Introduction to QTL mapping in model organisms

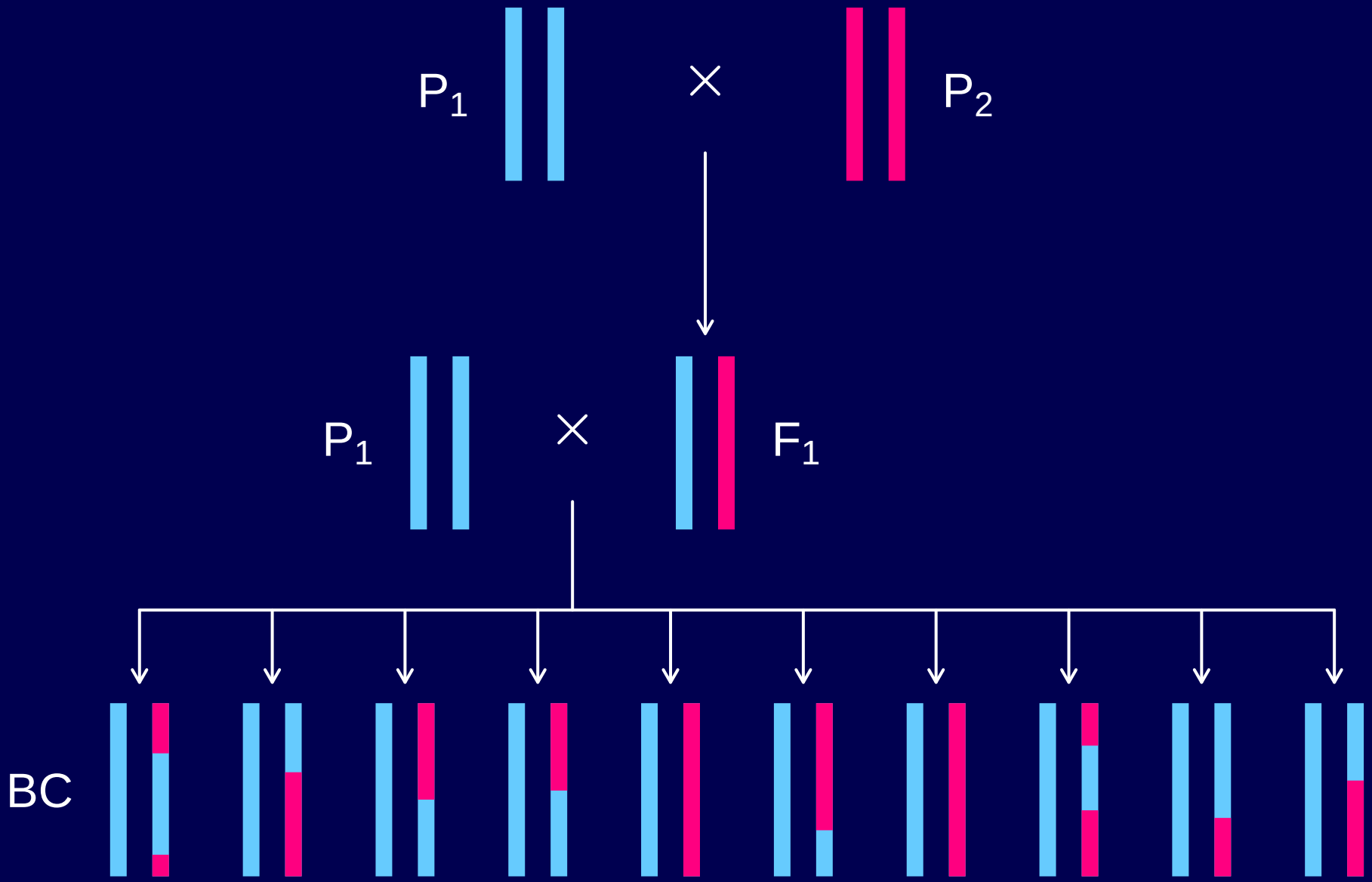
Karl W Broman

Department of Biostatistics and Medical Informatics
University of Wisconsin – Madison

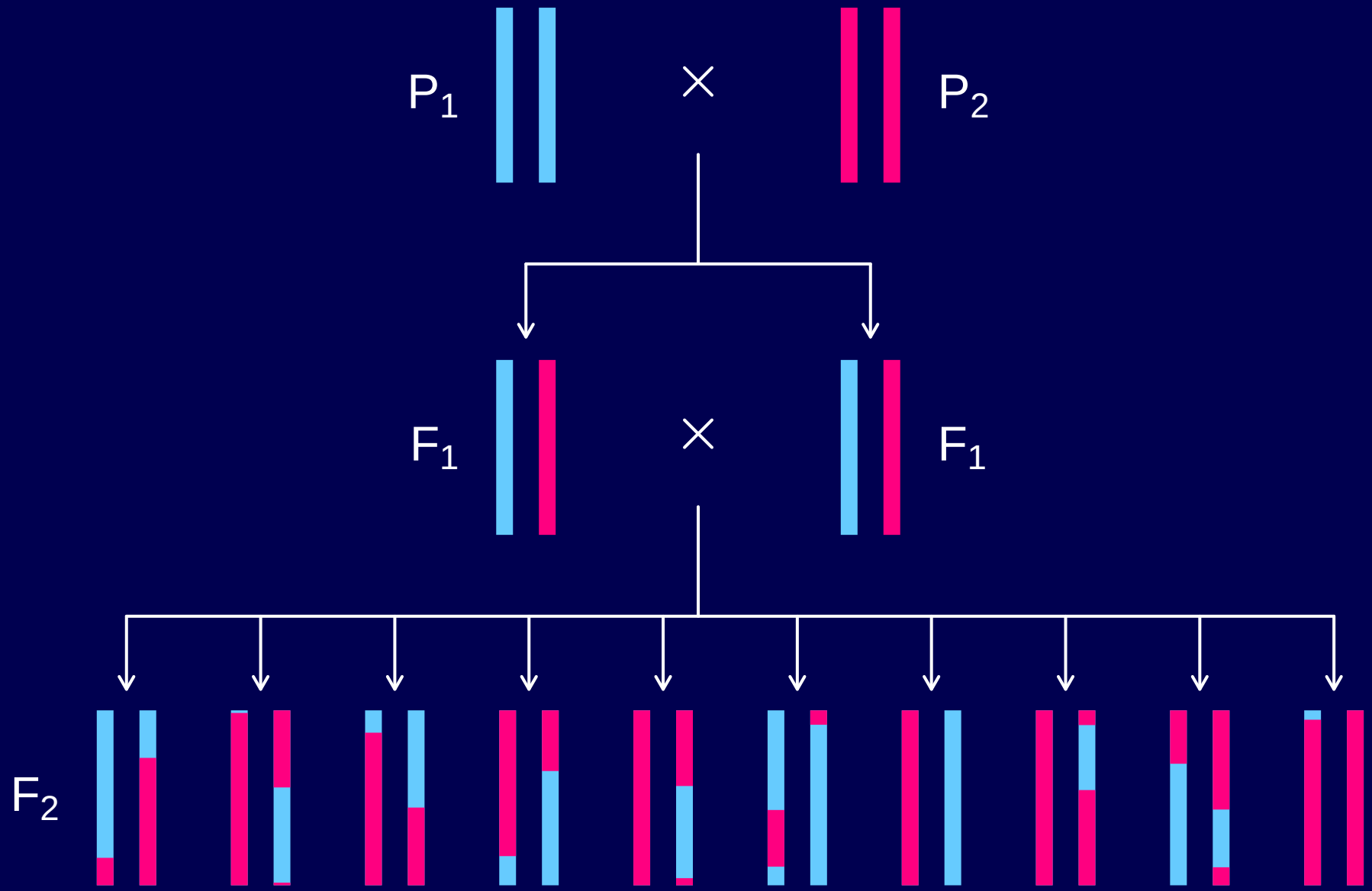
www.biostat.wisc.edu/~kbroman

[→ Teaching → Miscellaneous lectures]

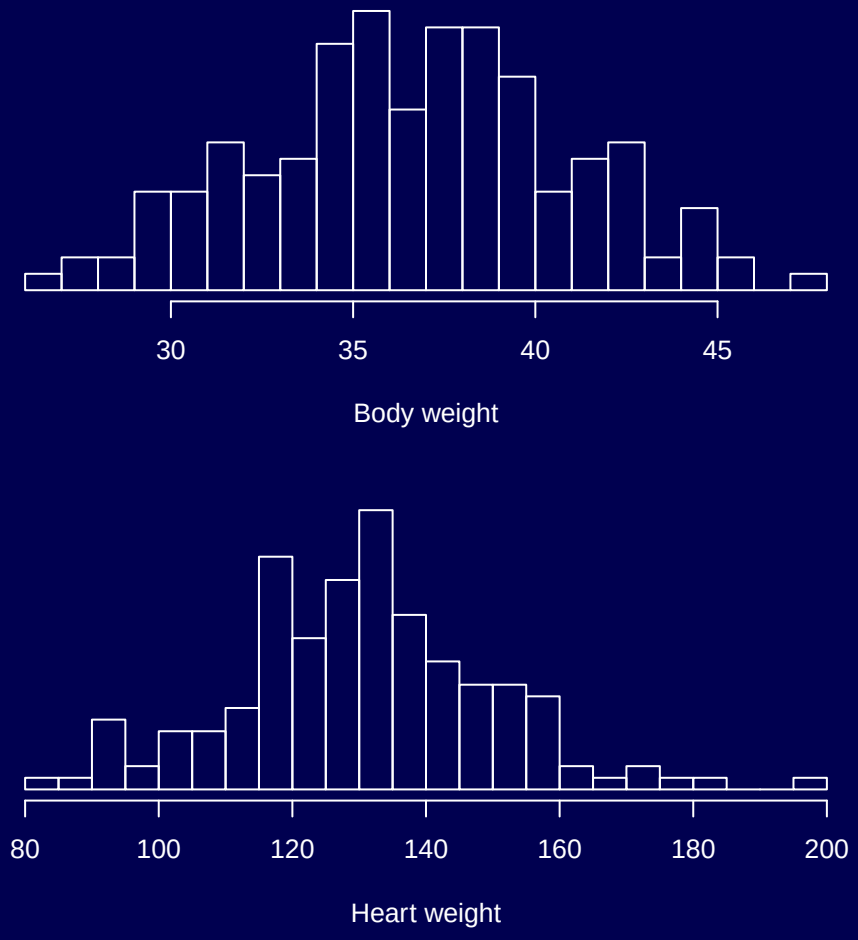
Backcross

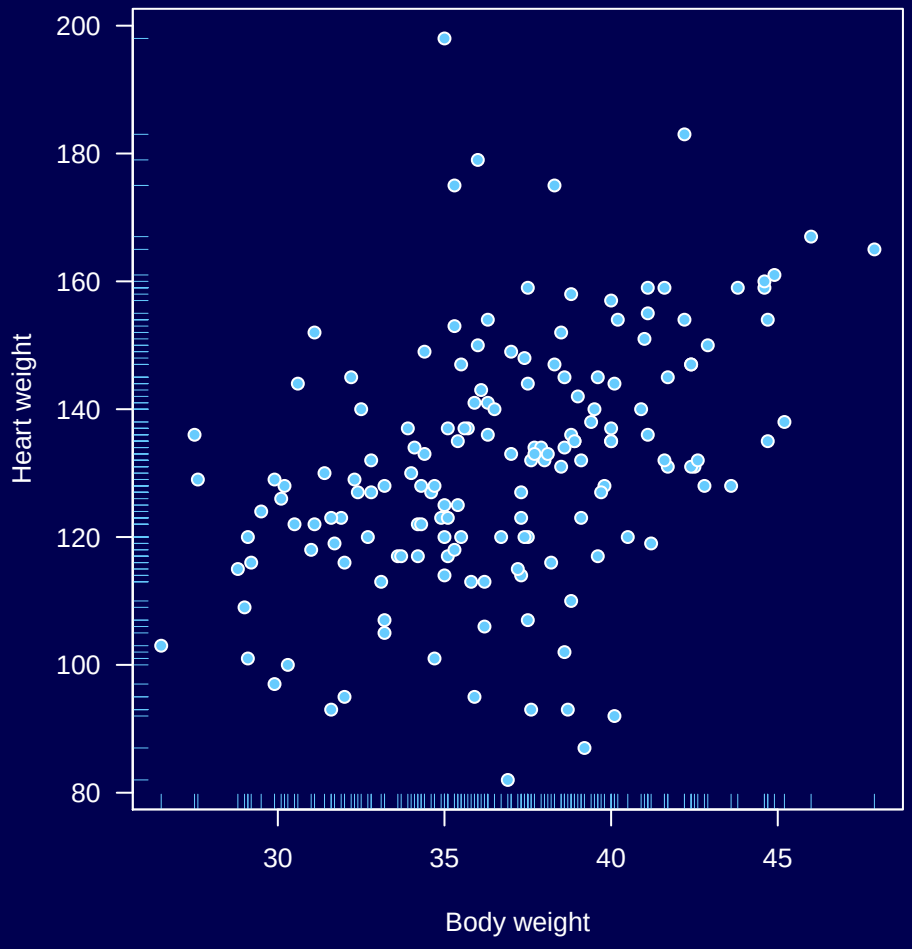


Intercross



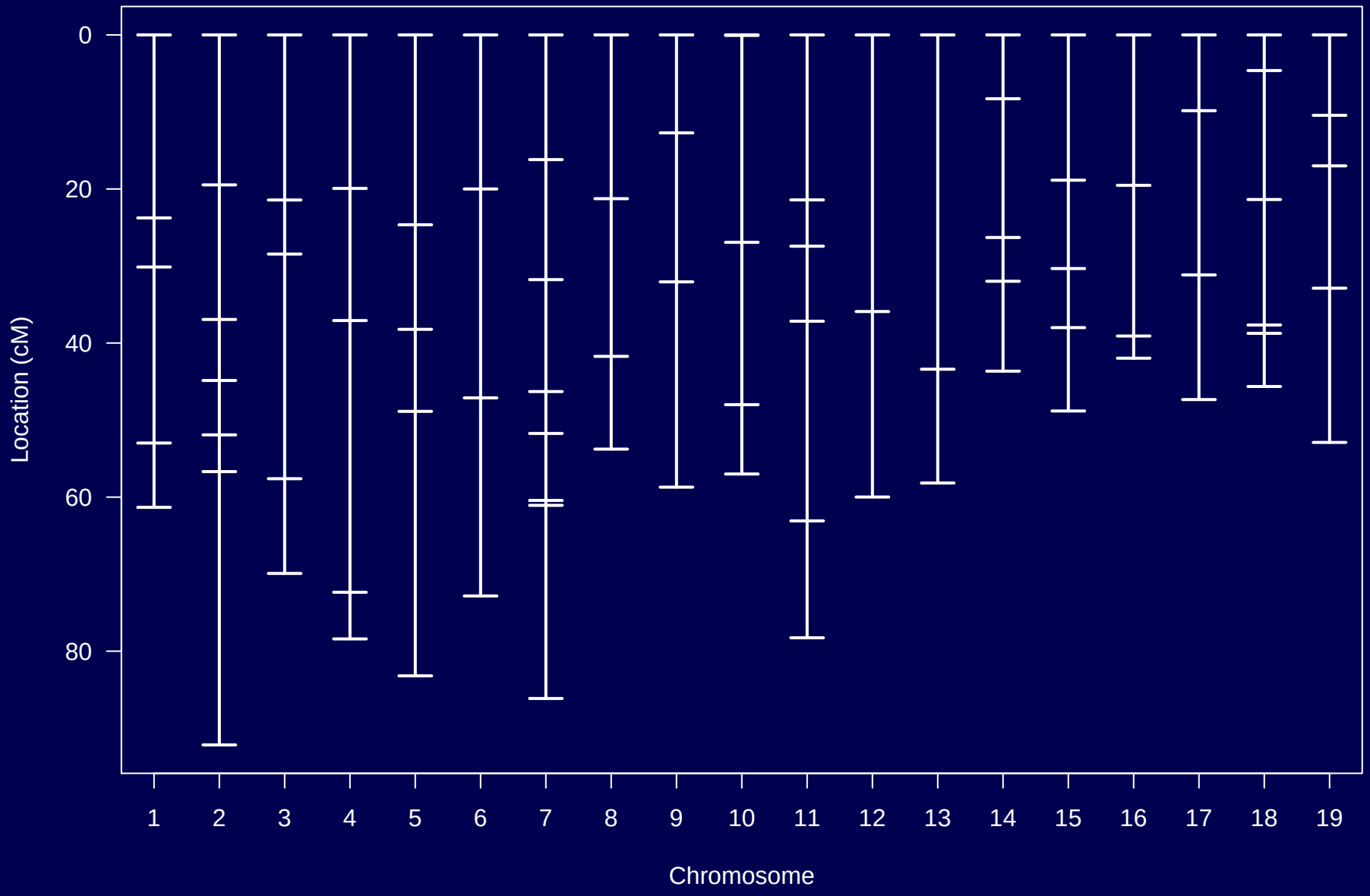
Phenotype data



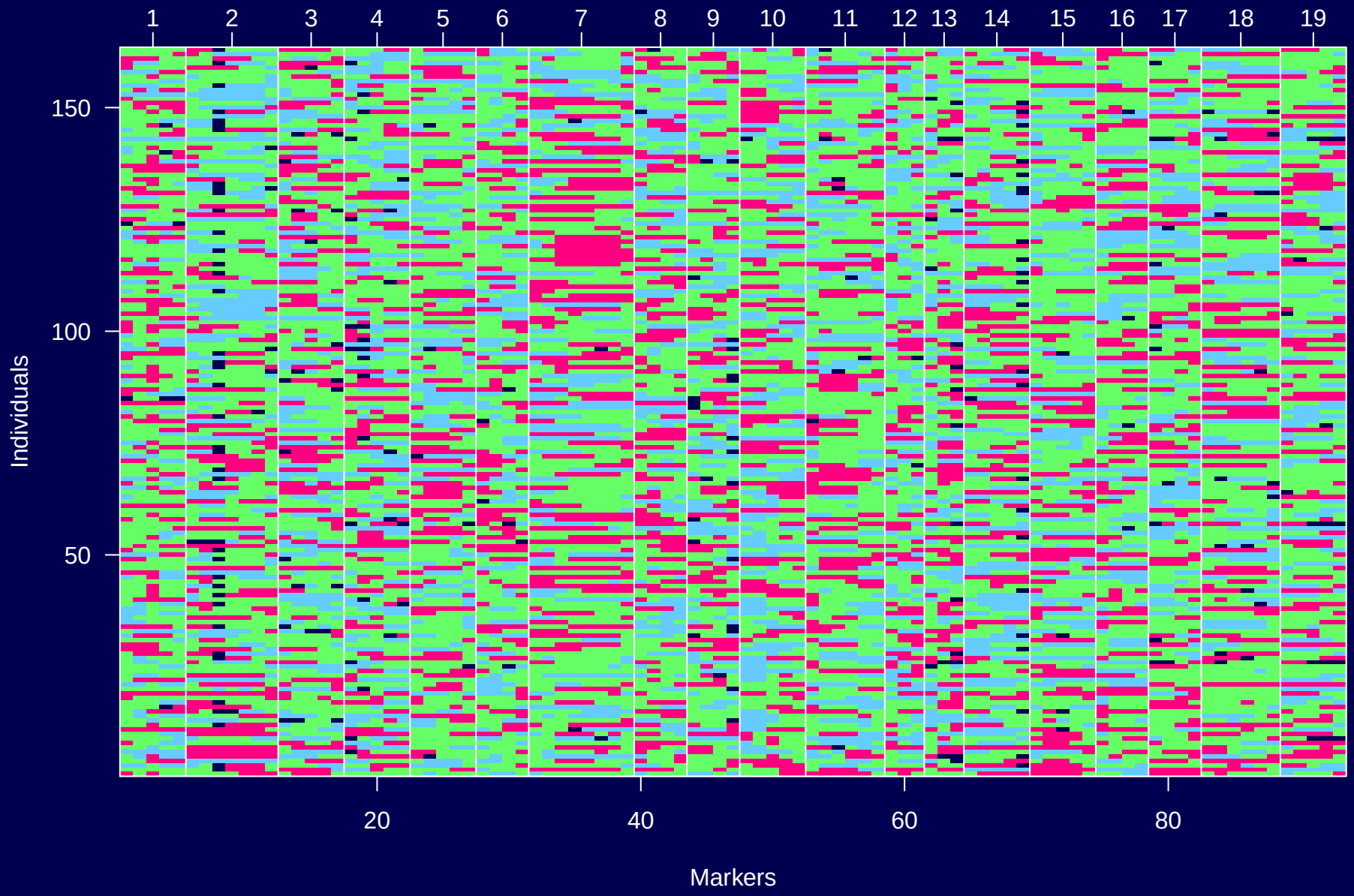


Sugiyama et al. (2002) Physiol Genomics 10:5–12

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

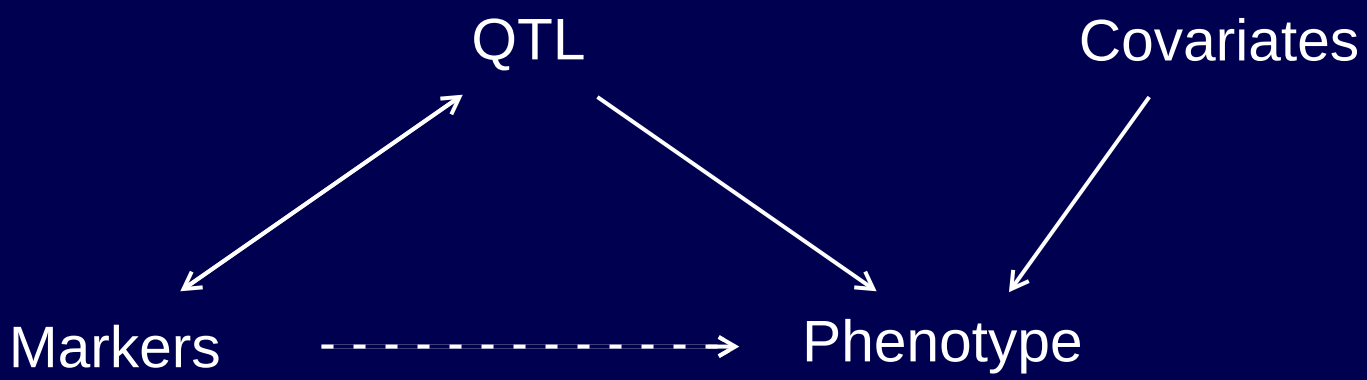
Genotype data



Goals

- Identify quantitative trait loci (QTL)
  (and interactions among QTL)
- Interval estimates of QTL location
- Estimated QTL effects

Statistical structure

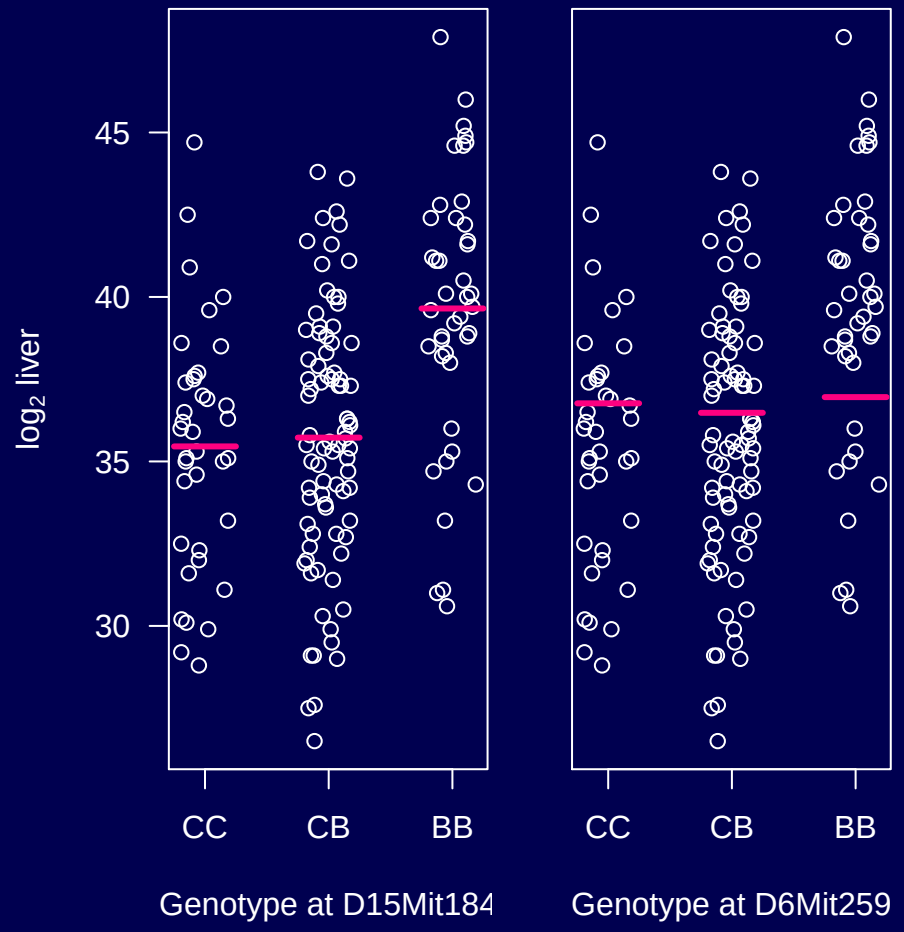


The missing data problem: Markers $\longleftrightarrow$ QTL

The model selection problem: QTL, covariates $\longrightarrow$ 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)

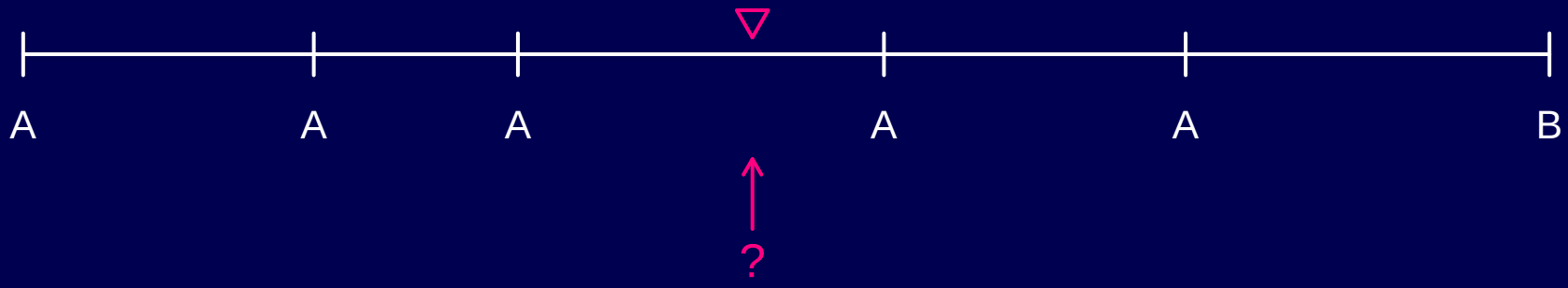
- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let q denote the (unobserved) QTL genotype

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

		QTL genotype	
M_1	M_2	BB	AB
BB	BB	$(1-r_L)(1-r_R)/(1-r)$	$r_L r_R/(1-r)$
BB	AB	$(1-r_L)r_R/r$	$r_L(1-r_R)/r$
AB	BB	$r_L(1-r_R)/r$	$(1-r_L)r_R/r$
AB	AB	$r_L r_R/(1-r)$	$(1-r_L)(1-r_R)/(1-r)$

r = recombination fractions between markers
r_L, r_R = recombination fractions between markers and QTL

Genotype probabilities



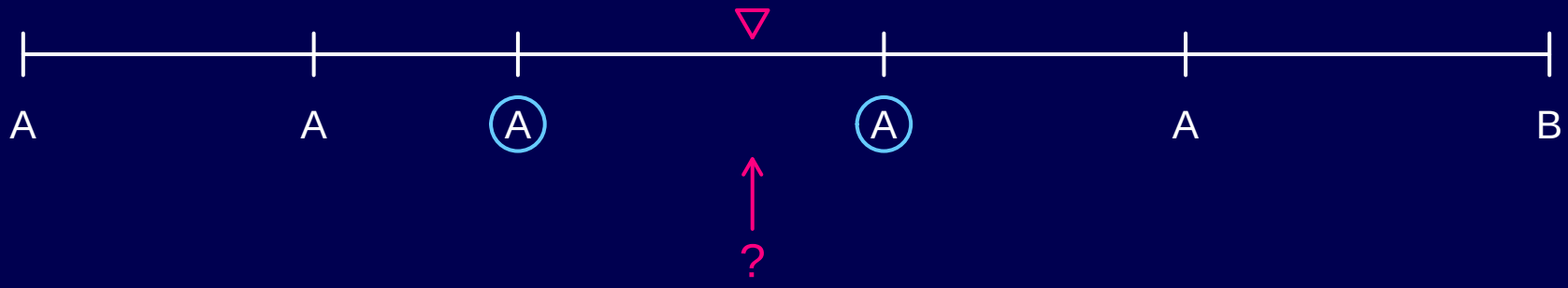
Calculate $\Pr(\mathrm{q} \mid \text{marker data})$, assuming

- 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



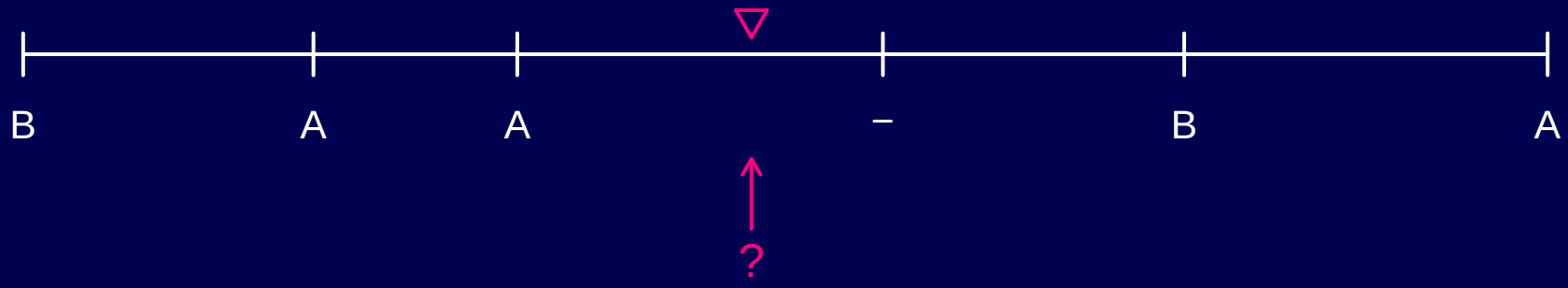
Calculate $\Pr(\mathrm{q} \mid \text{marker data})$, assuming

- 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



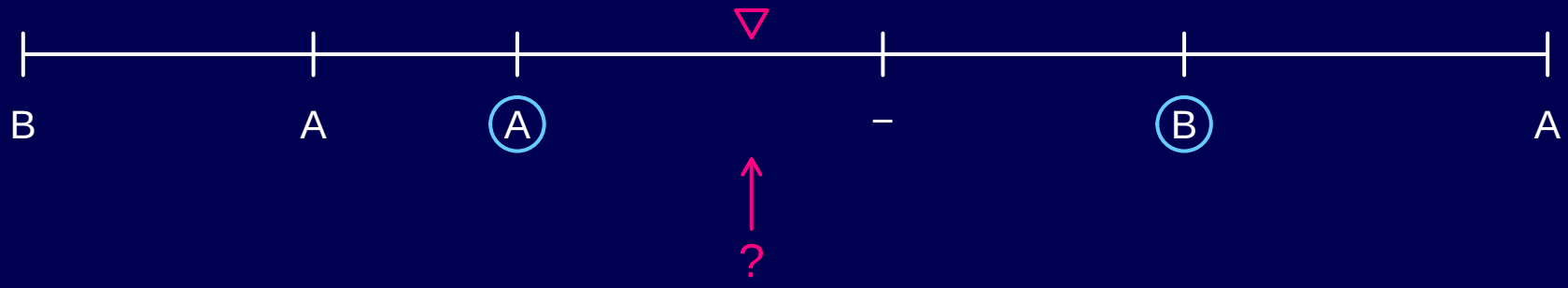
Calculate $\Pr(\mathrm{q} \mid \text{marker data})$, assuming

- 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



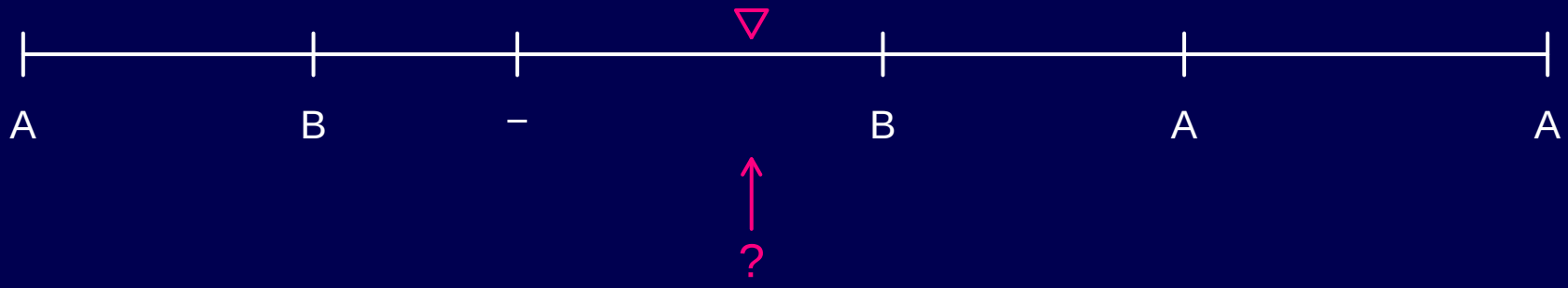
Calculate $\Pr(\mathrm{q} \mid \text{marker data})$, assuming

- 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



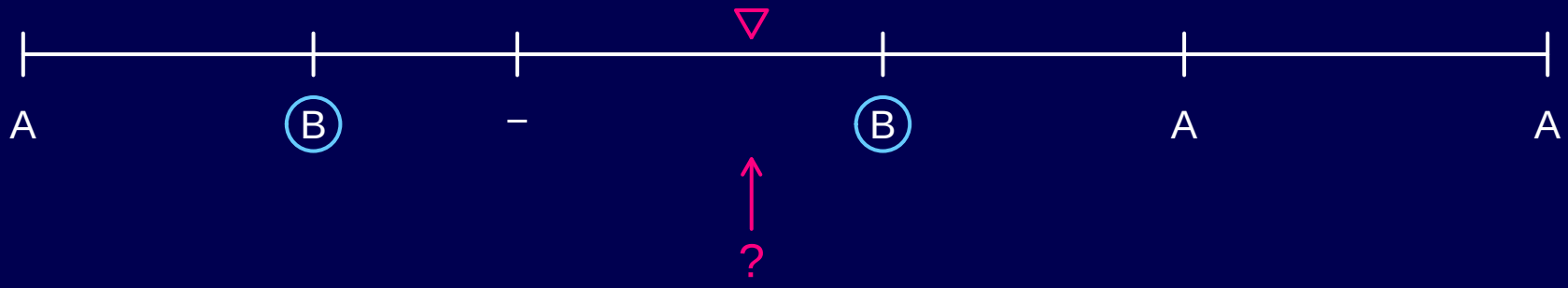
Calculate $\Pr(\mathrm{q} \mid \text{marker data})$, assuming

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

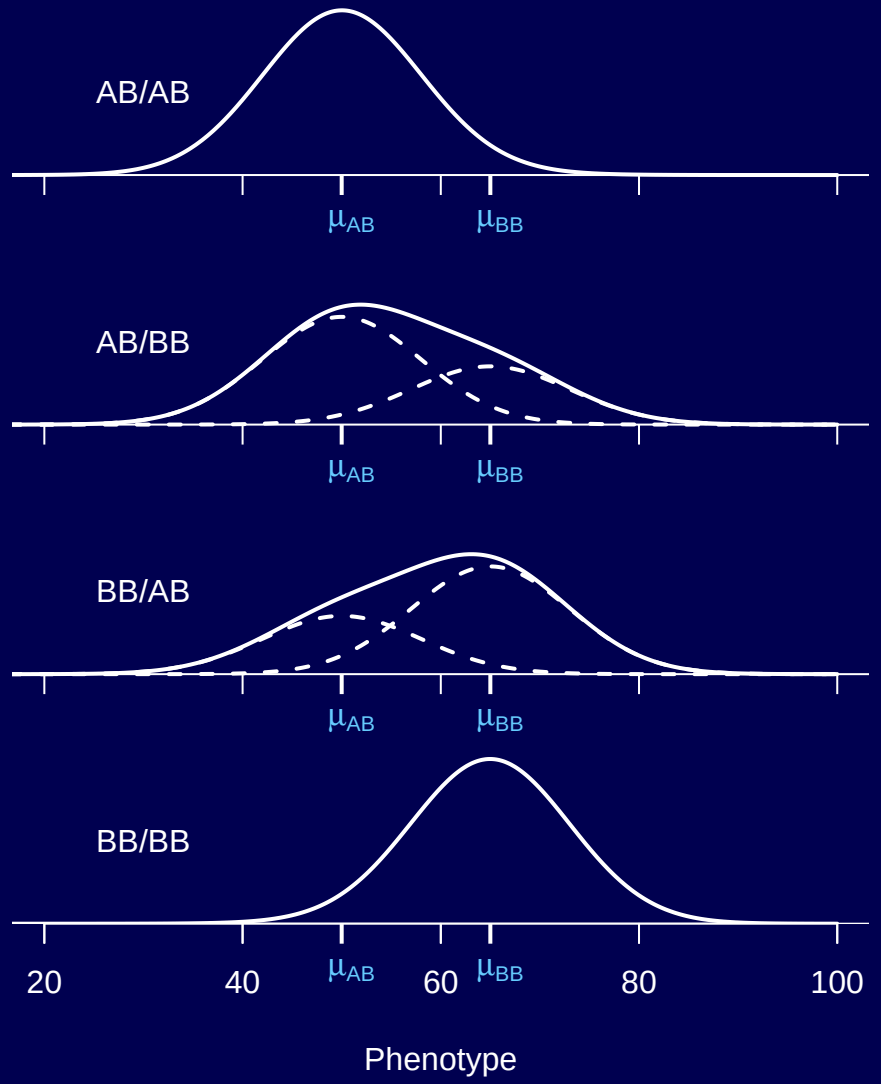
- No crossover interference
- No genotyping errors

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

- 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 $p_{ij} = \Pr(q_i = j|\text{marker data})$

$y_i|q_i \sim N(\mu_{q_i}, \sigma^2)$

$\Pr(y_i|\text{marker data}, \boldsymbol{\mu}, \sigma) = \sum_j p_{ij}\, f(y_i; \mu_j, \sigma)$

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

Log likelihood: $l(\boldsymbol{\mu}, \sigma) = \sum_i \log \Pr(y_i|\text{marker data}, \boldsymbol{\mu}, \sigma)$

Maximum likelihood estimates (MLEs) of $\boldsymbol{\mu}$, σ:

values for which $l(\boldsymbol{\mu}, \sigma)$ is maximized.

EM algorithm

Dempster et al. (1977)

E step:

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

$$= \frac{p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}{\sum_j p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}$$

M step:

Let $\hat{\mu}_j^{(k)} = \sum_i y_i w_{ij}^{(k)} / \sum_i w_{ij}^{(k)}$

$$\hat{\sigma}^{(k)} = \sqrt{\sum_i \sum_j w_{ij}^{(k)} (y_i - \hat{\mu}_j^{(k)})^2 / n}$$

The algorithm:

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

LOD scores

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

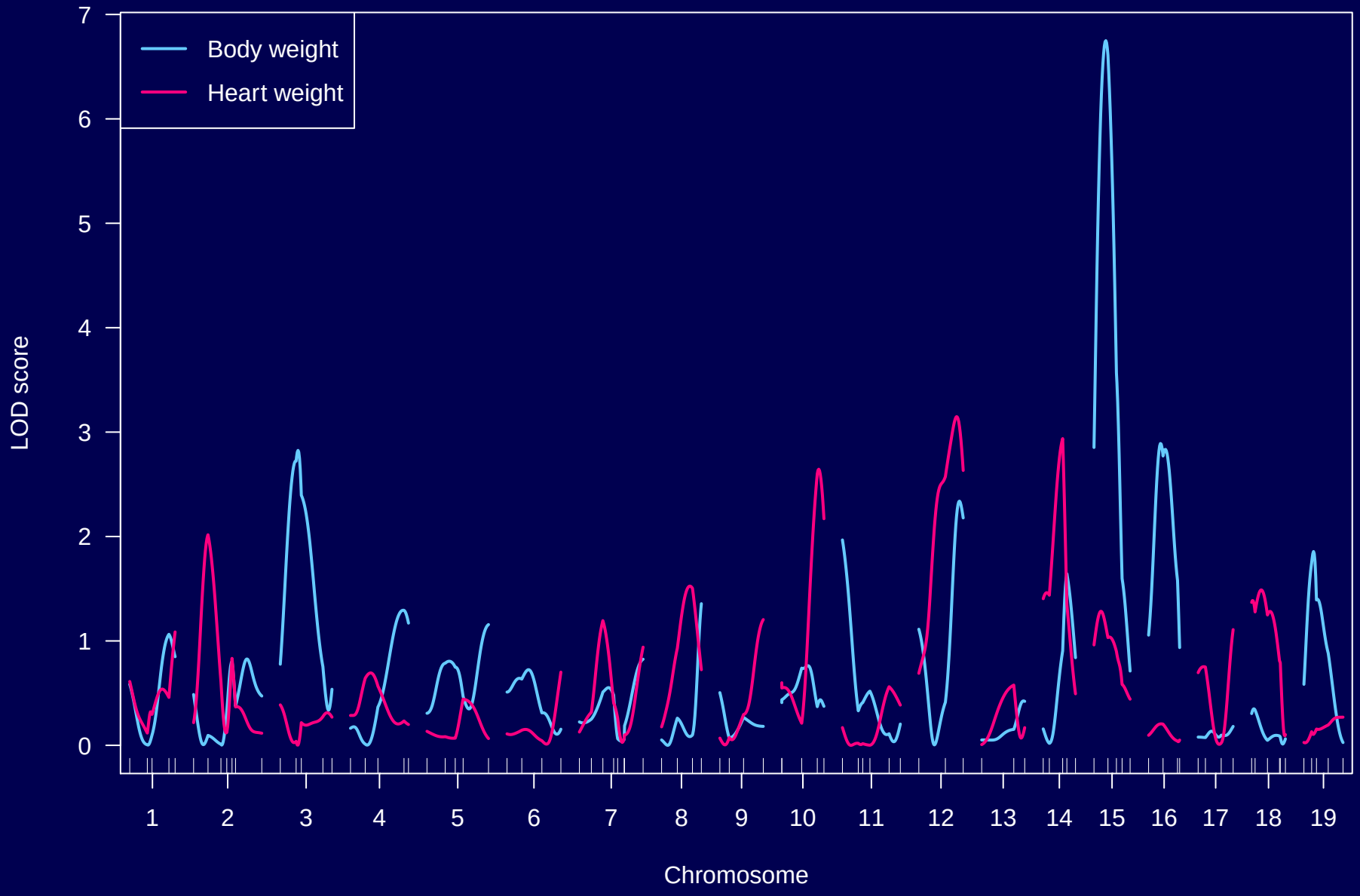
LOD(λ) $= \log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position λ versus that of no QTL

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

$\hat{\boldsymbol{\mu}}_\lambda, \hat{\sigma}_\lambda$ 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

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?

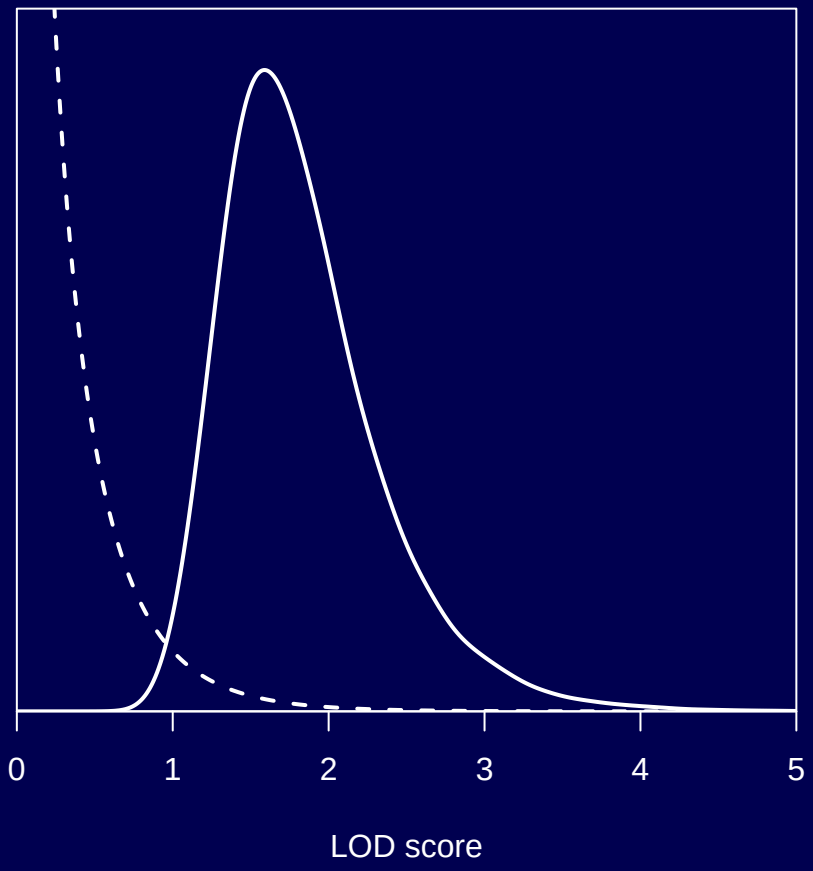
LOD threshold = 95 %ile of distr'n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:

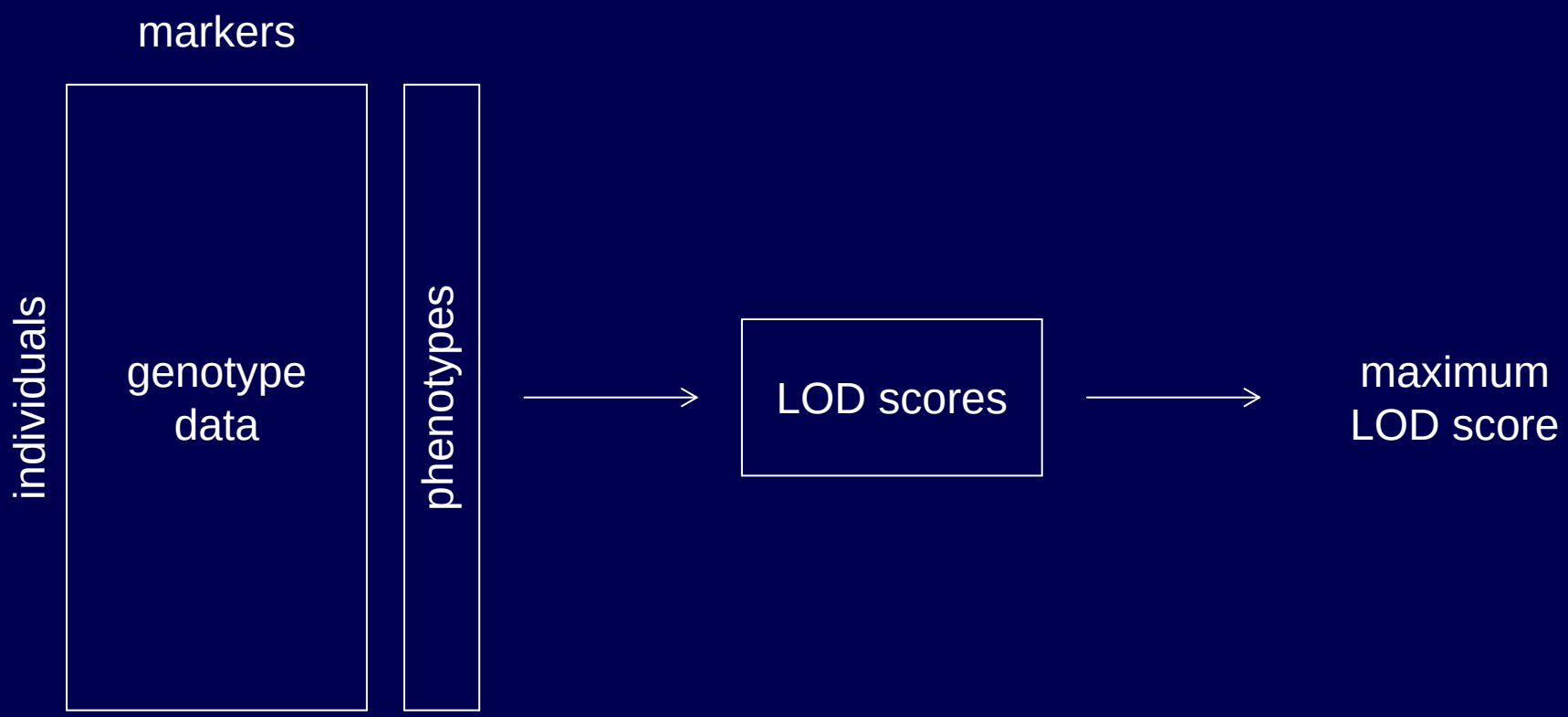
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)

Null distribution of the LOD score

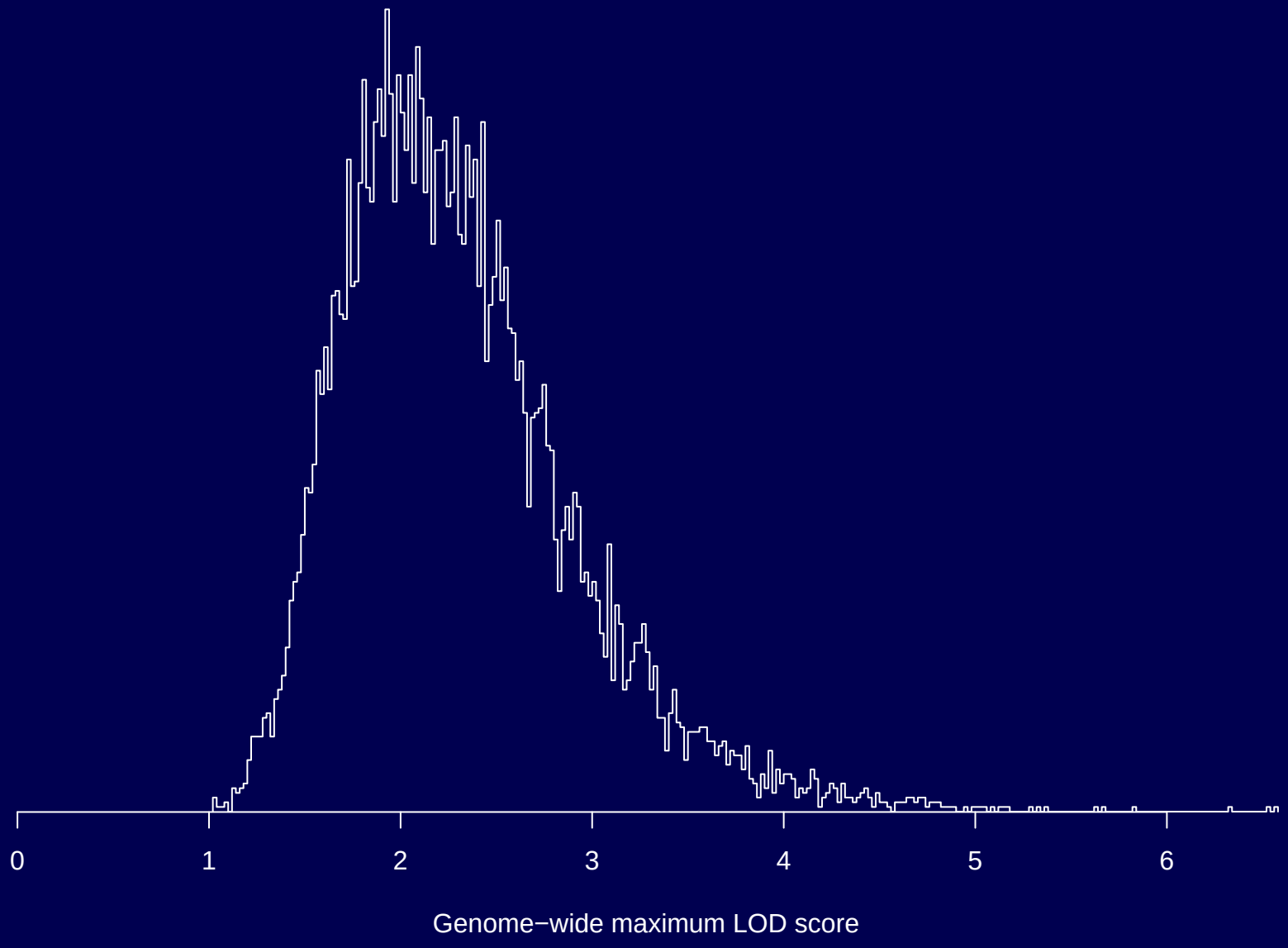
- Null distribution derived by computer simulation of backcross with genome of typical size.
- Dashed curve: distribution of LOD score at any one point.
- Solid curve: distribution of maximum LOD score, genome-wide.



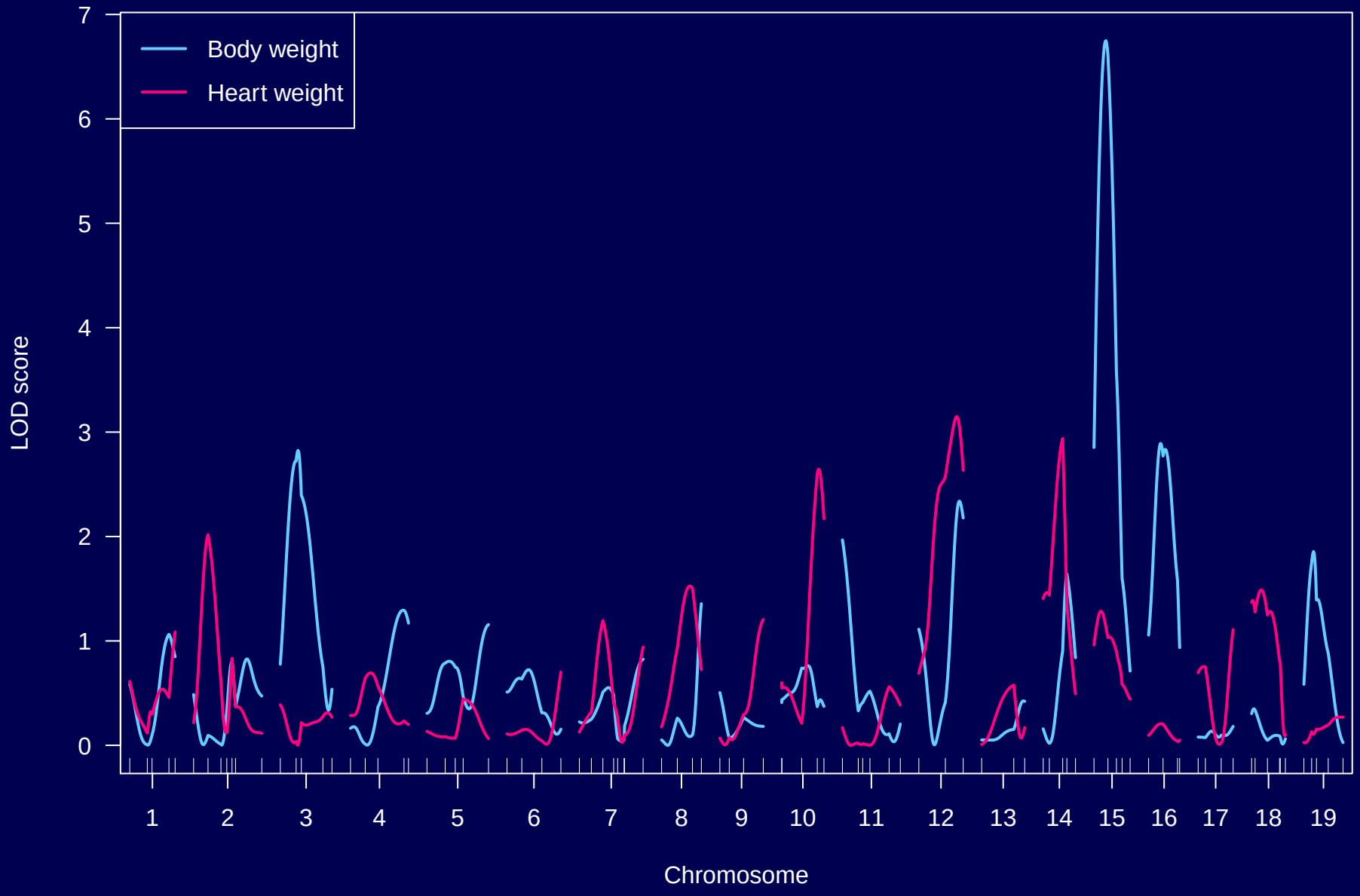
Permutation test



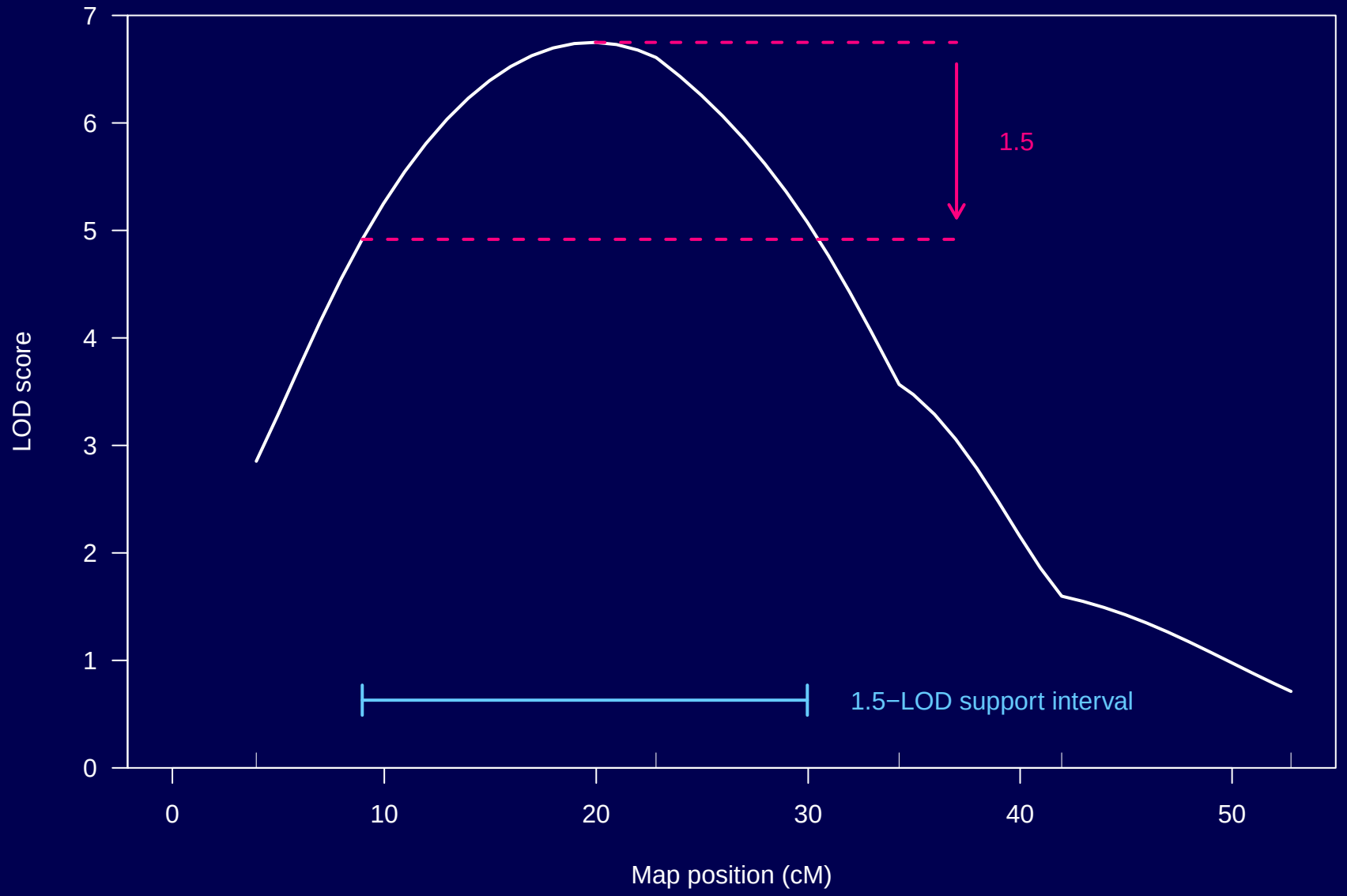
Permutation results



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

LOD support intervals



Haley-Knott regression

A quick approximation to Interval Mapping.

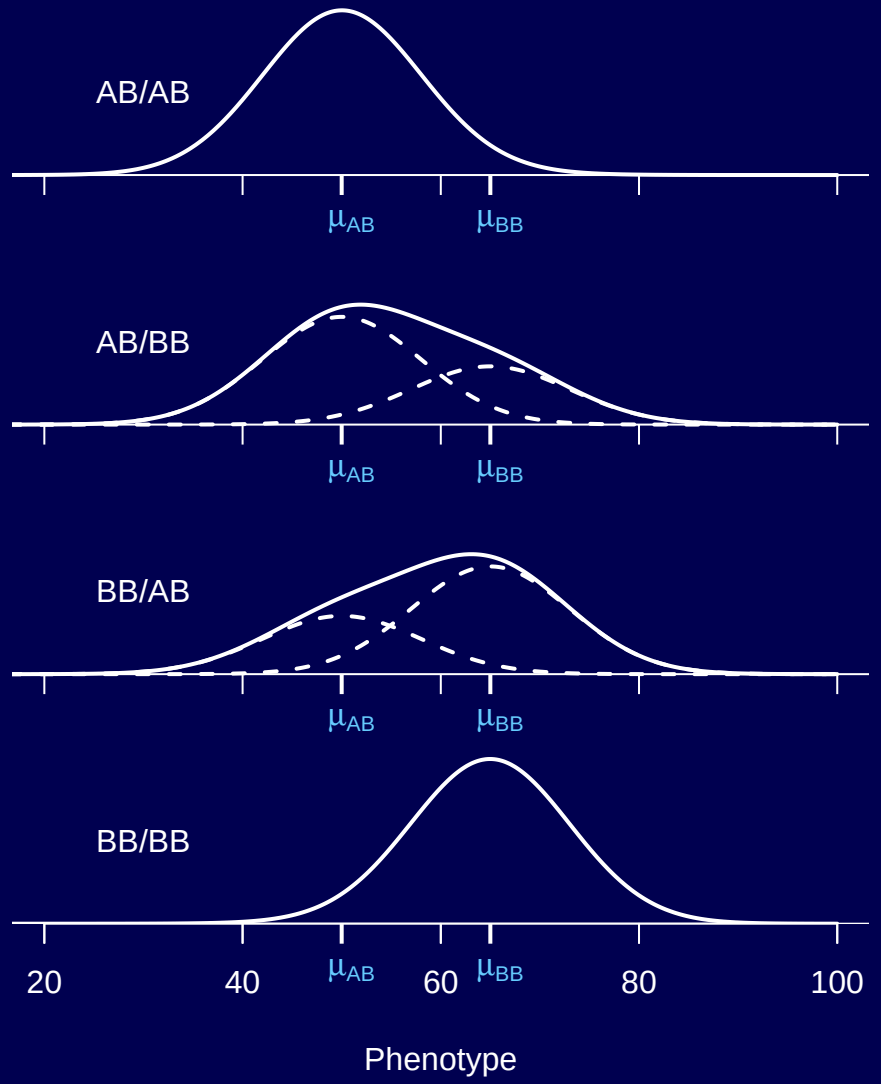
$$
\begin{aligned}
E(y_i|q_i) &= \mu_q \\
E(y_i|M_i) &= E[\, E(y_i|q_i) \,|M_i] \\
&= \textstyle\sum_j \Pr(q = j|M_i)\mu_j \\
&= \textstyle\sum_j p_{ij}\mu_j
\end{aligned}
$$

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

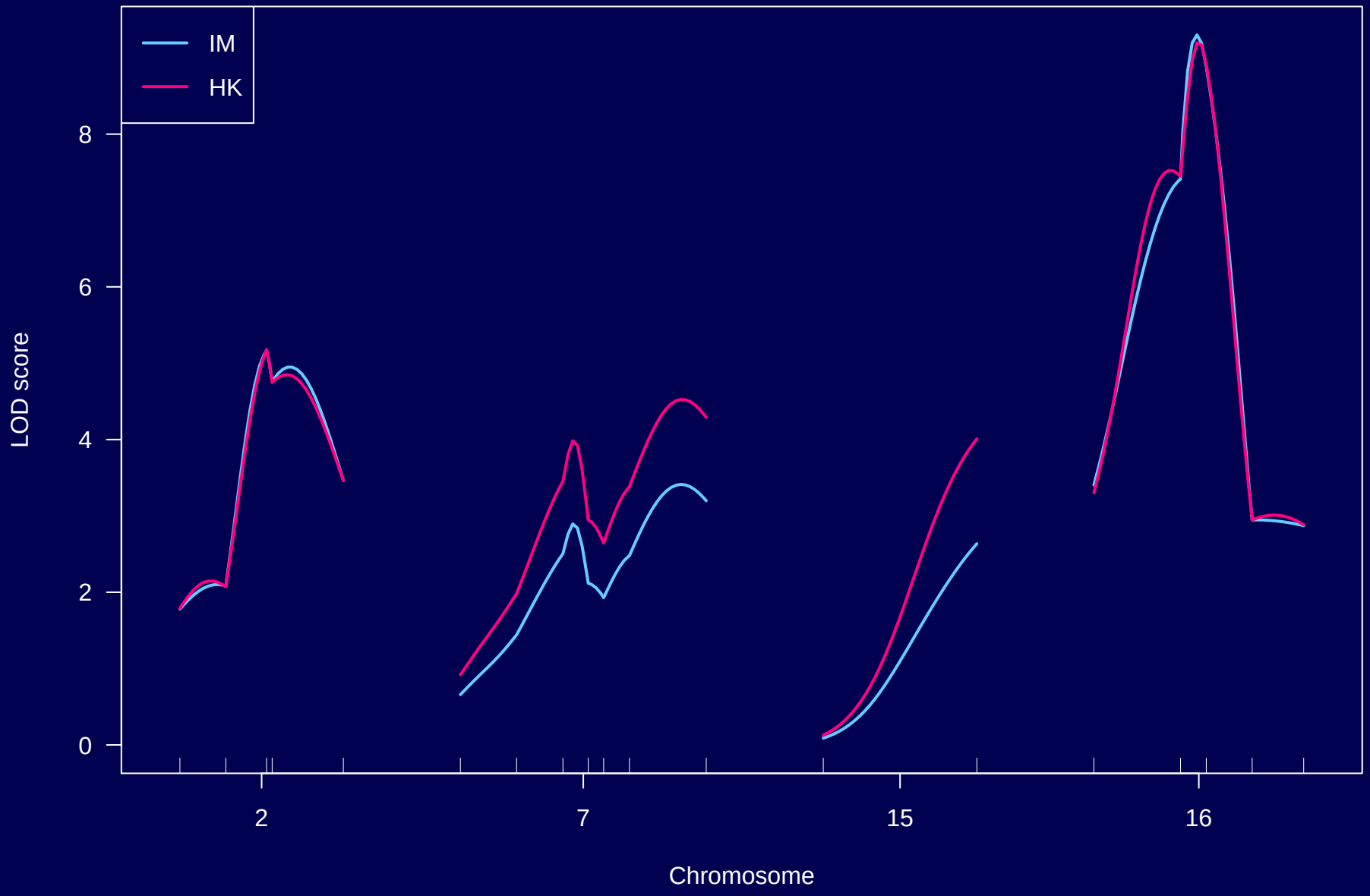
$$
\text{LOD} = \left(\frac{n}{2}\right) \log_{10}\left(\frac{RSS_0}{RSS_1}\right)
$$

The normal mixtures

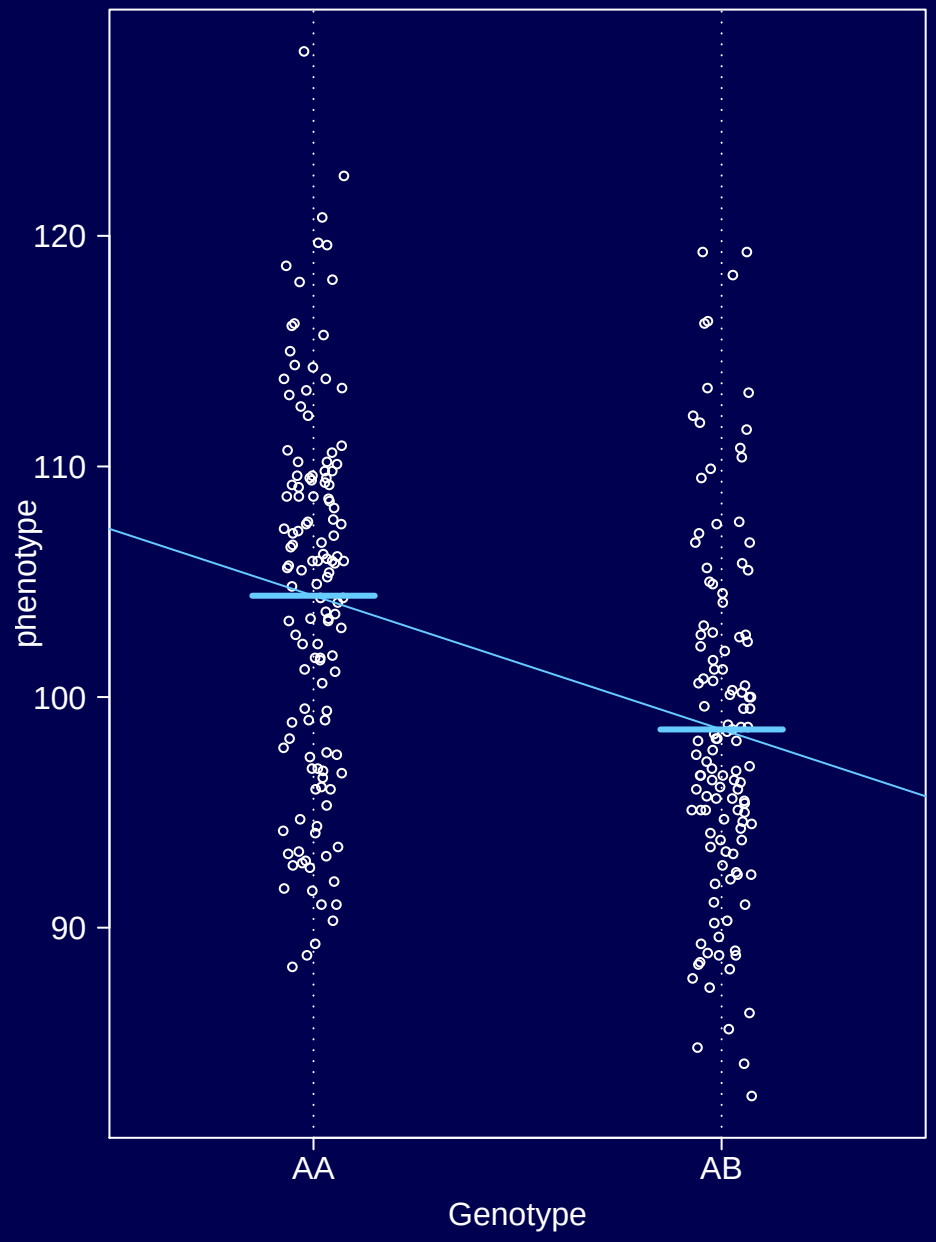
- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right shows the distributions of the phenotype conditional on the genotypes at the two markers.
- The dashed curves correspond to the components of the mixtures.

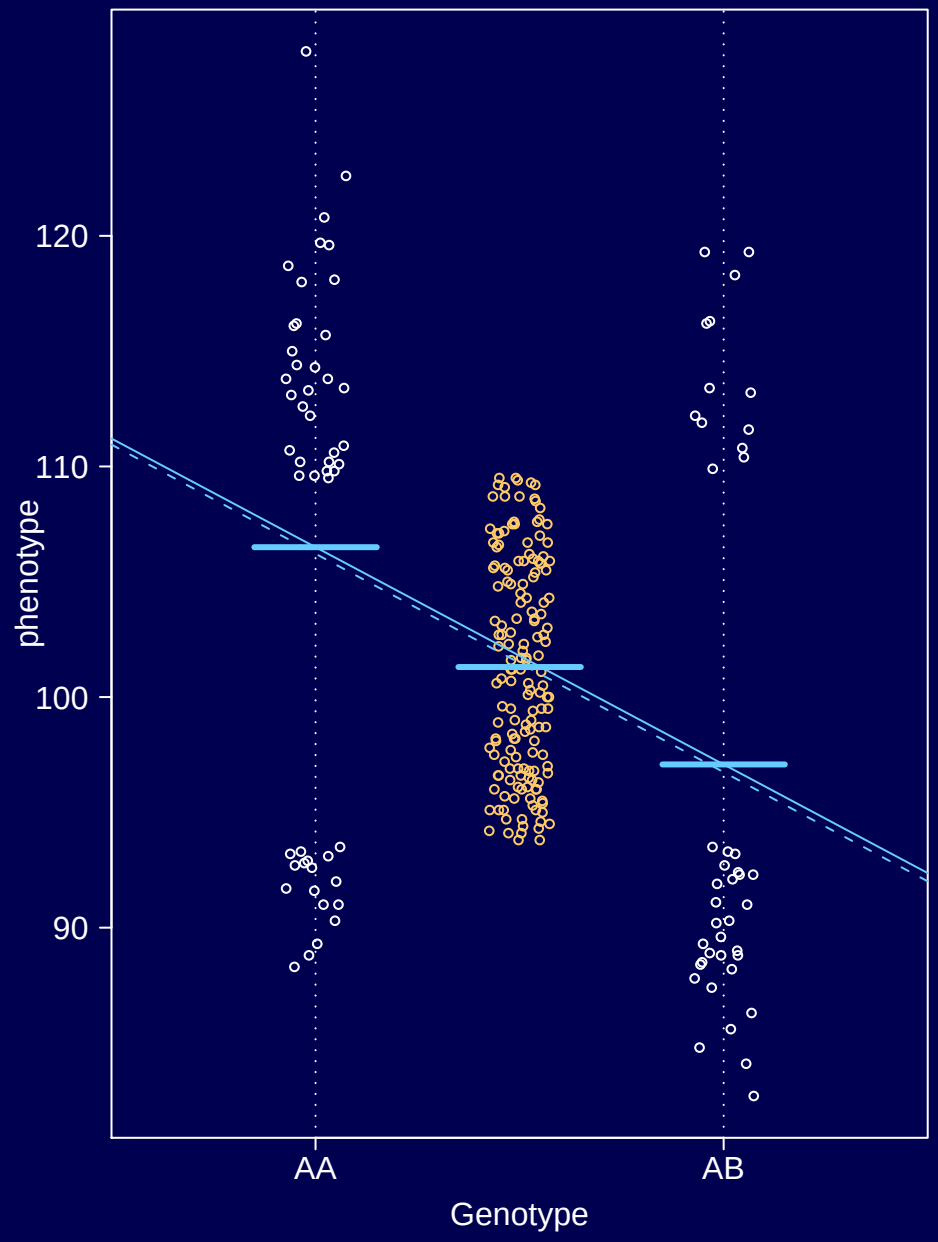


Haley-Knott results



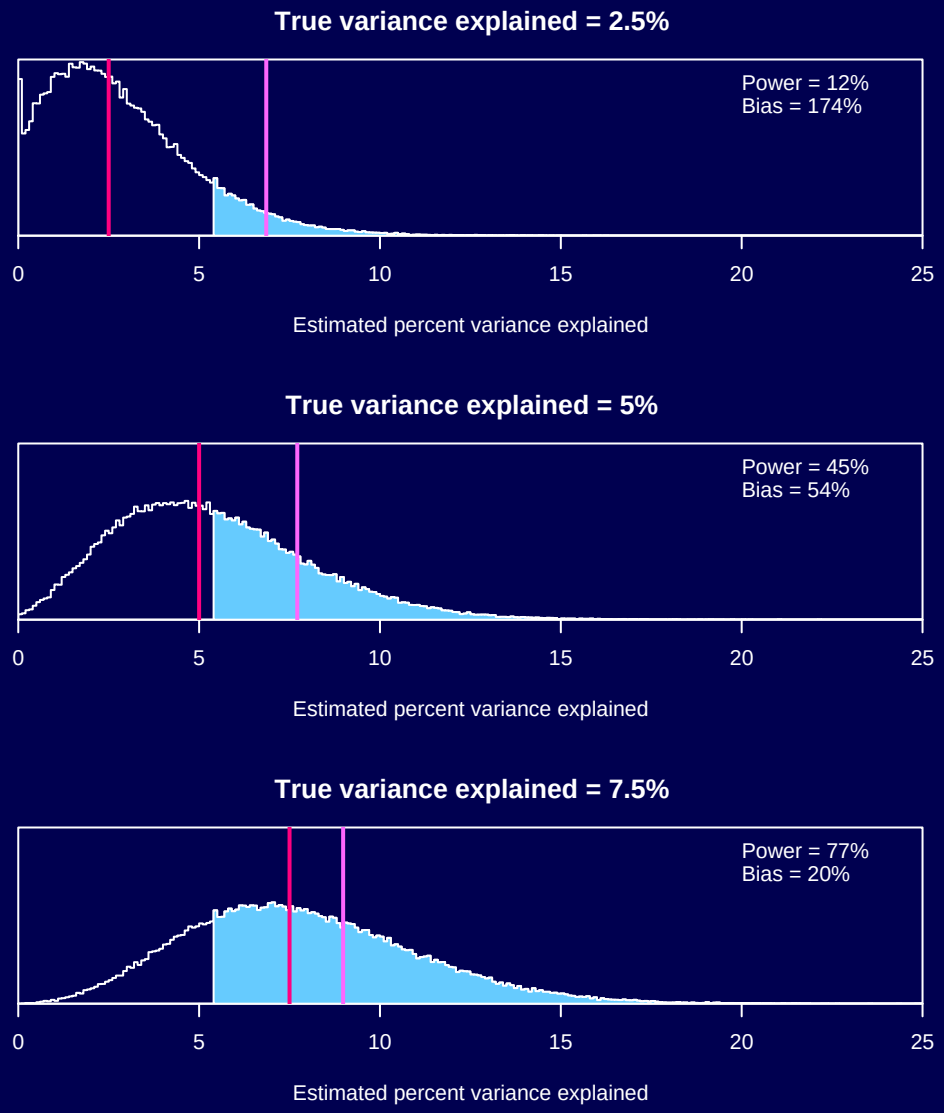
H-K with selective genotyping





Selection bias

- The estimated effect of a QTL will vary somewhat from its true effect.
- Only when the estimated effect is large will the QTL be detected.
- Among those experiments in which the QTL is detected, the estimated QTL effect will be, on average, larger than its true effect.
- This is selection bias.
- Selection bias is largest in QTLs with small or moderate effects.
- The true effects of QTLs that we identify are likely smaller than was observed.



Implications

- Estimated % variance explained by identified QTLs
- Repeating an experiment
- Congenics
- Marker-assisted selection

Non-normal traits

- Standard interval mapping assumes normally distributed residual variation. (Thus the phenotype distribution is a mixture of normals.)
- In reality: we see dichotomous traits, counts, skewed distributions, outliers, and all sorts of odd things.
- Interval mapping, with LOD thresholds derived from permutation tests, generally performs just fine anyway.
- Alternatives to consider:
  - Nonparametric approaches (Kruglyak & Lander 1995)
  - Transformations (*e.g.*, log, square root, normal quantiles)
  - Specially-tailored models (*e.g.*, a generalized linear model, the Cox proportional hazard model, and the model in Broman et al. 2000)

Data diagnostics

- Plot phenotypes
- Look for sample duplicates
- Look for excessive missing data
- Investigate segregation distortion
- Verify genetic maps/marker positions
- Look for genotyping errors
- Look at counts of crossovers

Summary

- Marker regression
  - do t-test or ANOVA at each marker
- Interval mapping
  - deals with missing genotypes at putative QTL
- LOD scores
  - measure of evidence for a QTL
- Permutation-based significance thresholds
  - to account for genome scan
- LOD support intervals
  - approximate confidence interval for QTL location

Summary

- Haley-Knott regression
  - quick approximation to interval mapping
- Selection bias
  - Estimated QTL effects generally biased
- Non-normal traits
  - Consider transformations
- Data diagnostics
  - critical component of QTL analysis