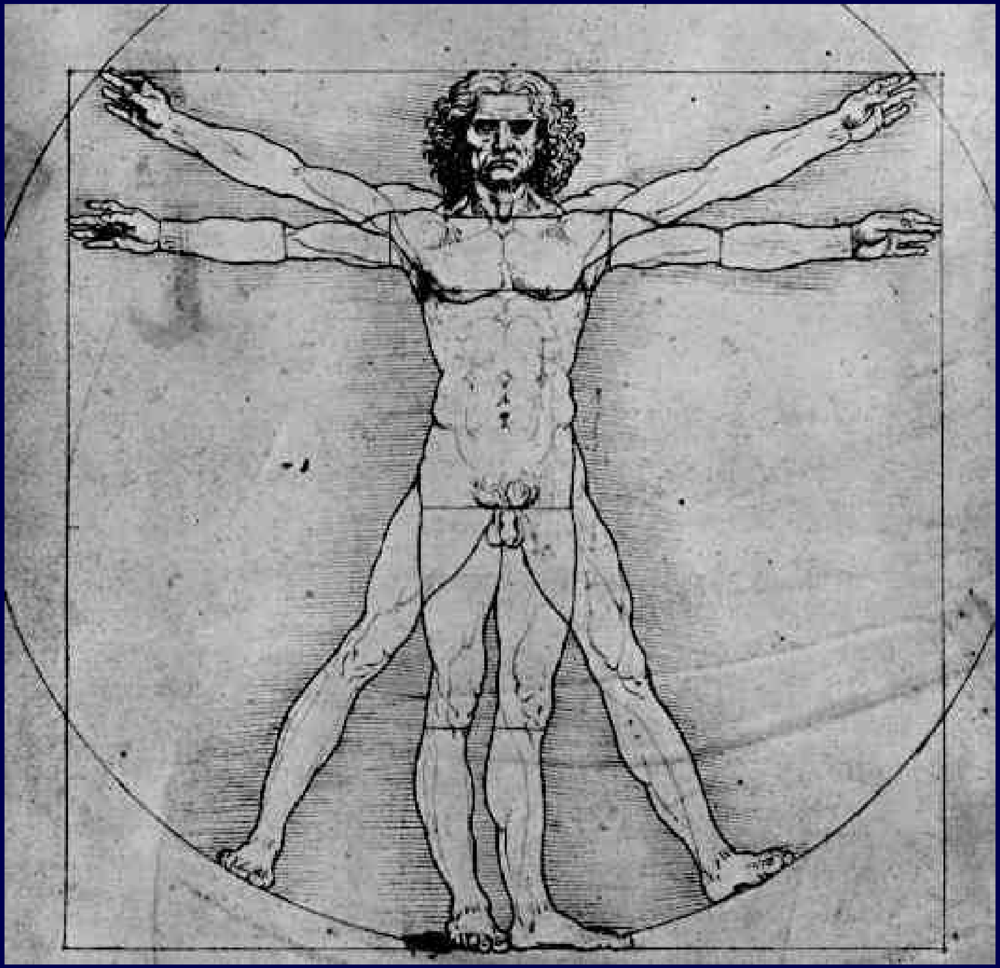
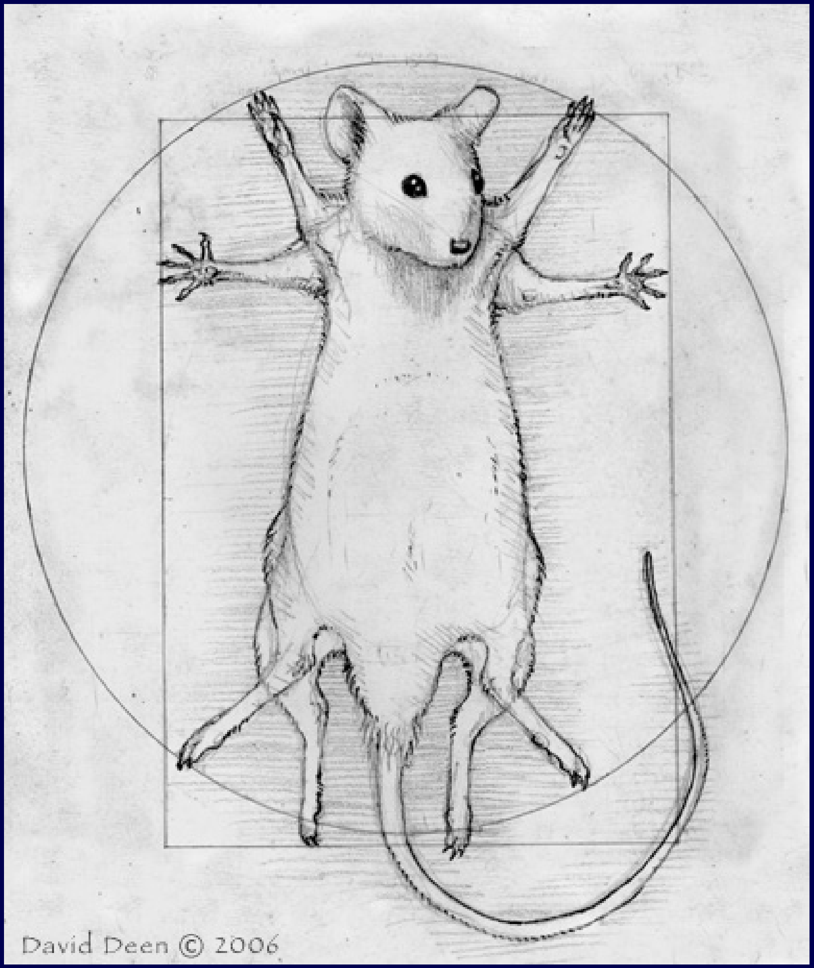
QTL mapping
in experimental crosses

Part I

Karl W Broman
kbroman.org
github.com/kbroman
@kwbroman

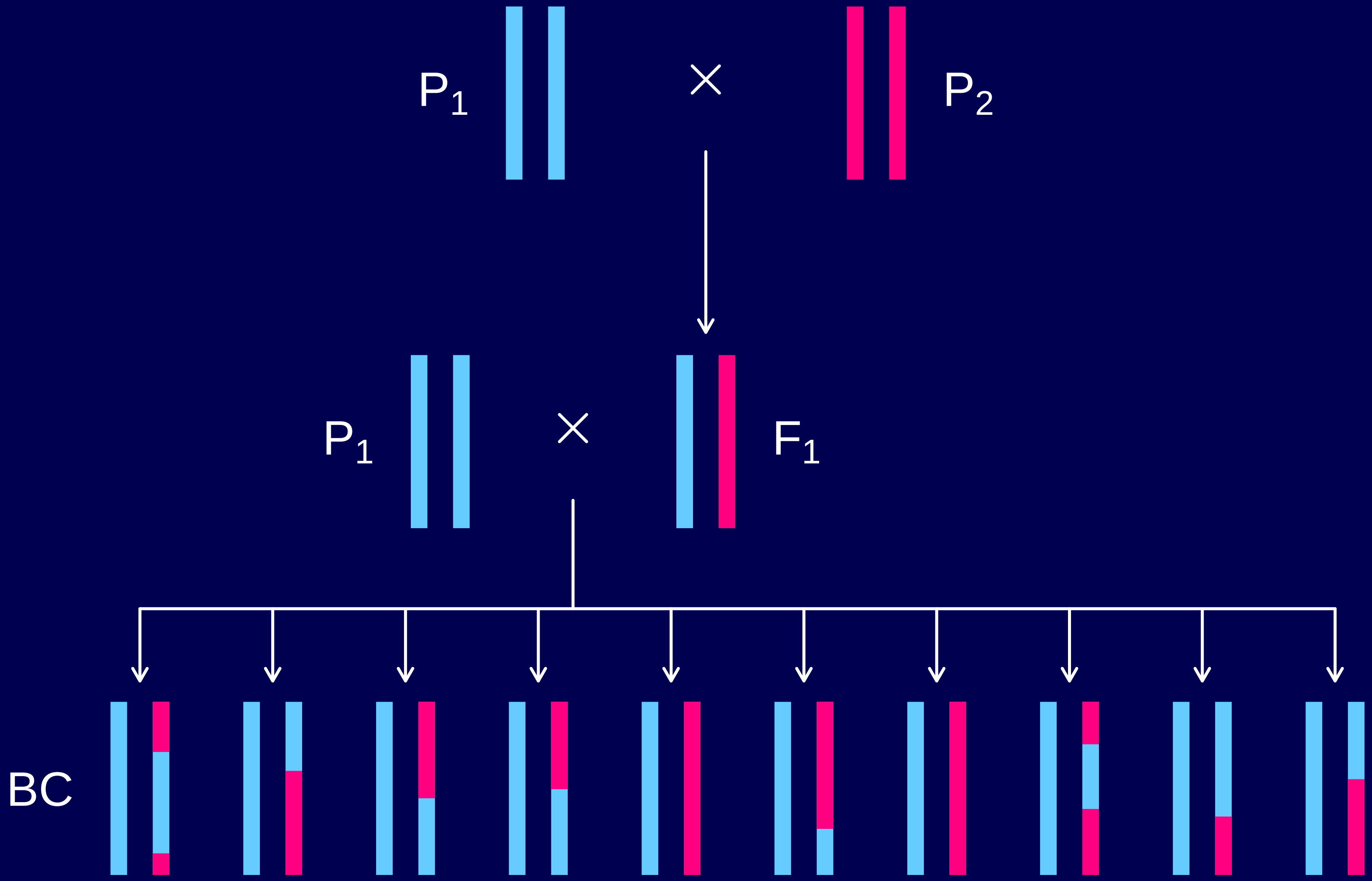
Human vs mouse



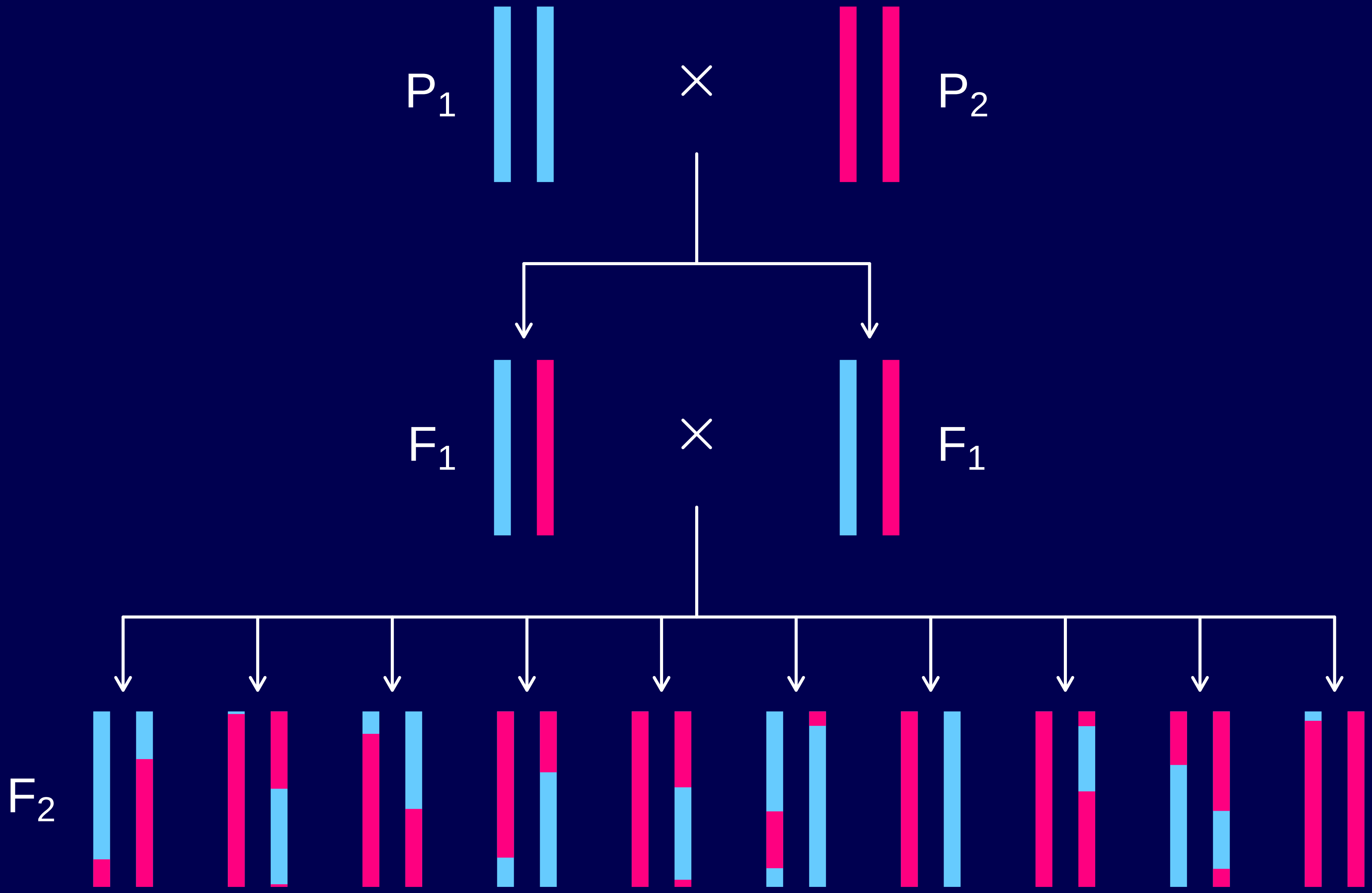


www.daviddeen.com

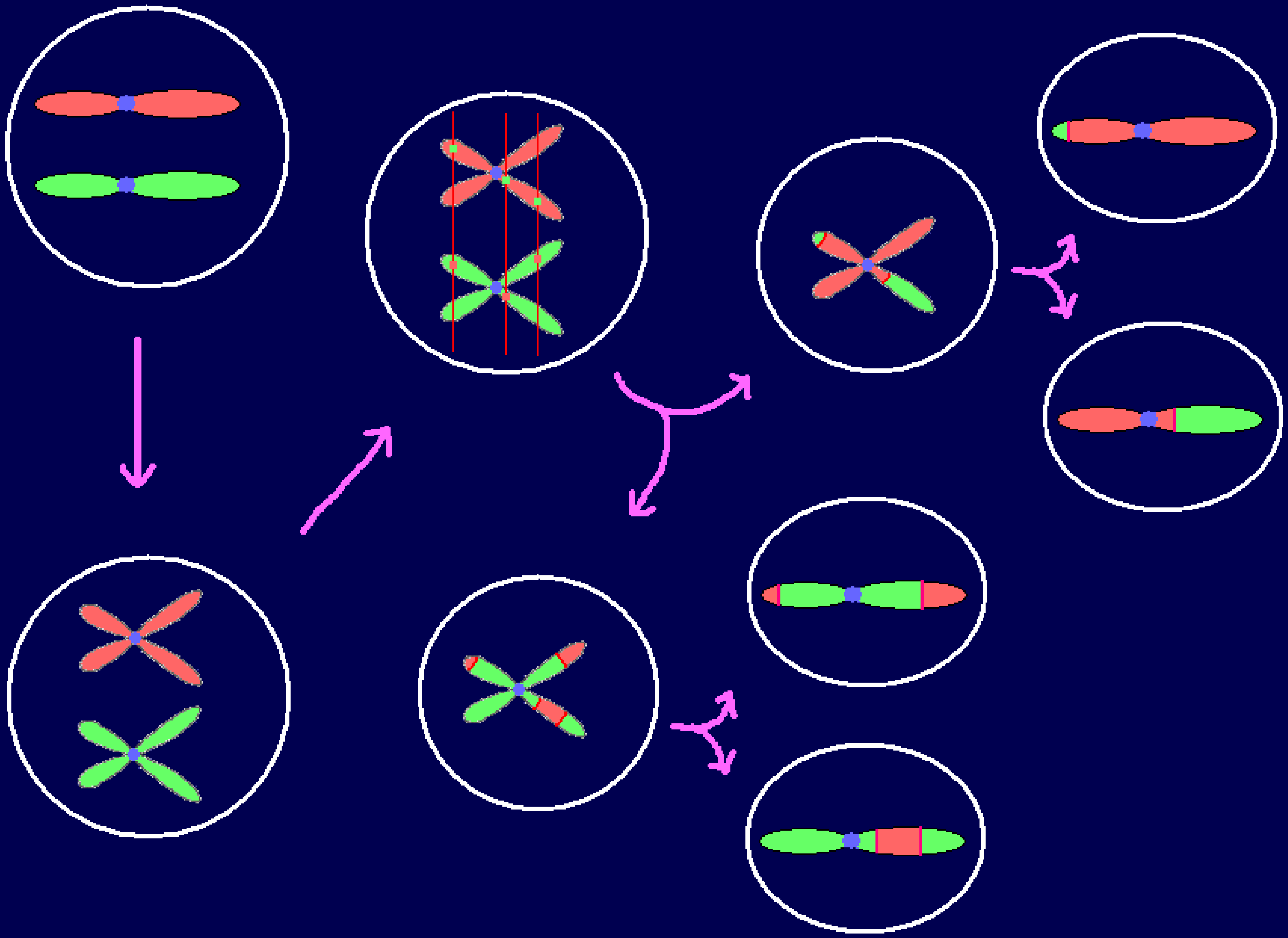
Backcross



Intercross



Meiosis



Genetic distance

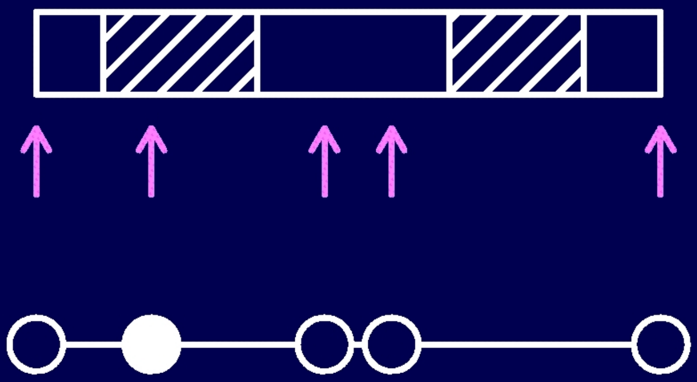
- Genetic distance between two markers (in cM) =

  Average number of crossovers in the interval in 100 meiotic products.

- “Intensity” of the crossover point process

- Recombination rate varies by
  - Organism
  - Sex
  - Chromosome
  - Position on chromosome

Recombination fraction



Crossover process

Marker data

We generally do not observe the locations of crossovers; rather, we observe the grandparental origin of DNA at a set of genetic markers.

Recombination across an interval indicates an odd number of crossovers.

Recombination fraction =

Pr(recombination in interval) = Pr(odd no. XOs in interval)

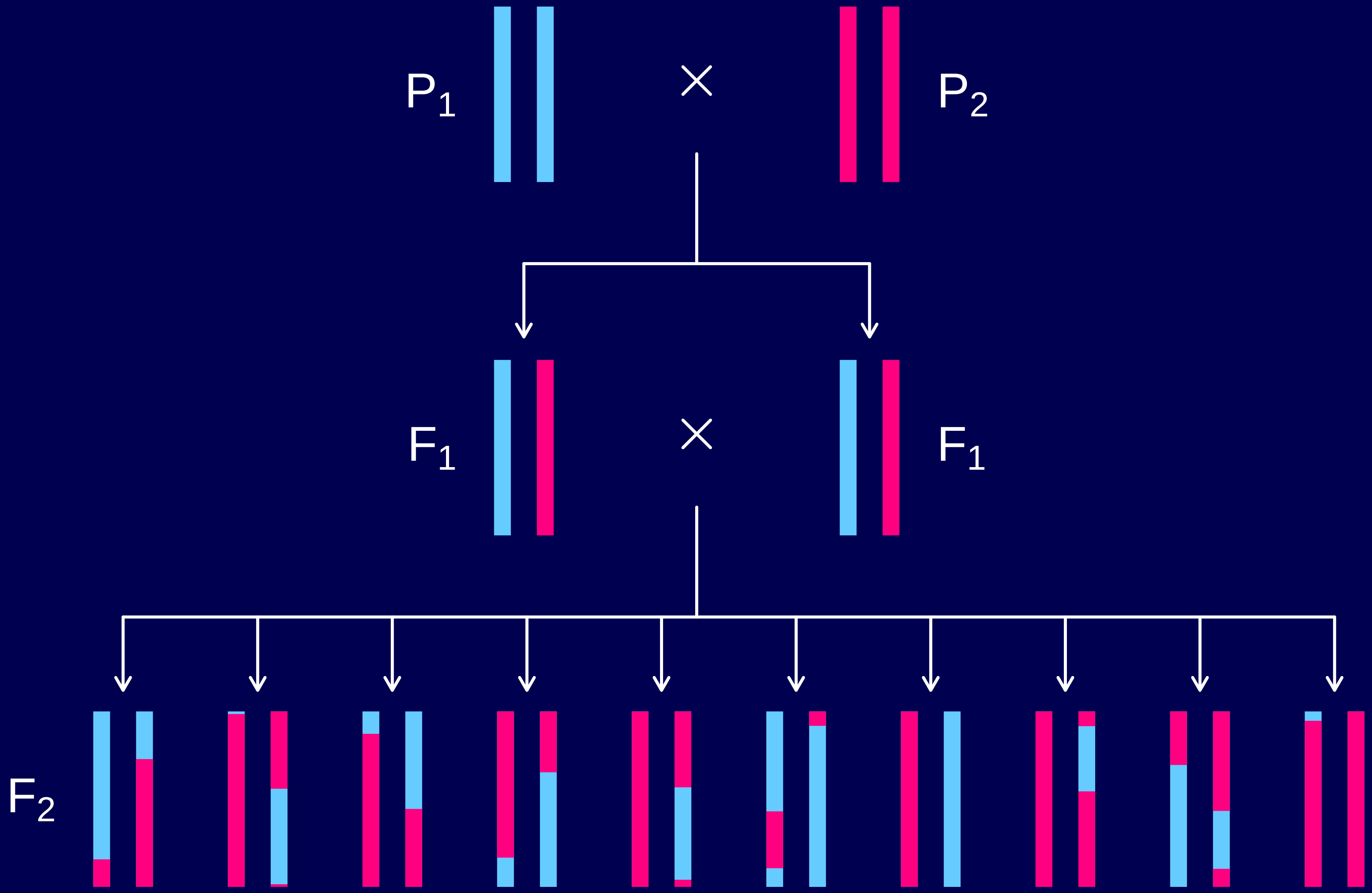
Map functions

- A map function relates the genetic length of an interval and the recombination fraction.

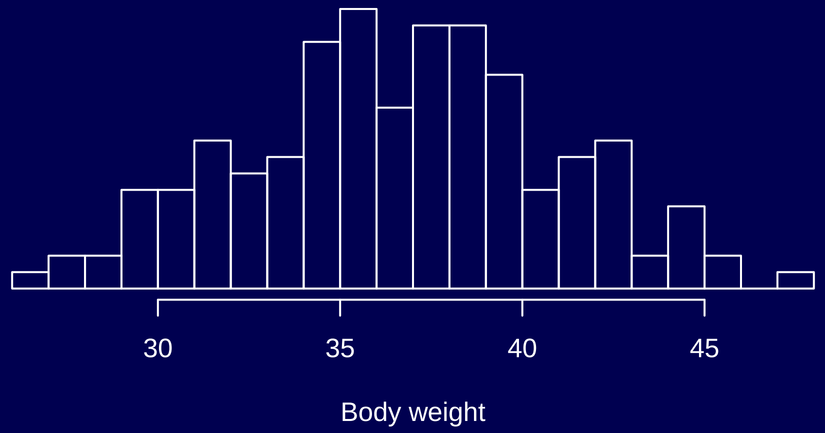
$$r = M(d)$$

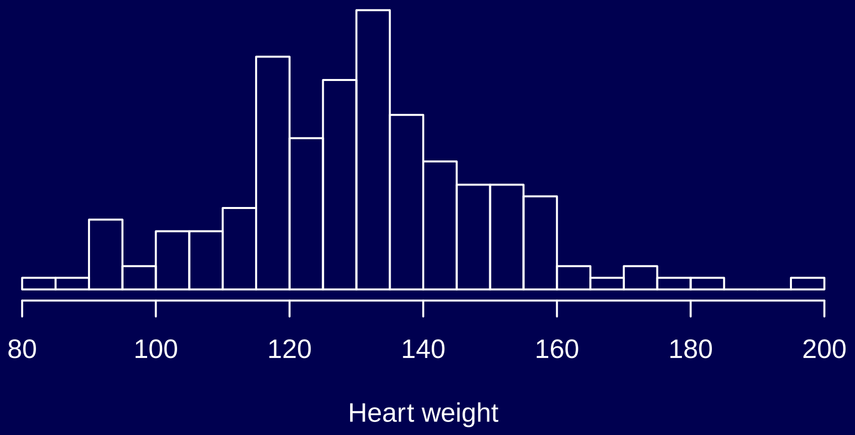
- Map functions are related to crossover interference, but a map function is not sufficient to define the crossover process.
- Haldane map function: no crossover interference
- Kosambi: similar to the level of interference in humans
- Carter-Falconer: similar to the level of interference in mice

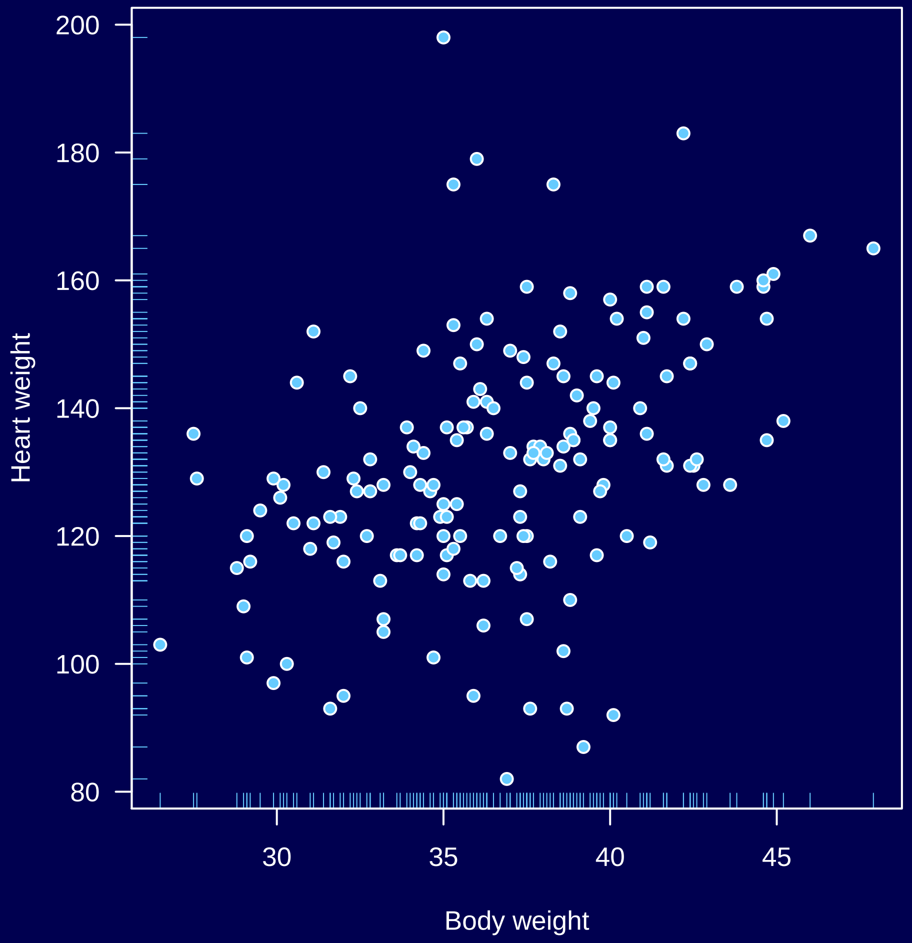
Intercross



Phenotype data

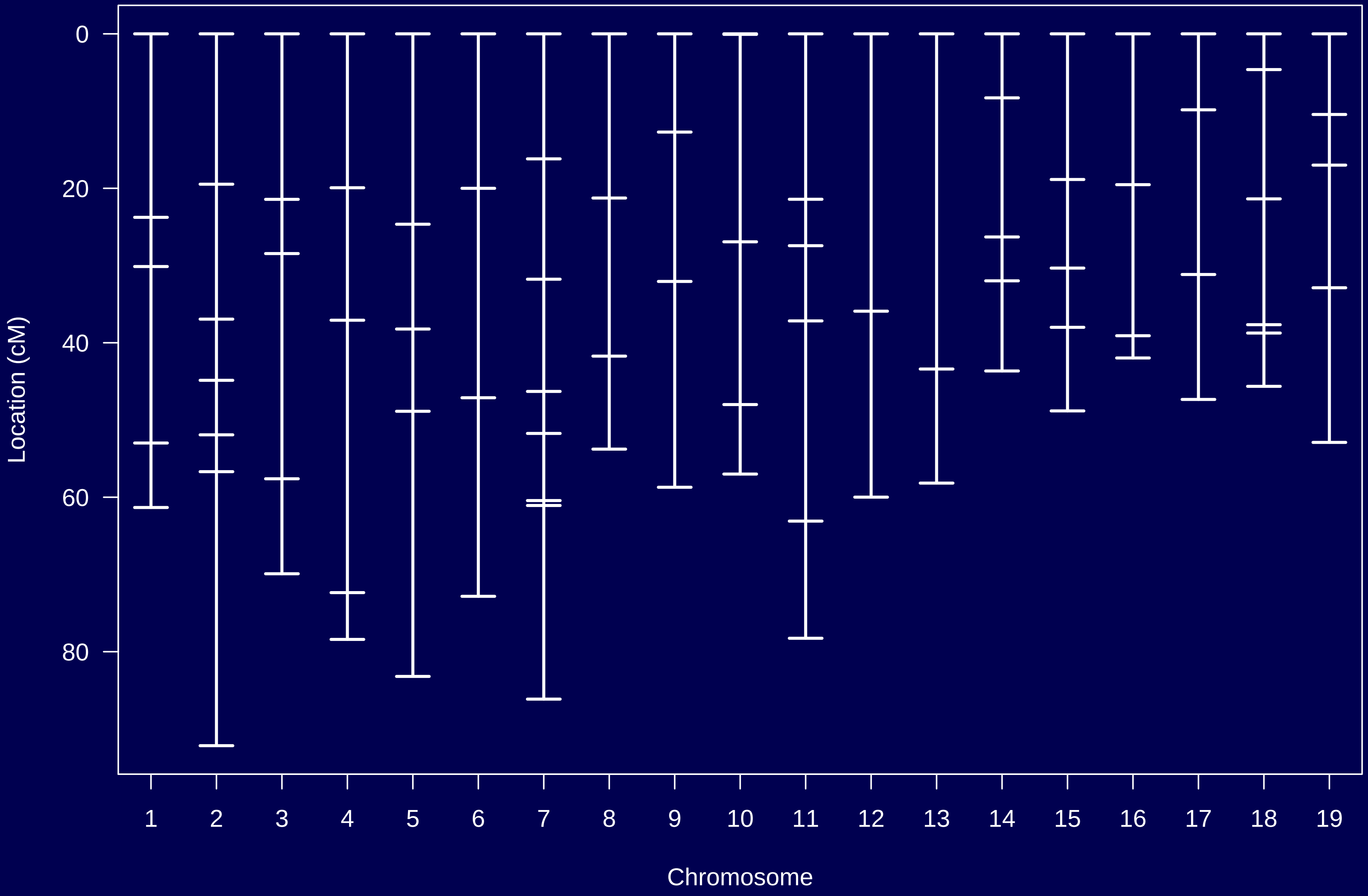




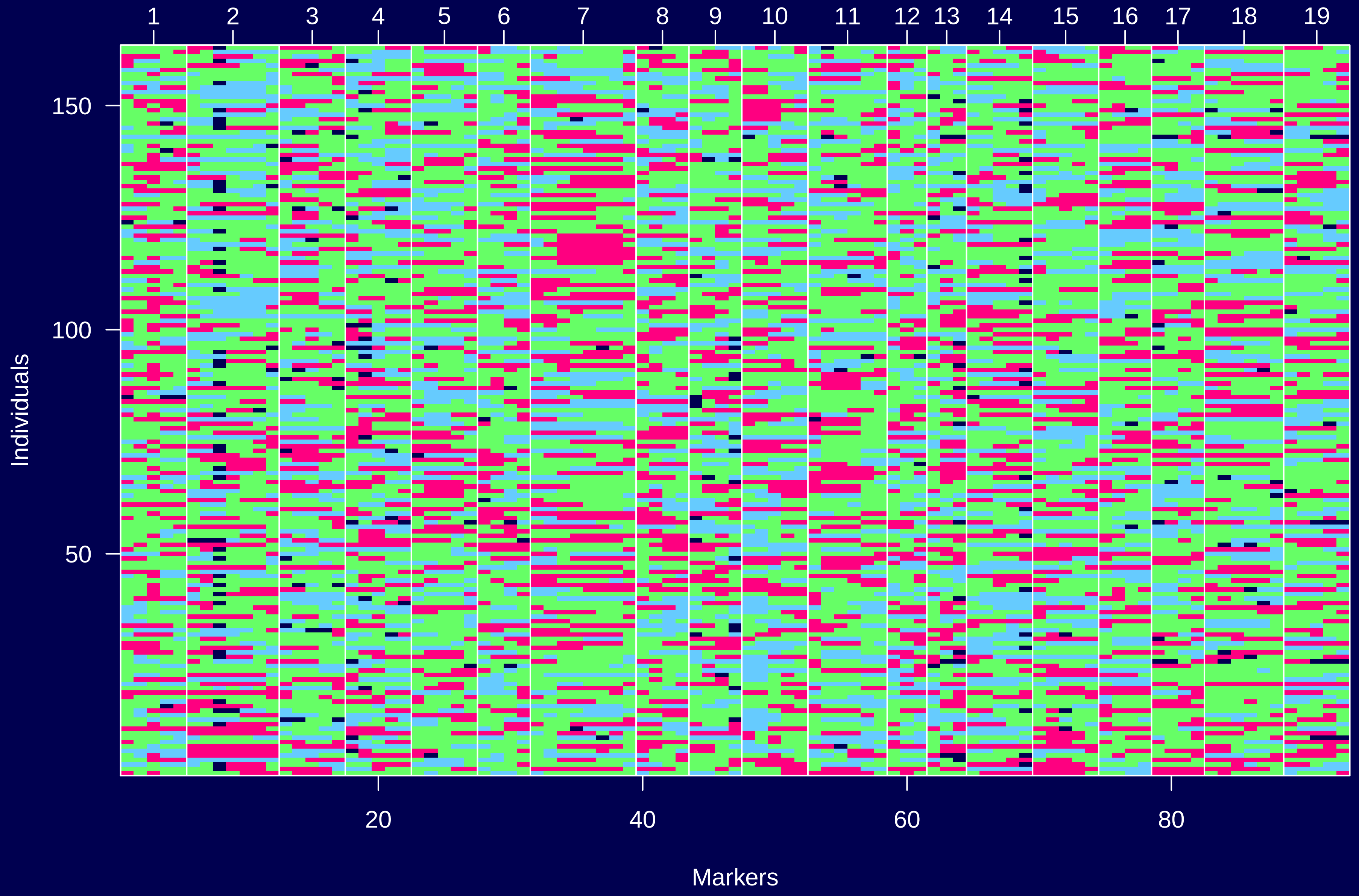


Sugiyama et al. (2002) Physiol Genomics 10:5–12

Genetic map


0
20
40
60
80
Location (cM)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Chromosome

Genotype data



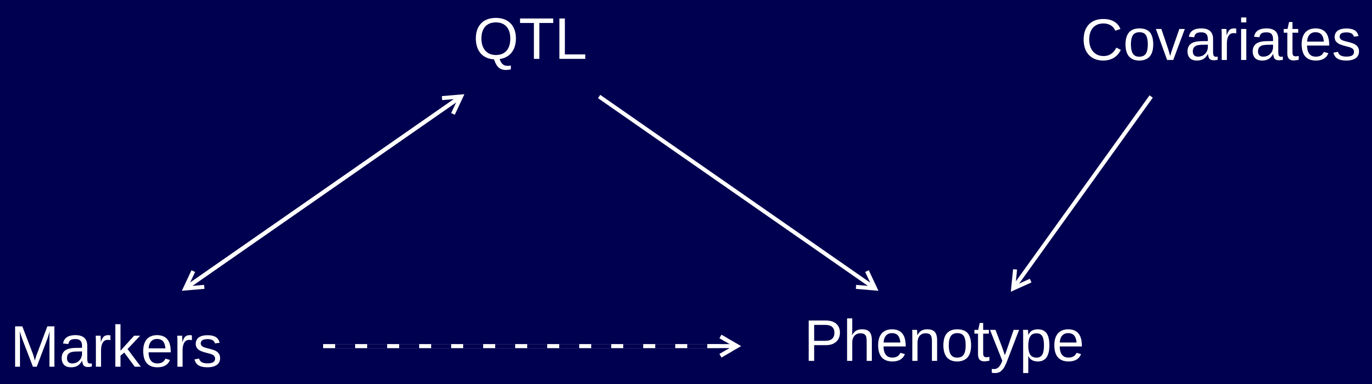
Goals

- Identify quantitative trait loci (QTL)
  (and interactions among QTL)

- Interval estimates of QTL location

- Estimated QTL effects

Statistical structure

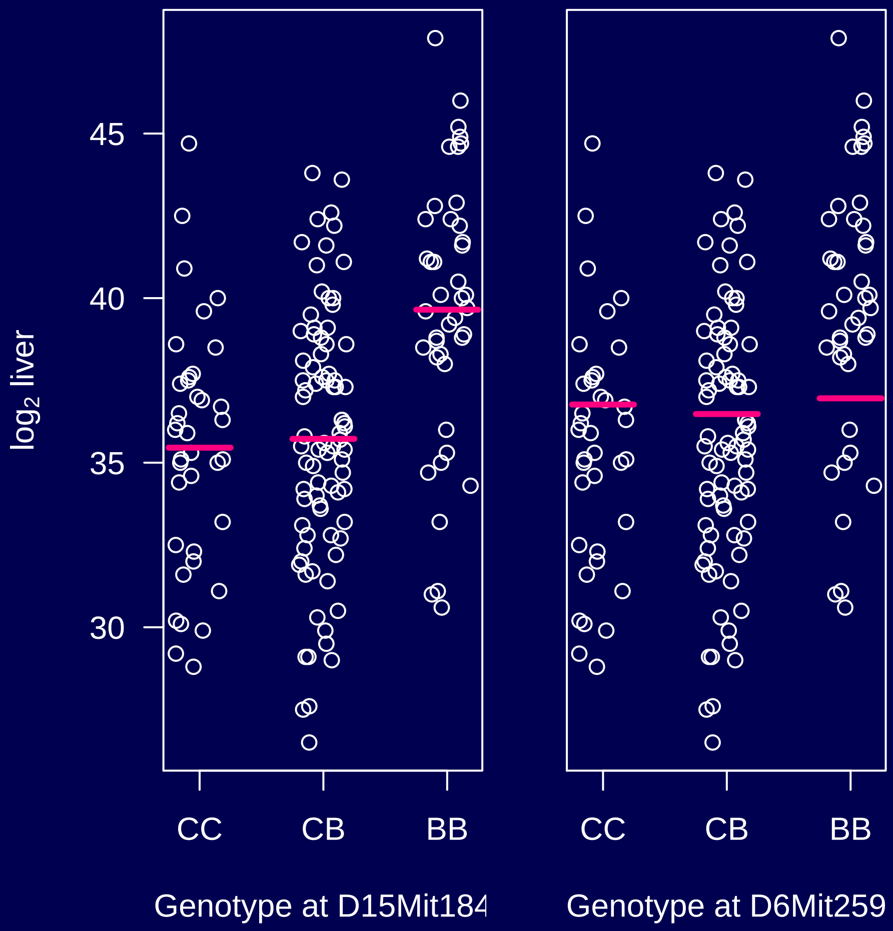


The missing data problem: Markers ⟷ QTL

The model selection problem: QTL, covariates ⟶ phenotype

ANOVA at marker loci

- Also known as marker regression.
- Split mice into groups according to genotype at a marker.
- Do a t-test / ANOVA.
- Repeat for each marker.



ANOVA at marker loci

Advantages

- Simple.
- Easily incorporates covariates.
- Easily extended to more complex models.
- Doesn't require a genetic map.

Disadvantages

- Must exclude individuals with missing genotype data.
- Imperfect information about QTL location.
- Suffers in low density scans.
- Only considers one QTL at a time.

Interval mapping

Lander & Botstein (1989)

- Assume a single QTL model.

- Each position in the genome, one at a time, is posited as the putative QTL.

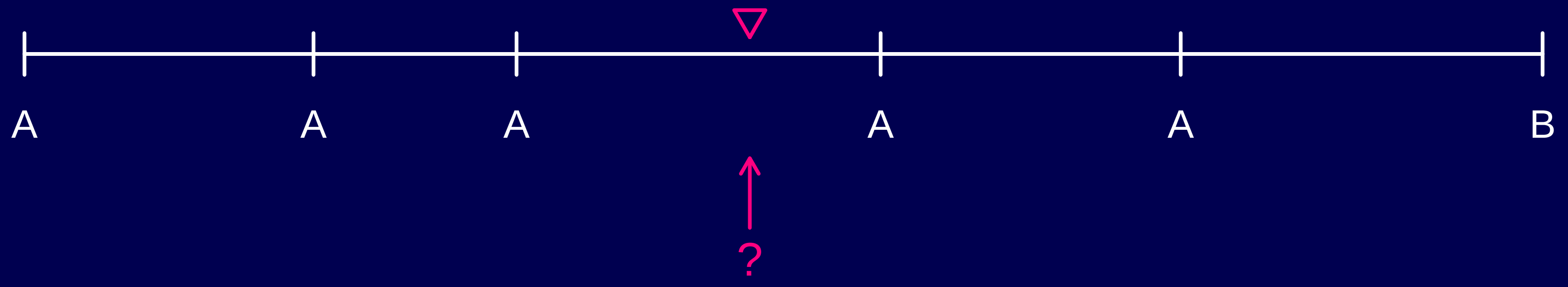
- Let $\mathsf{q} = 1/0$ if the (unobserved) QTL genotype is BB/AB.

  (Or 2/1/0 if the QTL genotype is BB/AB/AA in an intercross.)

  Assume $\mathsf{y}|\mathsf{q} \sim \mathsf{N}(\mu_\mathsf{q}, \sigma)$

- Given genotypes at linked markers, $\mathsf{y} \sim$ mixture of normal dist'ns with mixing proportions $\Pr(\mathsf{q} \mid \text{marker data})$

Genotype probabilities



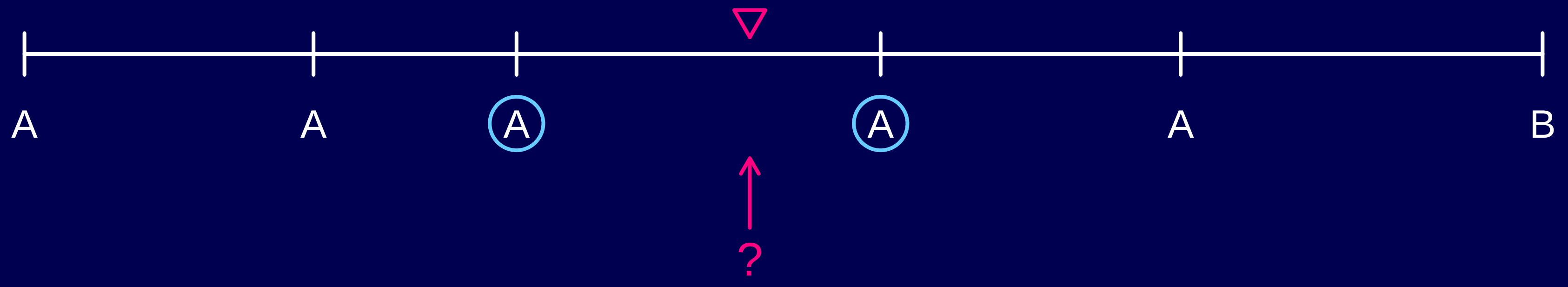
Calculate $\text{Pr}(\mathbf{q} \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

Genotype probabilities



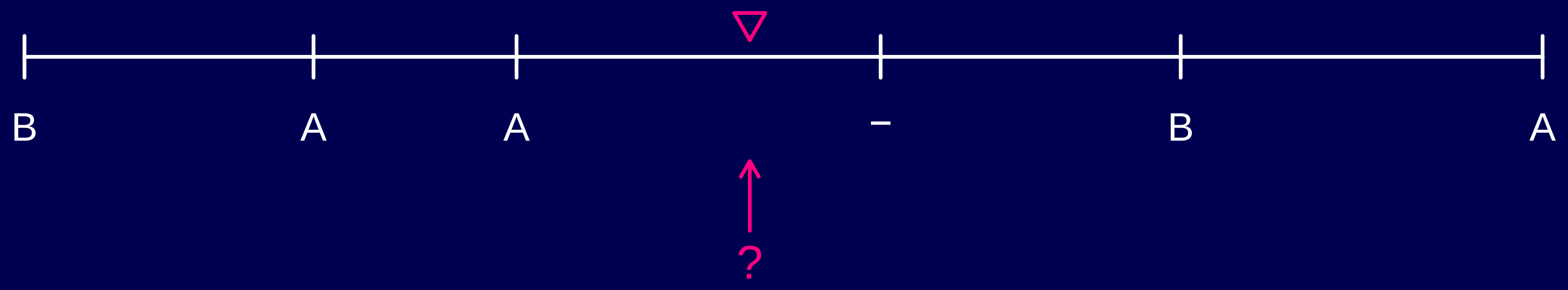
Calculate $\Pr(\mathbf{q} \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

Genotype probabilities



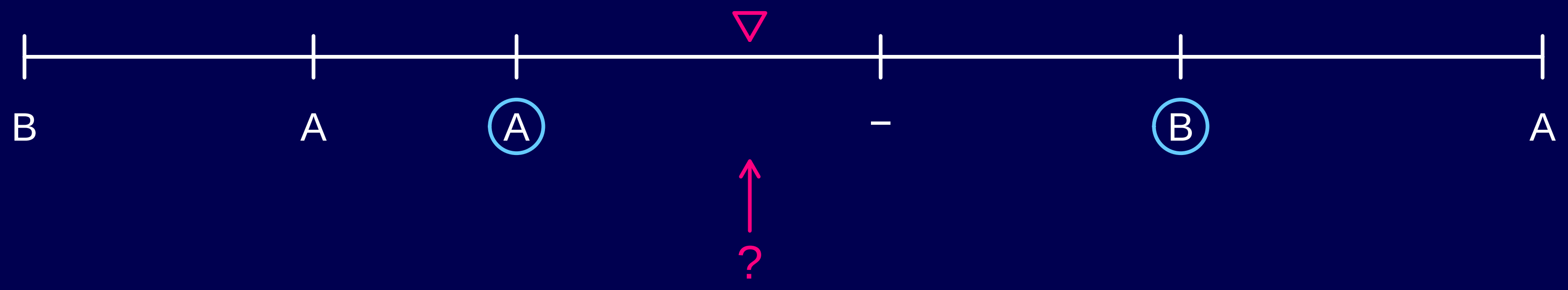
Calculate $\Pr(\mathbf{q} \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

Genotype probabilities



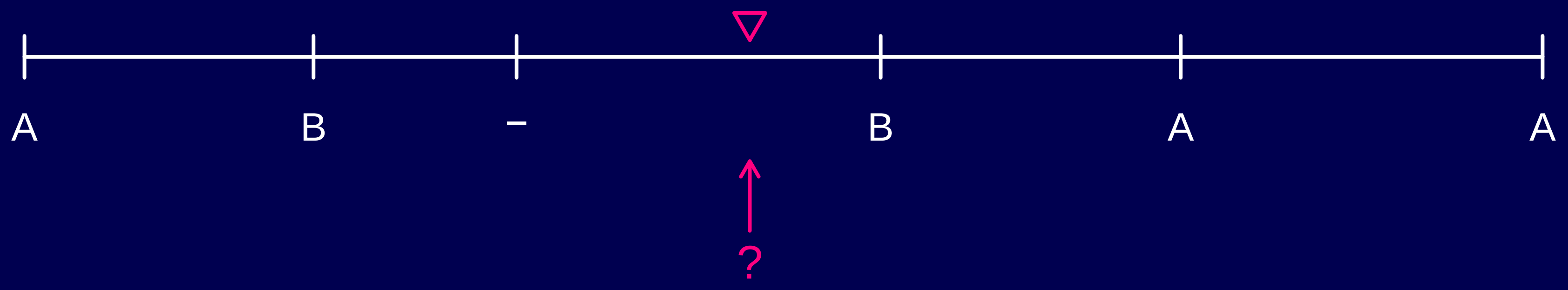
Calculate $\Pr(\mathbf{q} \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

Genotype probabilities



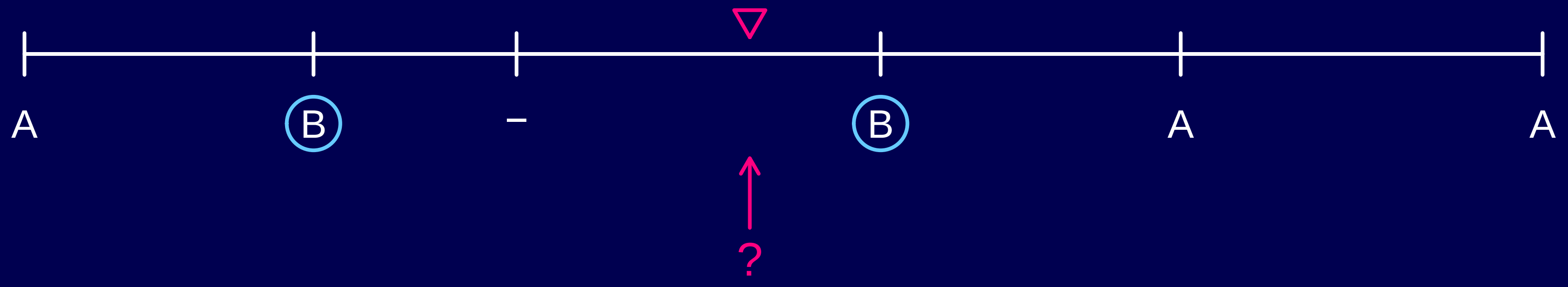
Calculate $\mathrm{Pr}(\mathbf{q} \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

Genotype probabilities



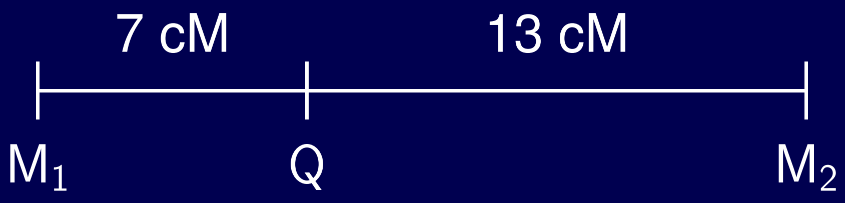
Calculate $\mathrm{Pr}(\mathbf{q} \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

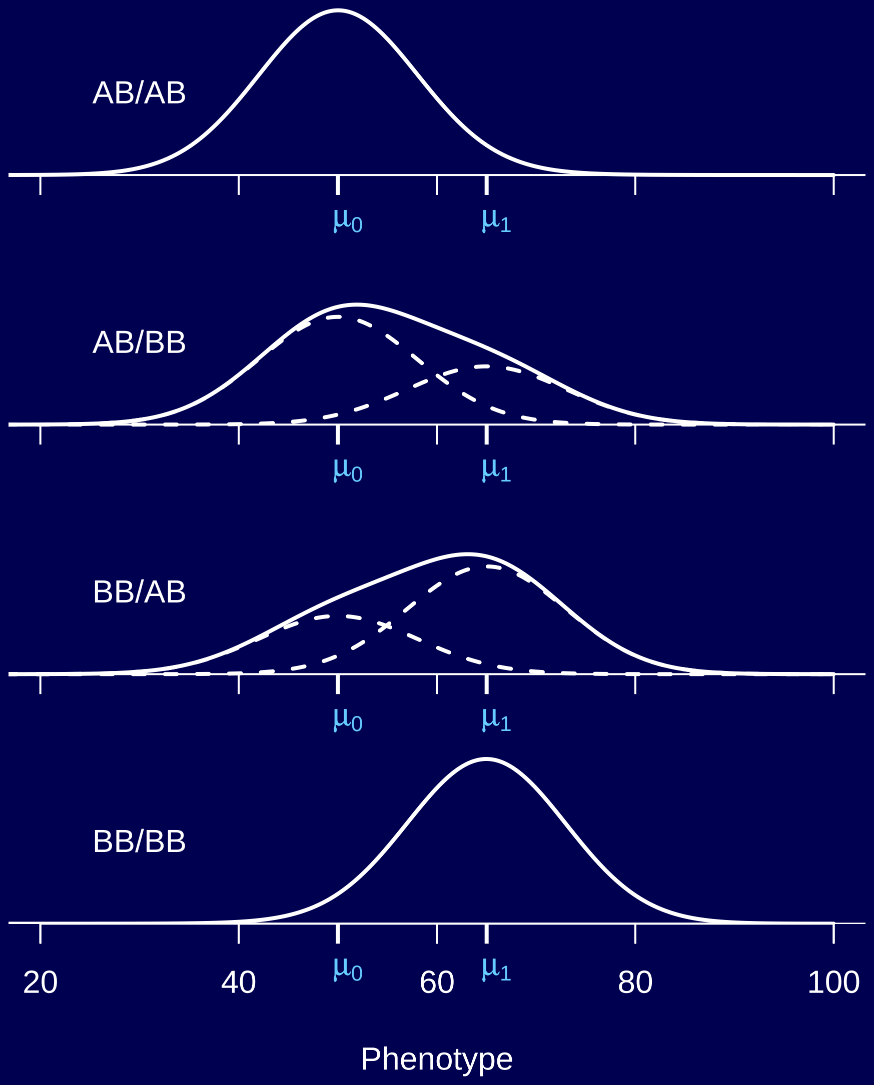
The normal mixtures



7 cM 13 cM

M_1 Q M_2

- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right shows the distributions of the phenotype conditional on the genotypes at the two markers.
- The dashed curves correspond to the components of the mixtures.



Interval mapping

Let $\mathsf{p}_{\mathsf{ij}} = \Pr(\mathsf{q}_{\mathsf{i}} = \mathsf{j} | \text{marker data})$

$\mathsf{y}_{\mathsf{i}} | \mathsf{q}_{\mathsf{i}} \sim \mathsf{N}(\mu_{\mathsf{q}_{\mathsf{i}}}, \sigma^2)$

$\Pr(\mathsf{y}_{\mathsf{i}} | \text{marker data}, \mu_0, \mu_1, \sigma) = \sum_{\mathsf{j}} \mathsf{p}_{\mathsf{ij}} \, \mathsf{f}(\mathsf{y}_{\mathsf{i}}; \mu_{\mathsf{j}}, \sigma)$

where $\mathsf{f}(\mathsf{y}; \mu, \sigma) = \exp[-(\mathsf{y} - \mu)^2/(2\sigma^2)]/\sqrt{2\pi\sigma^2}$

Log likelihood: $\mathsf{l}(\mu_0, \mu_1, \sigma) = \sum_{\mathsf{i}} \log \Pr(\mathsf{y}_{\mathsf{i}} | \text{marker data}, \mu_0, \mu_1, \sigma)$

Maximum likelihood estimates (MLEs) of μ_0, μ_1, σ:

values for which $\mathsf{l}(\mu_0, \mu_1, \sigma)$ is maximized.

EM algorithm

Dempster et al. (1977)

E step:

Let $$\mathrm{w}_{\mathrm{ij}}^{(\mathrm{k})} = \Pr(\mathrm{q_i} = \mathrm{j} | \mathrm{y_i}, \text{marker data}, \hat{\mu}_0^{(\mathrm{k}-1)}, \hat{\mu}_1^{(\mathrm{k}-1)}, \hat{\sigma}^{(\mathrm{k}-1)})$$

$$= \frac{\mathrm{p_{ij}}\, \mathrm{f}(\mathrm{y_i}; \hat{\mu}_\mathrm{j}^{(\mathrm{k}-1)}, \hat{\sigma}^{(\mathrm{k}-1)})}{\sum_\mathrm{j} \mathrm{p_{ij}}\, \mathrm{f}(\mathrm{y_i}; \hat{\mu}_\mathrm{j}^{(\mathrm{k}-1)}, \hat{\sigma}^{(\mathrm{k}-1)})}$$

M step:

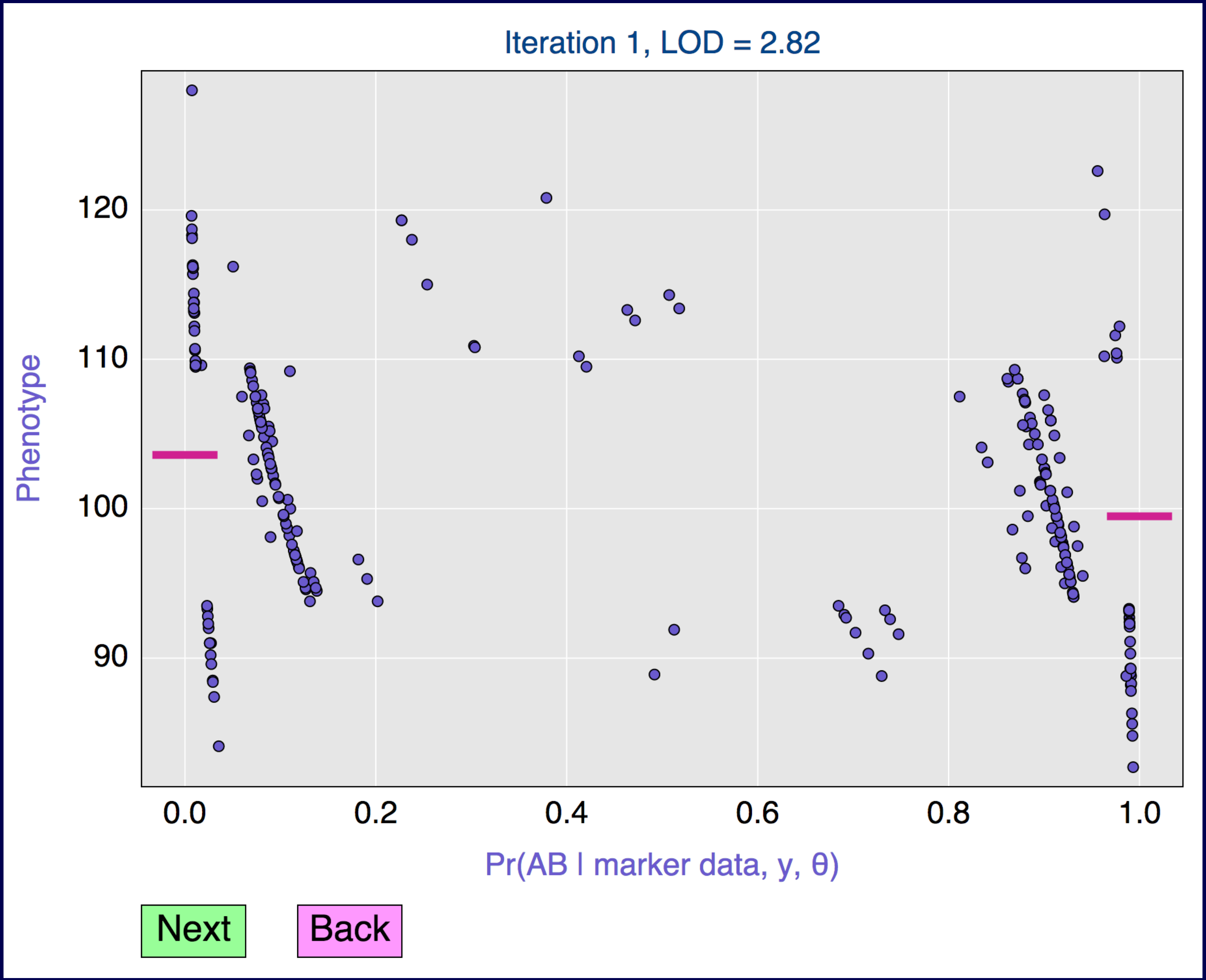
Let $$\hat{\mu}_\mathrm{j}^{(\mathrm{k})} = \sum_\mathrm{i} \mathrm{y_i} \mathrm{w}_{\mathrm{ij}}^{(\mathrm{k})} / \sum_\mathrm{i} \mathrm{w}_{\mathrm{ij}}^{(\mathrm{k})}$$

$$\hat{\sigma}^{(\mathrm{k})} = \sqrt{\sum_\mathrm{i} \sum_\mathrm{j} \mathrm{w}_{\mathrm{ij}}^{(\mathrm{k})} (\mathrm{y_i} - \hat{\mu}_\mathrm{j}^{(\mathrm{k})})^2 / \mathrm{n}}$$

The algorithm:

Start with $\mathrm{w}_{\mathrm{ij}}^{(1)} = \mathrm{p_{ij}}$; iterate the E & M steps until convergence.

Interactive illustration



LOD scores

The LOD score is a measure of the strength of evidence for the presence of a QTL at a particular location.

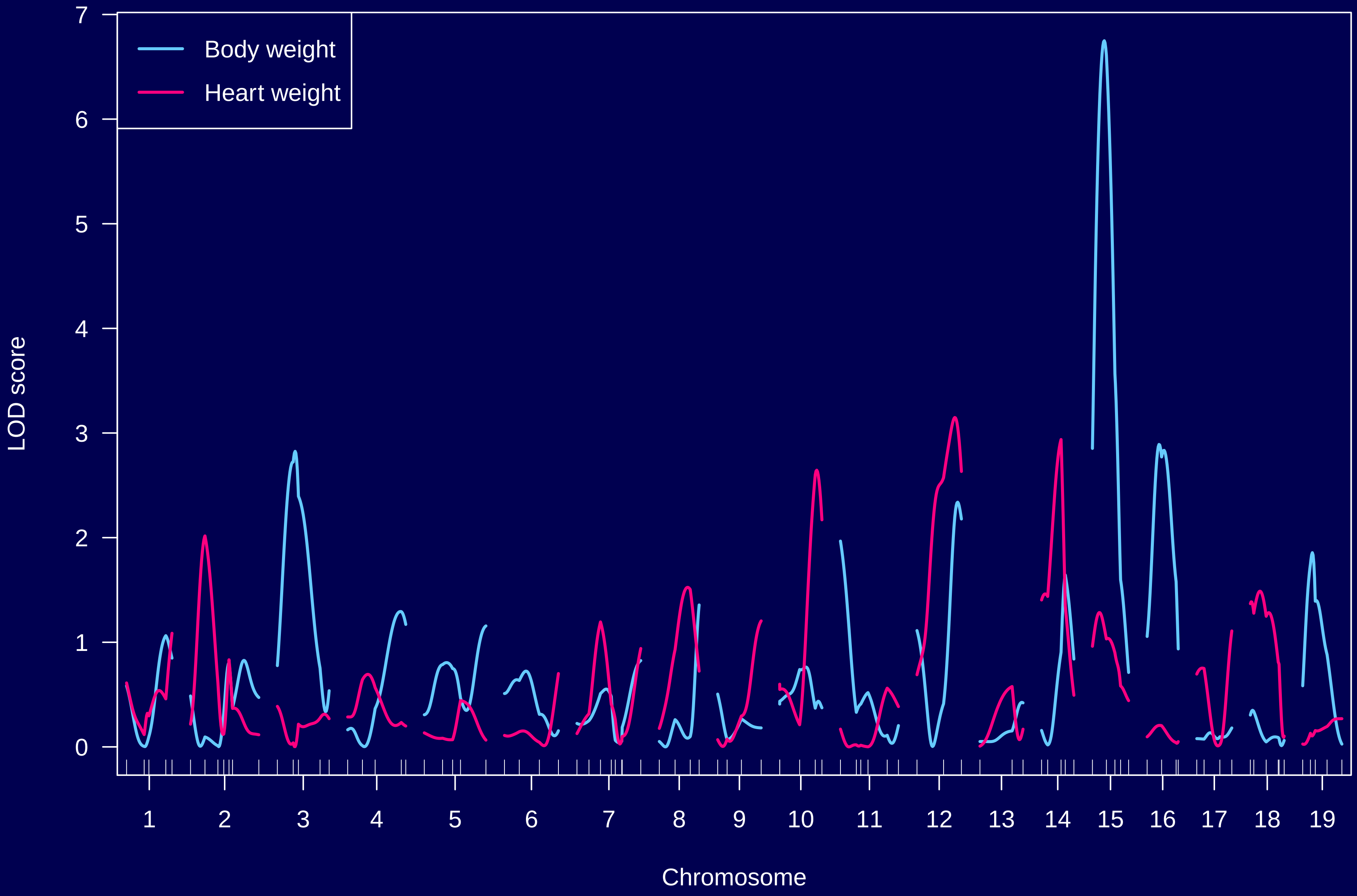
LOD$(\lambda) = \log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position λ versus that of no QTL

$$= \log_{10} \left\{ \frac{\Pr(\text{y}|\text{QTL at } \lambda, \hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_{\lambda})}{\Pr(\text{y}|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

$\hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_{\lambda}$ are the MLEs, assuming a single QTL at position λ.

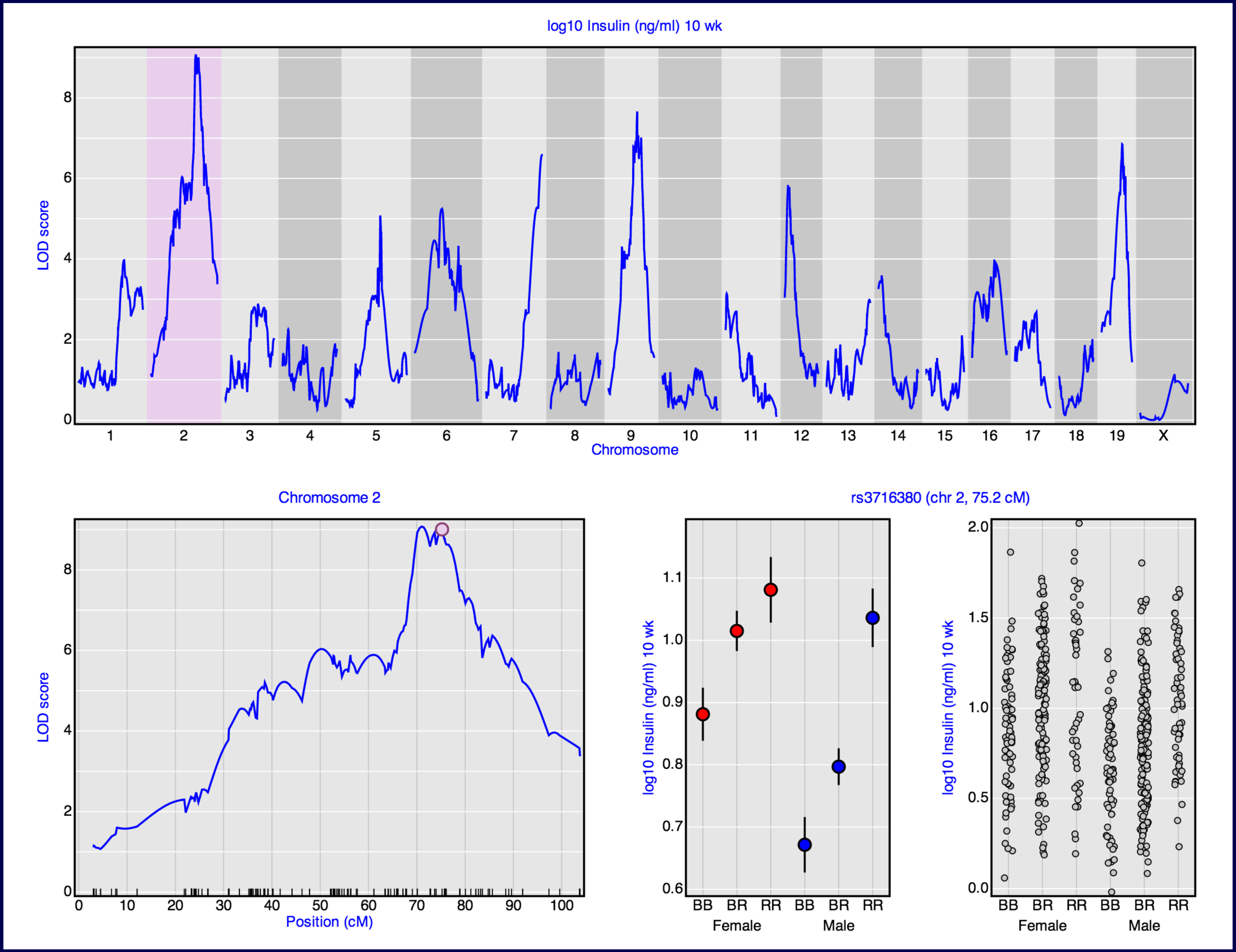
No QTL model: The phenotypes are independent and identically distributed (iid) N(μ, σ^2).

LOD curves



Body weight
Heart weight
LOD score
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Chromosome

Interactive plot



Interval mapping

Advantages

- Takes proper account of missing data.
- Allows examination of positions between markers.
- Gives improved estimates of QTL effects.
- Provides pretty graphs.

Disadvantages

- Increased computation time.
- Requires specialized software.
- Difficult to generalize.
- Only considers one QTL at a time.

LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

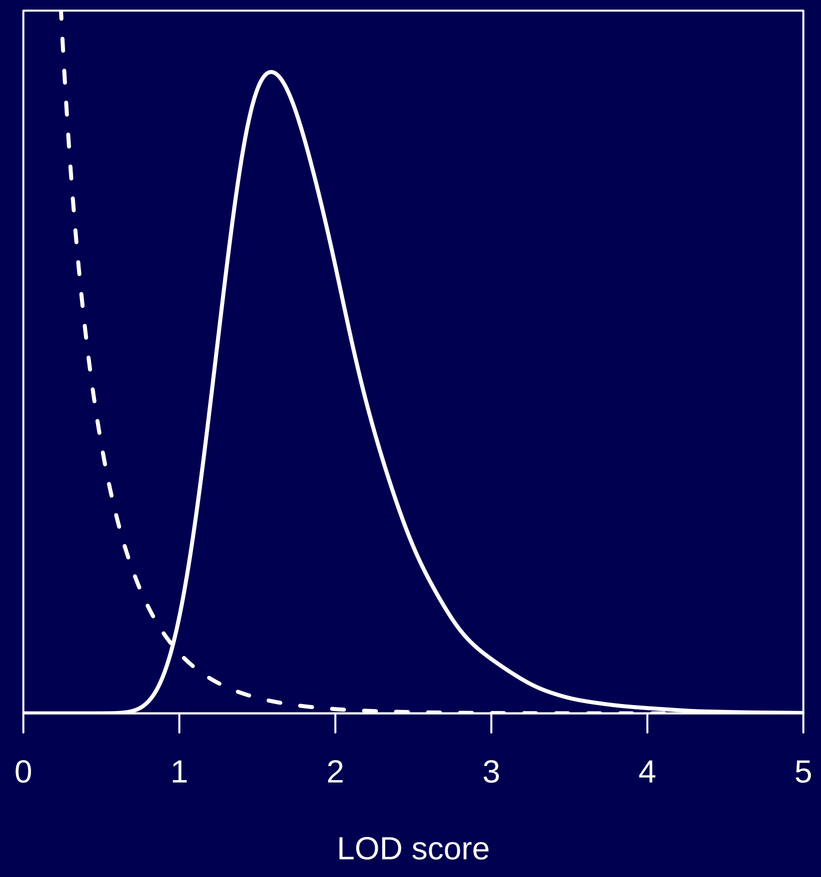
LOD threshold = 95 %ile of distr’n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:

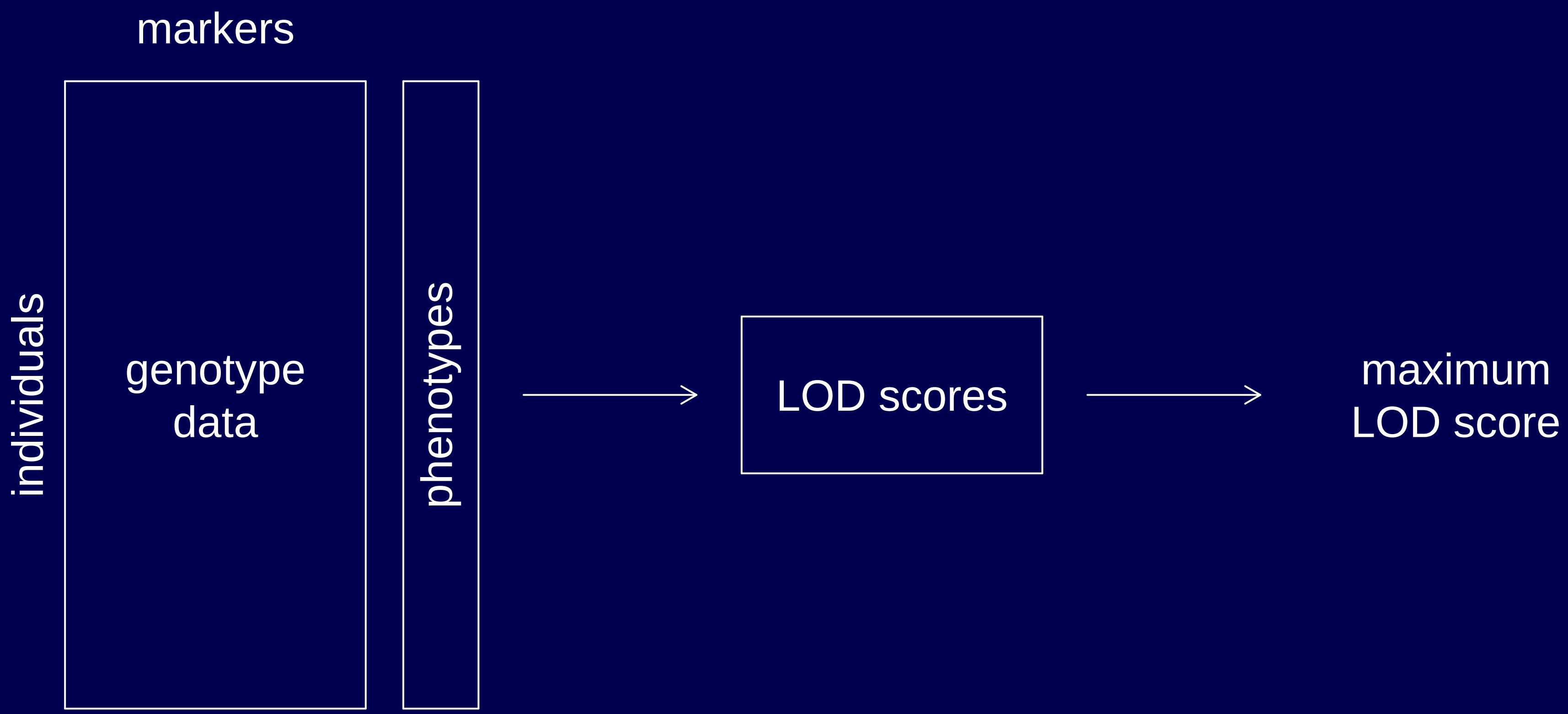
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)

Null distribution of the LOD score

- Null distribution derived by computer simulation of backcross with genome of typical size.
- Dashed curve: distribution of LOD score at any one point.
- Solid curve: distribution of maximum LOD score, genome-wide.



Permutation test



markers

individuals

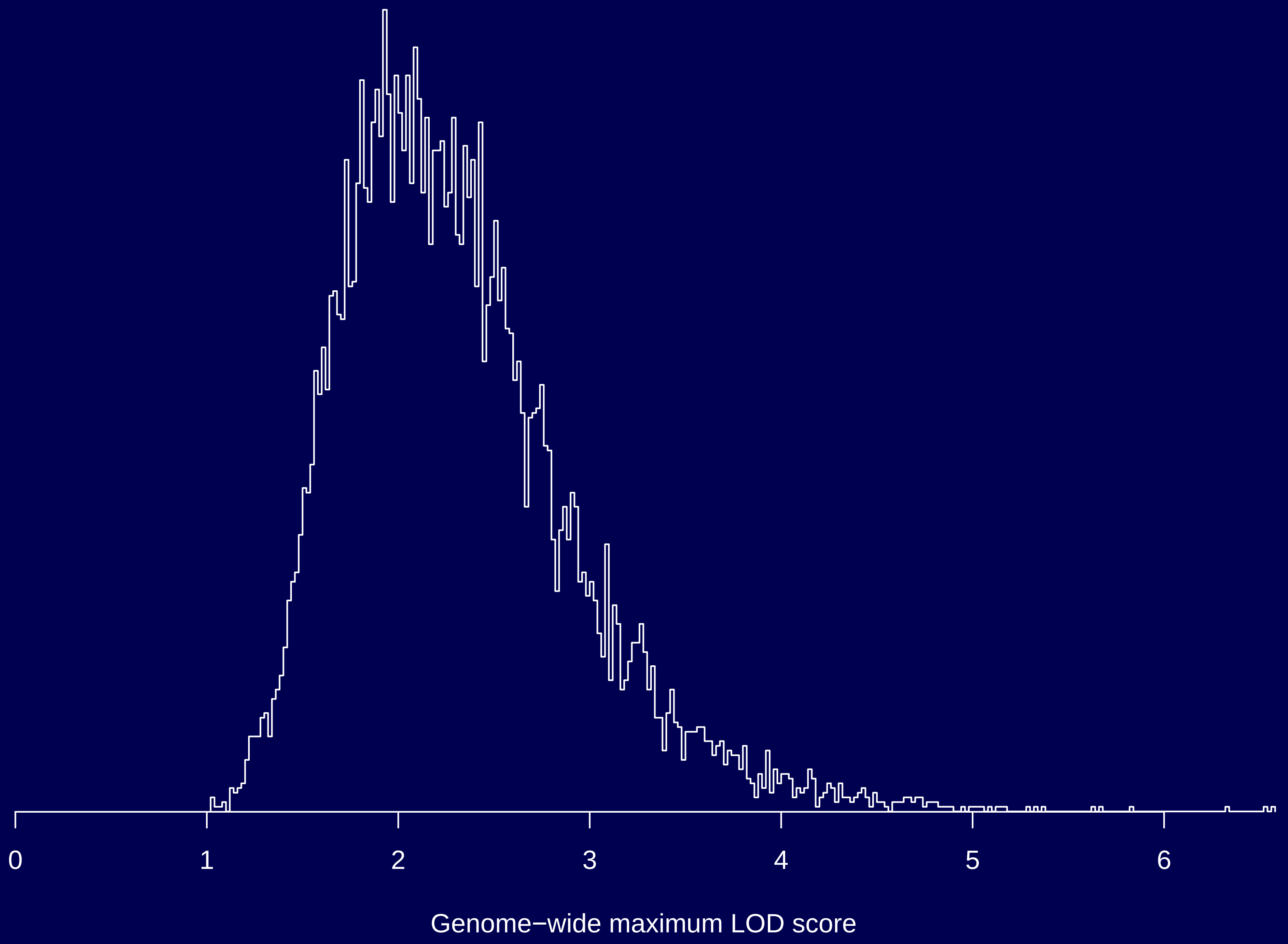
genotype data

phenotypes

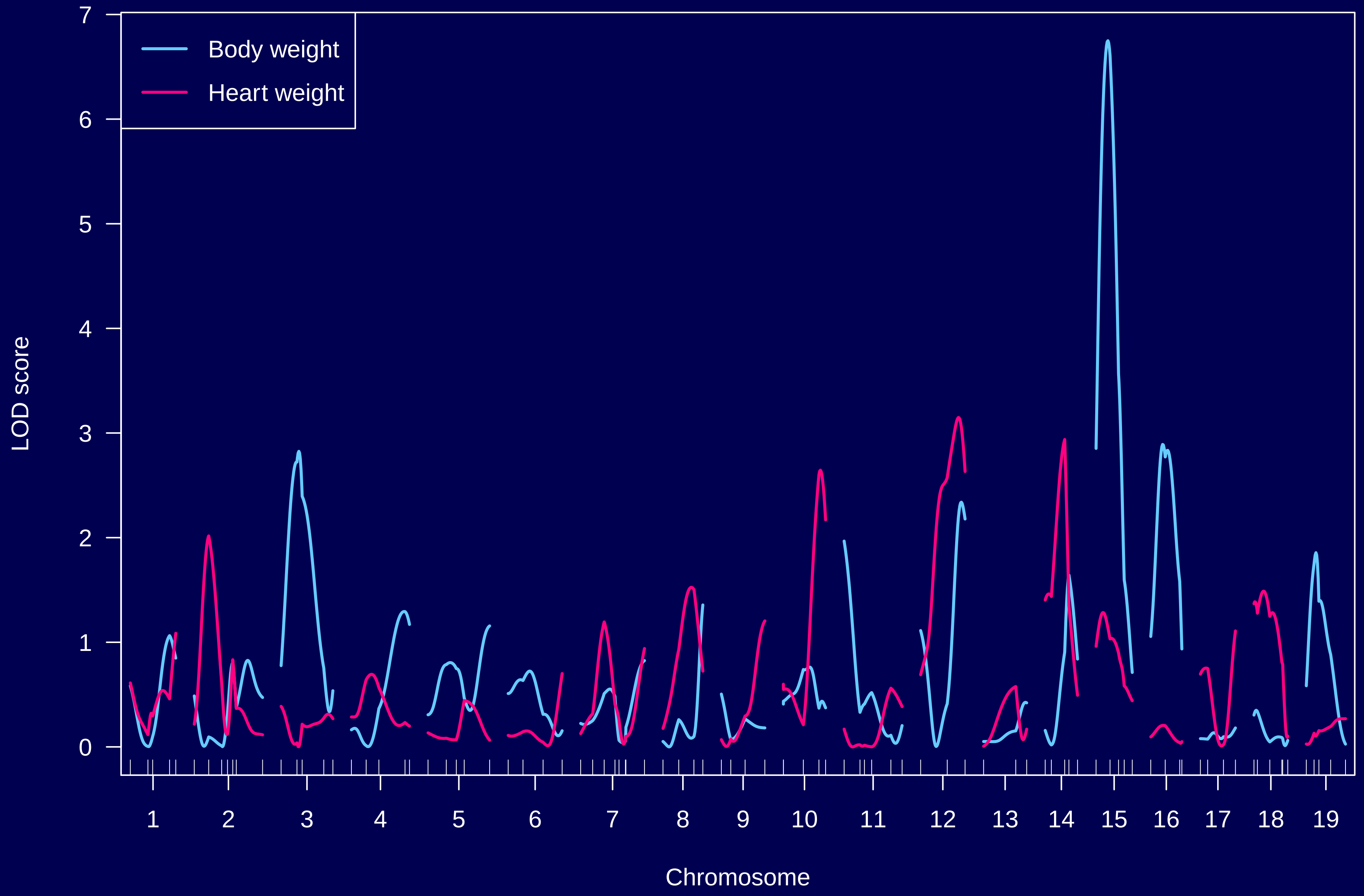
LOD scores

maximum LOD score

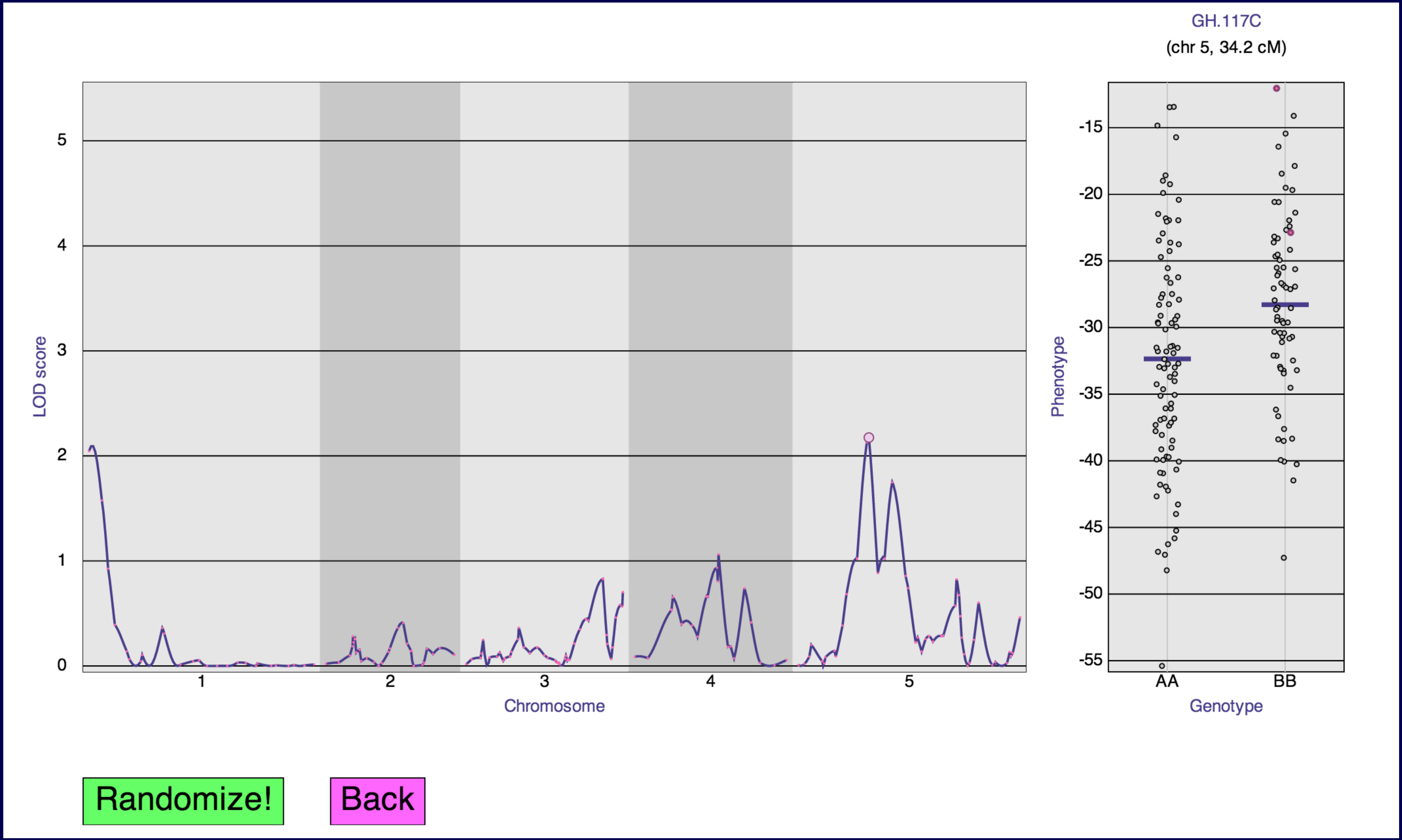
Permutation results



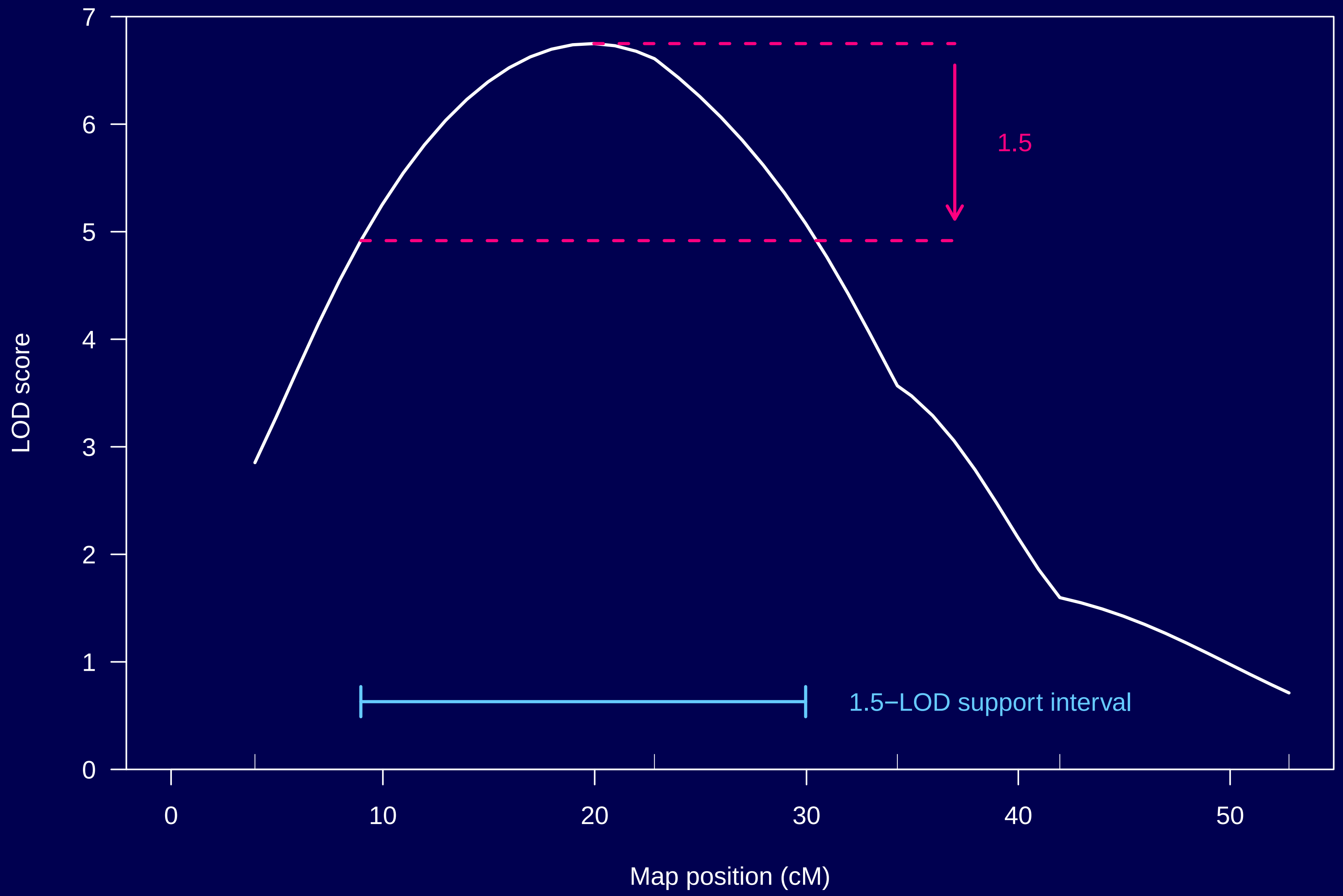
LOD curves


Body weight
Heart weight
LOD score
0
1
2
3
4
5
6
7
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Interactive plot



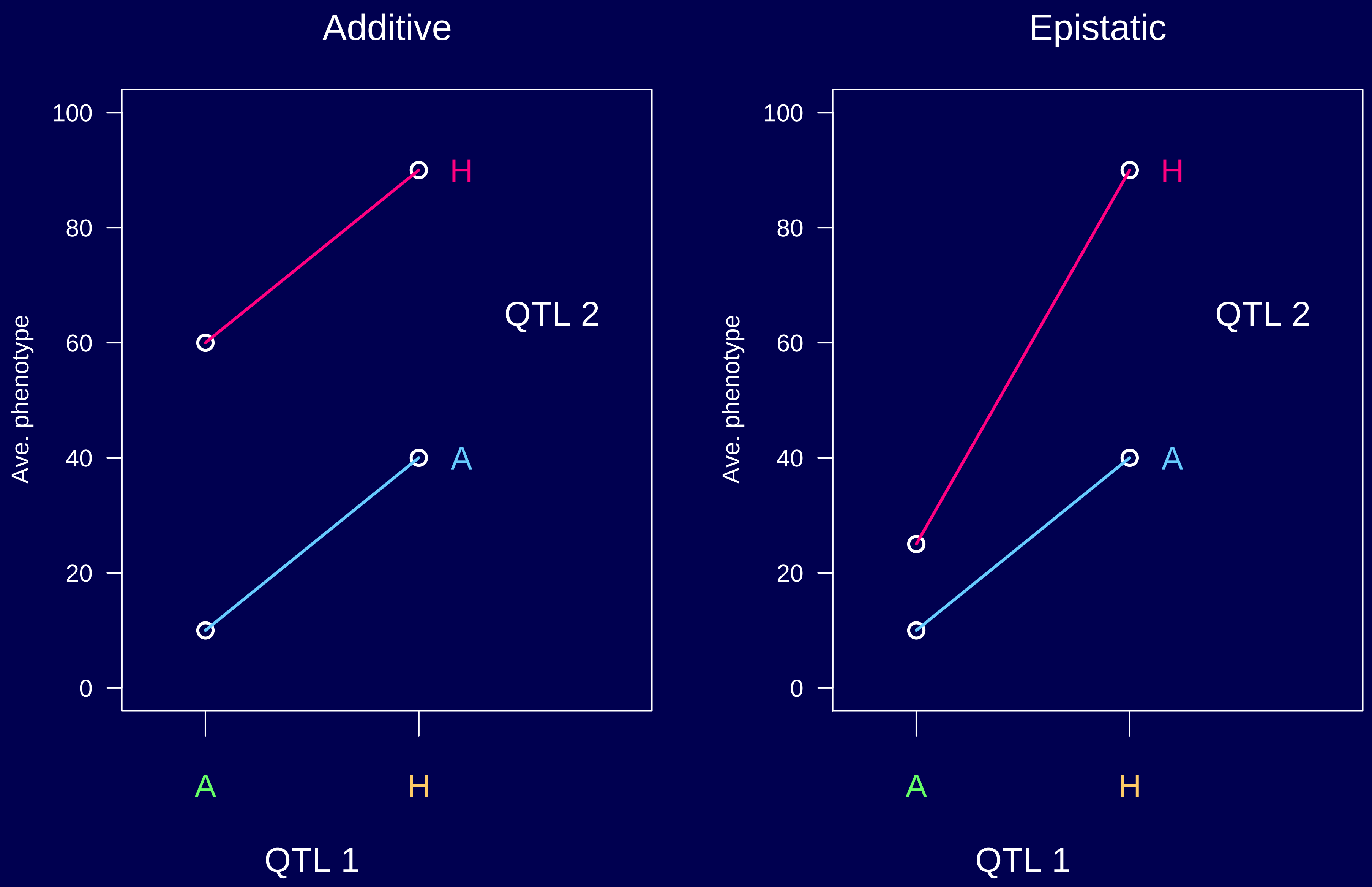
LOD support intervals



Modelling multiple QTL

- Reduce residual variation $\Longrightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL

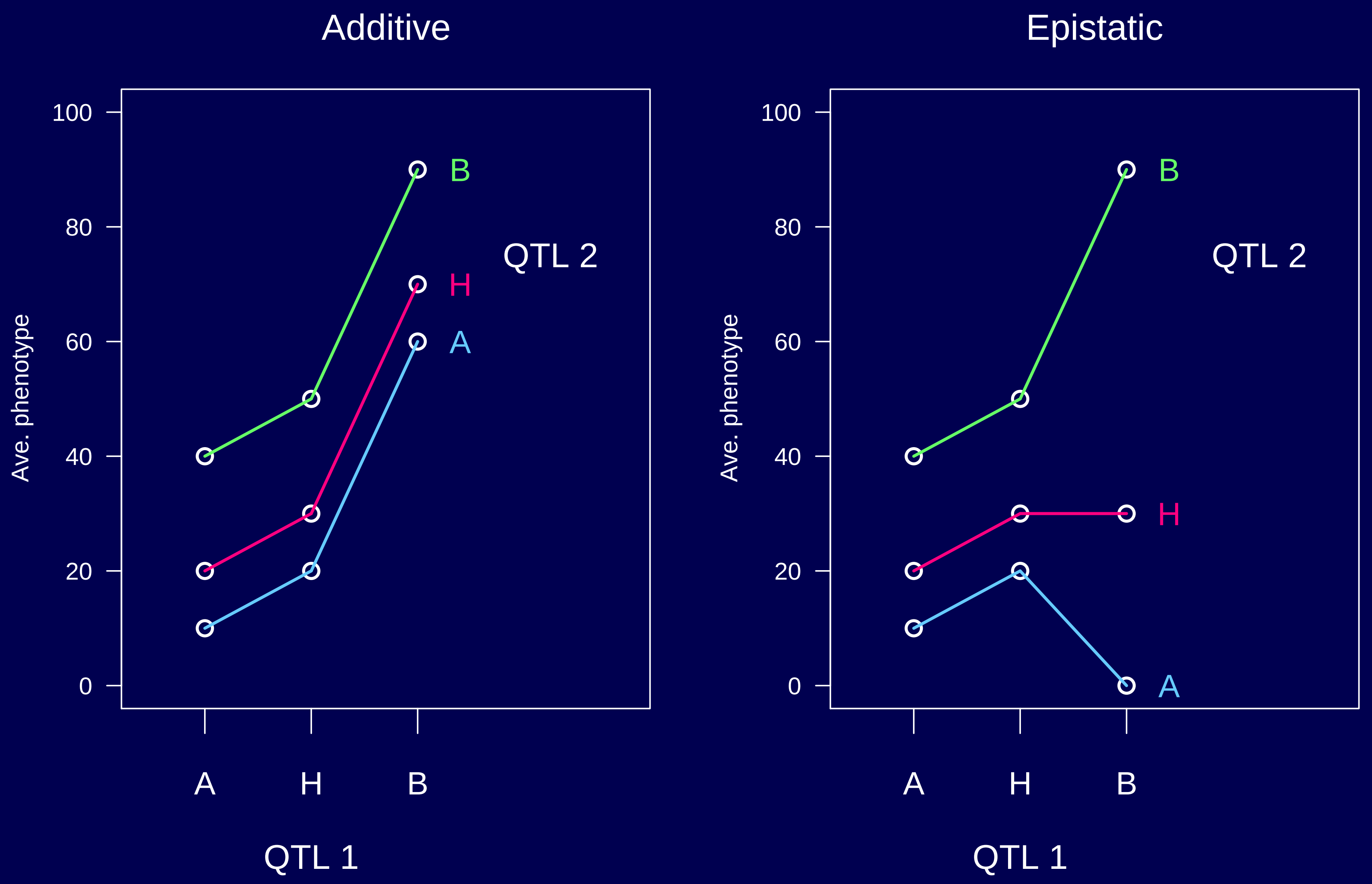
Epistasis in BC



Additive

Epistatic

Epistasis in F_2



Haley-Knott regression

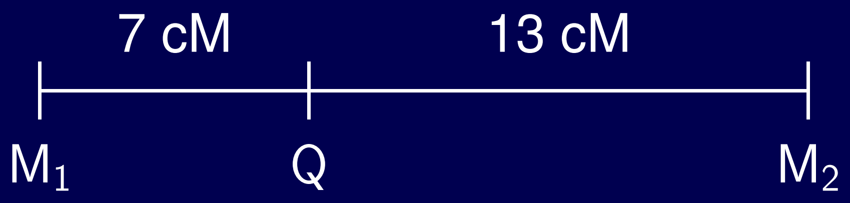
A quick approximation to Interval Mapping.

$$\begin{aligned} \mathsf{E}(\mathsf{y_i}|\mathsf{q_i}) &= \mu_\mathsf{q} \\ \mathsf{E}(\mathsf{y_i}|\mathsf{M_i}) &= \mathsf{E}[\ \mathsf{E}(\mathsf{y_i}|\mathsf{q_i})\ |\mathsf{M_i}] = \textstyle\sum_\mathsf{j} \Pr(\mathsf{q} = \mathsf{j}|\mathsf{M_i})\mu_\mathsf{j} \\ &= \textstyle\sum_\mathsf{j} \mathsf{p_{ij}}\mu_\mathsf{j} \end{aligned}$$

Regress y on p_i, pretending the residual variation is normally distributed (with constant variance).

$$\mathsf{LOD} = \frac{\mathsf{n}}{2}\log_{10}\left(\frac{\mathsf{RSS_0}}{\mathsf{RSS_1}}\right)$$

The normal mixtures



7 cM

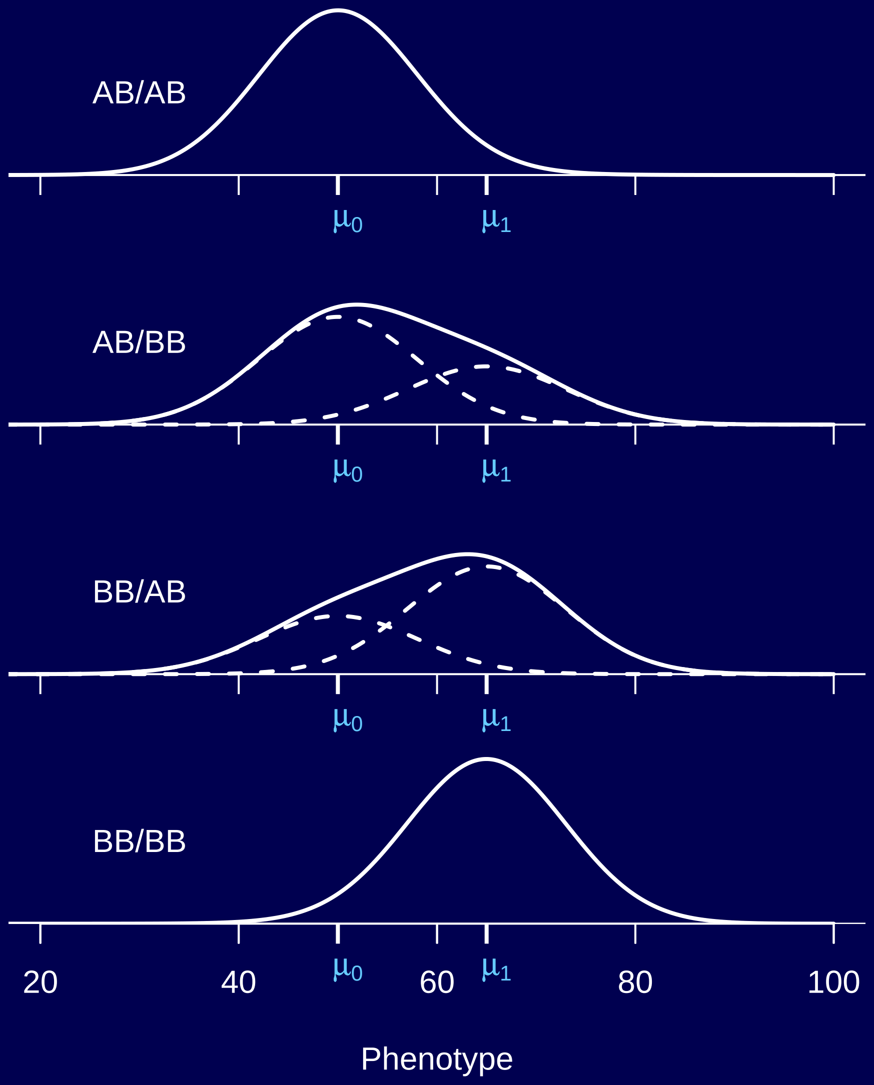
13 cM

M_1

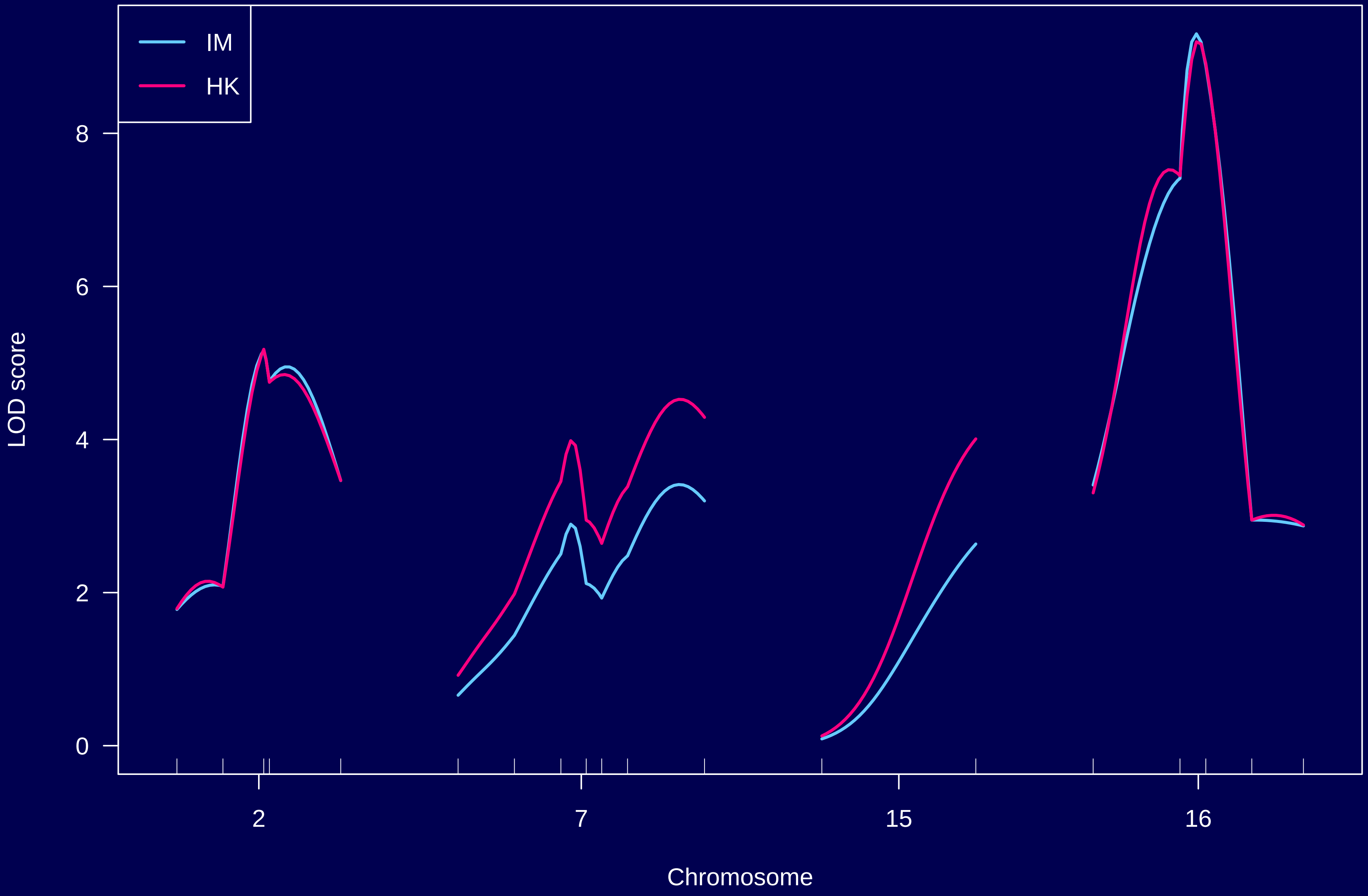
Q

M_2

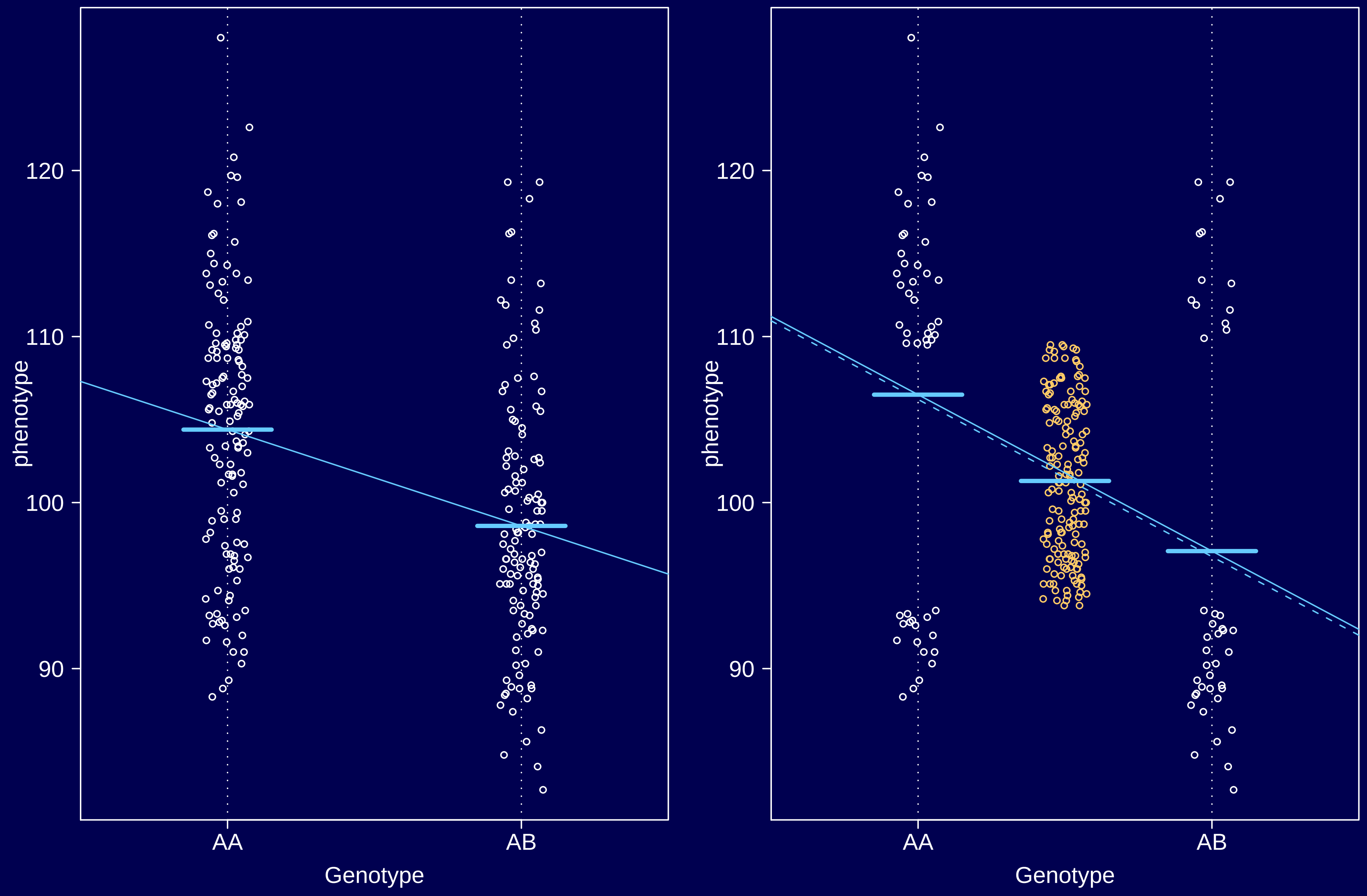
- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right shows the distributions of the phenotype conditional on the genotypes at the two markers.
- The dashed curves correspond to the components of the mixtures.



Haley-Knott results



H-K with selective genotyping



References

- Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab Animal 30:44–52

  A review for non-statisticians.

- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, MA, chapter 15

  Chapter on QTL mapping.

- Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

  The seminal paper.

- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

  LOD thresholds by permutation tests.

- Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

  Haley-Knott regression.

- Strickberger MW (1985) *Genetics*, 3rd edition. Macmillan, New York, chapter 11.

  An old but excellent general genetics textbook with a very interesting discussion of epistasis.