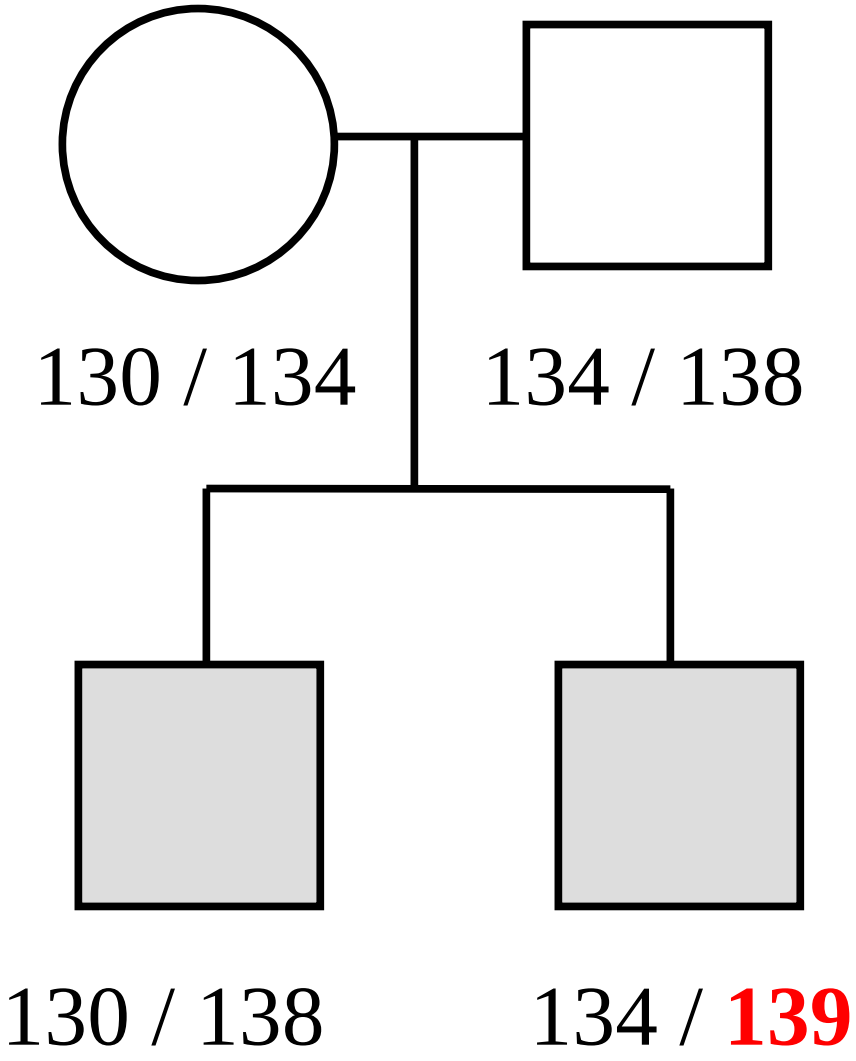

Chromosome 6
Family 884
3m
4m
5m
6m
7m
8m
9m
10m
11m
12m
13m
14m
0
100
200
300
400

CEPH: homozygous regions

Individual	Number of regions	Length (cM)	Total length (cM)
884-01 (father)	10	2 – 23	116
884-02 (mother)	11	5 – 40	165
884: grandparents	2 7 – 10	7 1 – 29	14 58 – 120
884: 12 progeny	5 – 18	2 – 36	62 – 204
102: 14 progeny	4 – 13	1 – 39	85 – 254
1331-12 (GP)	2	1 – 6	7
1416-14 (GP)	4	6 – 29	76
35/100 of the others	1 – 2	0 – 7	

Present

- Finish the past
- Yellowstone wolves
- Sib pairs where parents are unavailable
- Various disease projects


130 / 134
134 / 138
130 / 138
134 / 139

Future

- Large genetic/epidemiological studies
  - Many people
  - Many phenotypes
  - Many environmental covariates
- Highly computational statistical genetic methods
- Various disease projects