

Introduction to QTL mapping

Karl Broman

Biostatistics and Medical Informatics
University of Wisconsin – Madison

rqtl.org

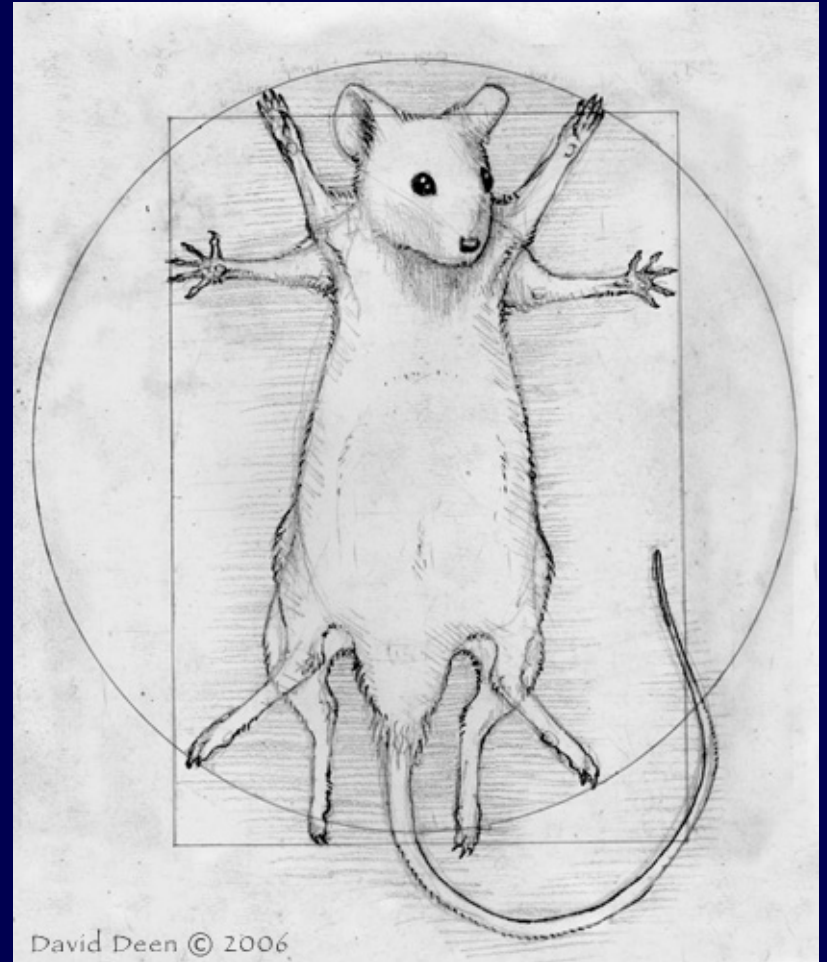
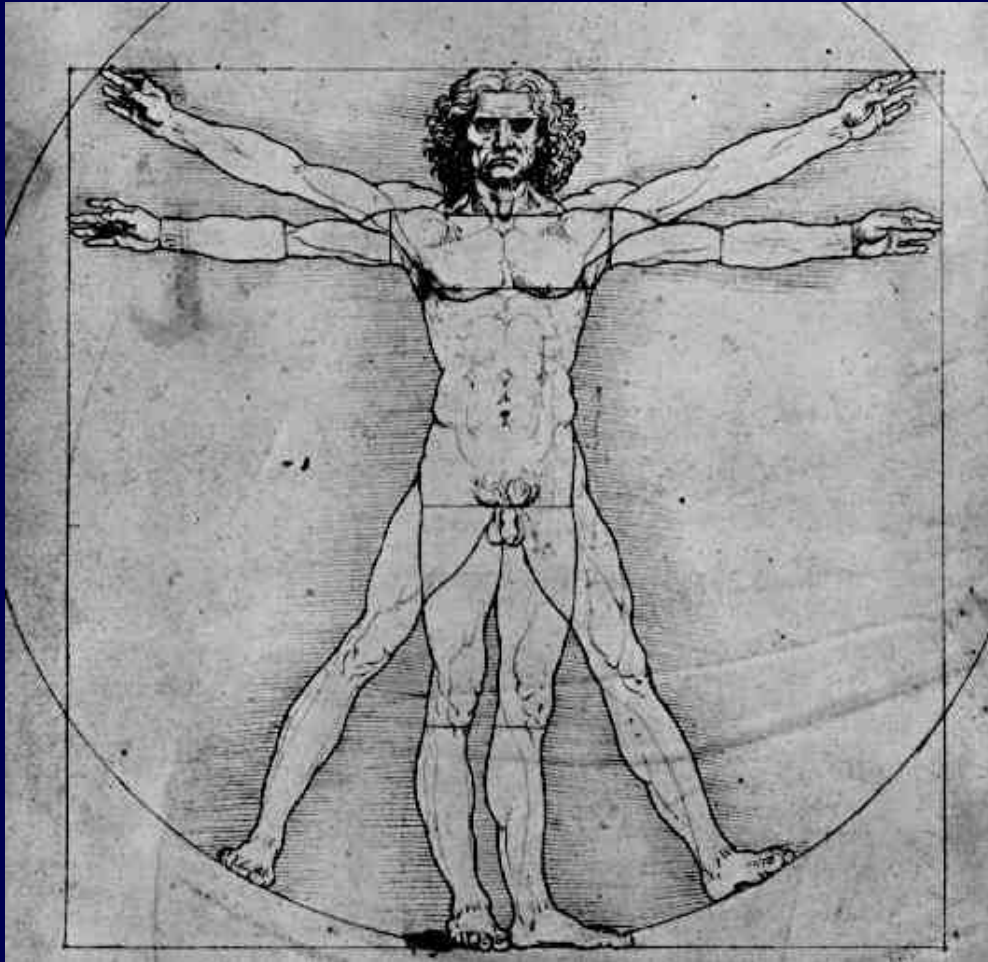
kbroman.org

github.com/kbroman

@kwbroman

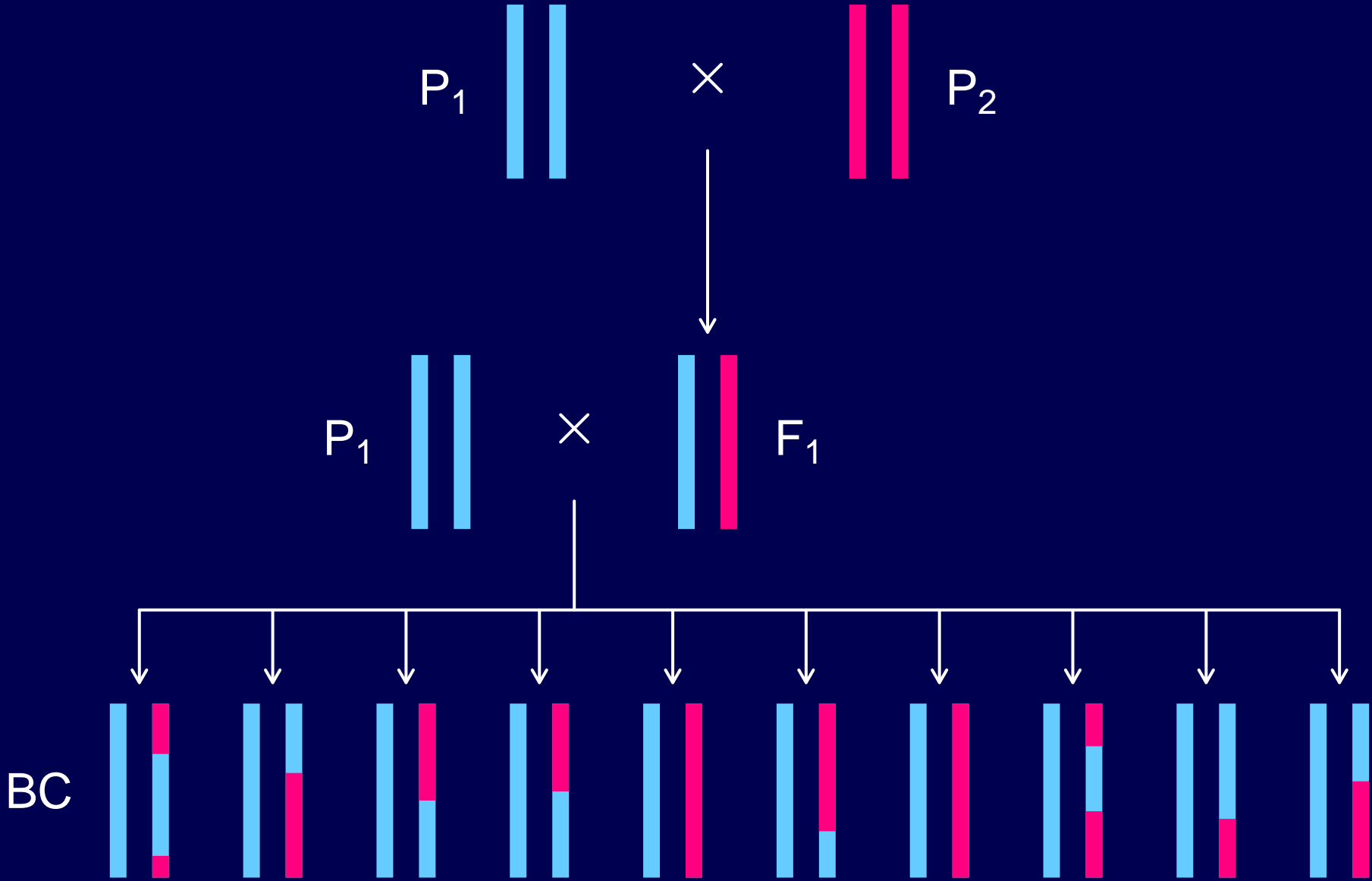


Human vs mouse

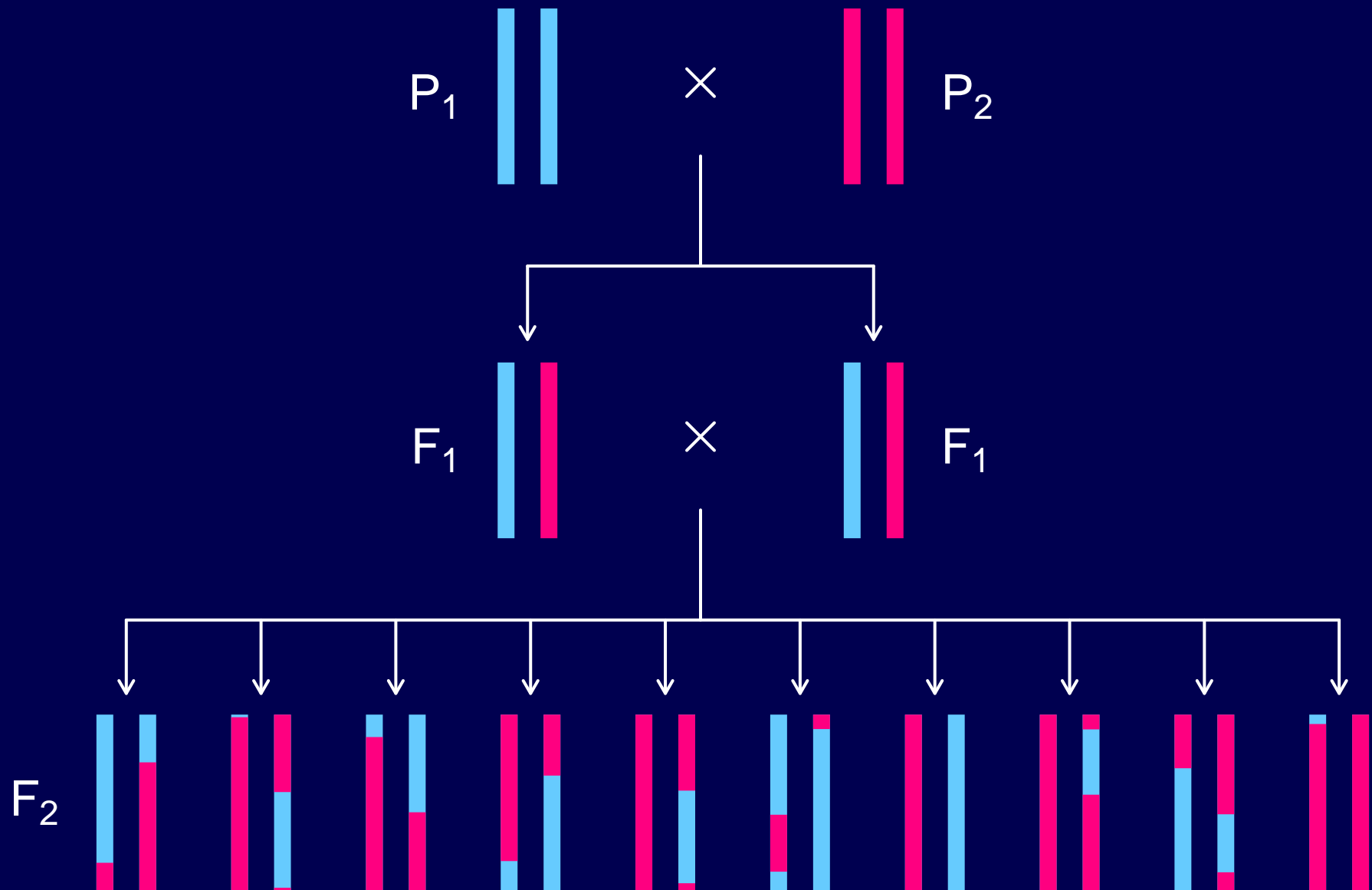


www.daviddeen.com

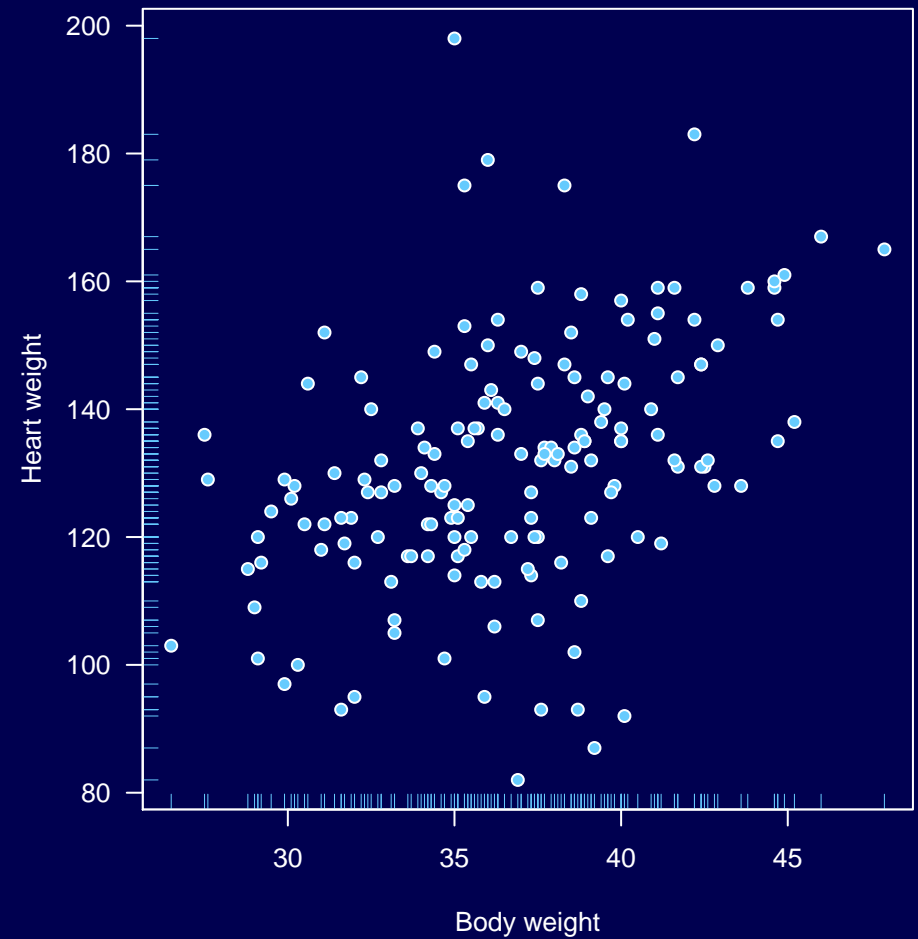
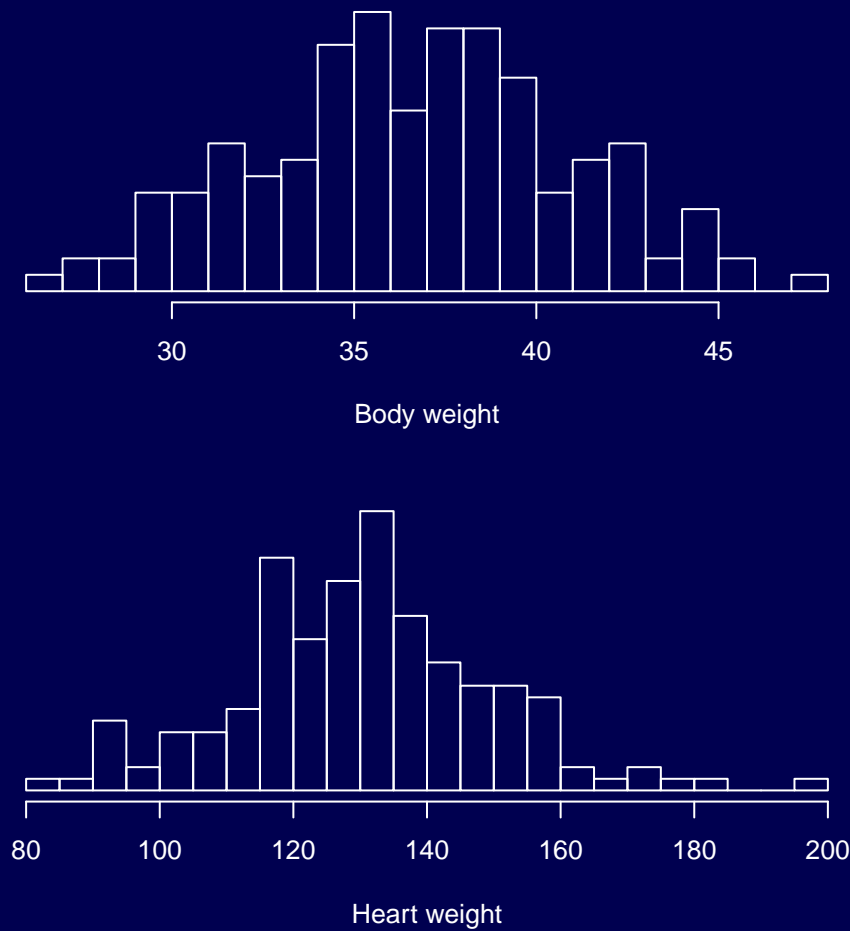
Backcross



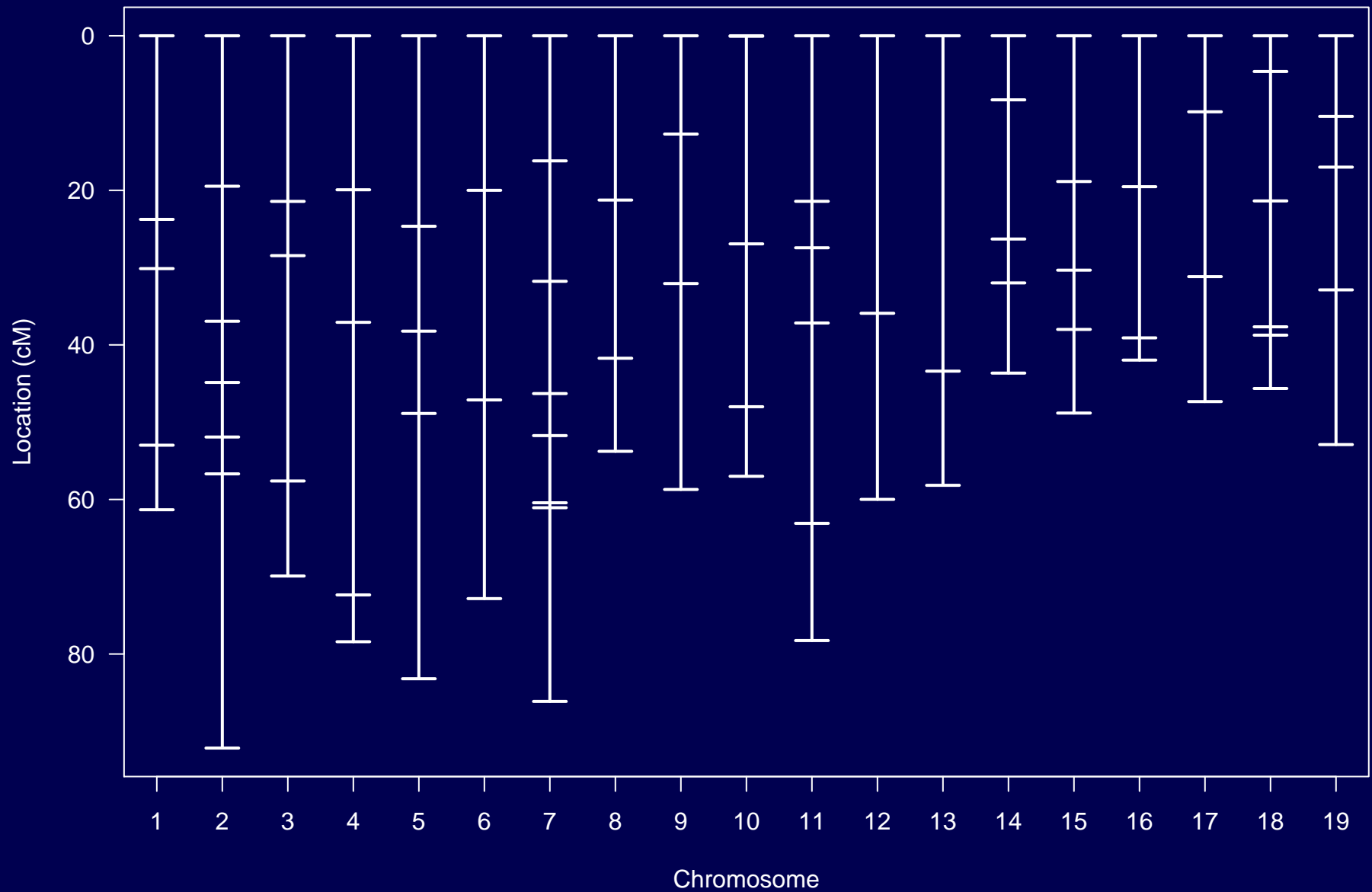
Intercross



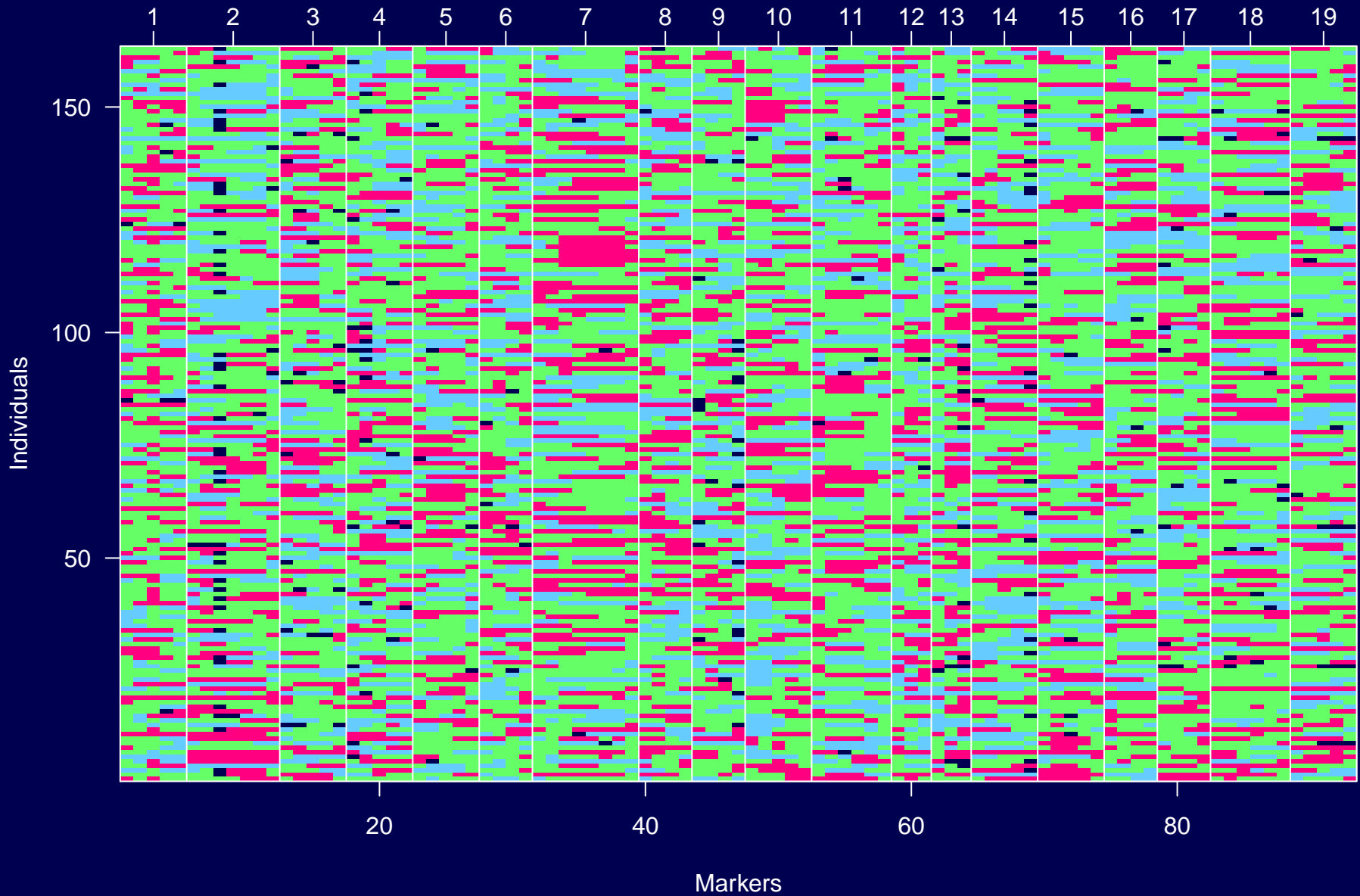
Phenotype data



Genetic map



Genotype data



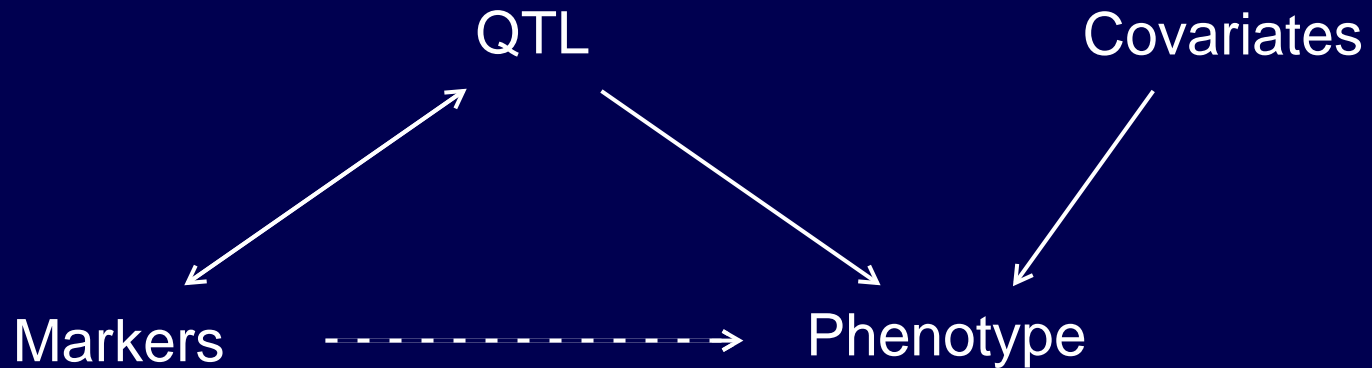
Typical data file

	A	B	C	D	E	F	G	H	I
1	liver	spleen	sex	pgm	D1Mit18	D1Mit80	D1Mit17	D2Mit379	D2Mit75
2					1	1	1	2	2
3					27.3	51.4	110.4	38.3	48.1
4	61.92	153.16	m	1	BB	SB	SB	SB	SB
5	88.33	178.58	m	1	-	-	-	BB	BB
6	58	131.91	m	1	BB	SB	SB	SB	SB
7	78.06	126.13	m	1	SB	SB	BB	SS	SS
8	65.31	181.05	m	1	-	-	-	SB	SB
9	59.26	191.54	m	1	-	-	-	SS	SS
10	59.47	154.88	m	1	BB	BB	BB	SB	SB
11	65.63	184.12	m	1	-	-	-	SB	SB
12	38.64	133.05	m	1	SB	BB	SB	SB	SB
13	60.94	275.63	m	1	-	-	-	SB	BB
14	51.48	395.25	m	1	-	-	-	SB	BB
15	47.12	260.45	m	1	BB	SB	SB	BB	BB

Goals

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

Statistical structure



The missing data problem:

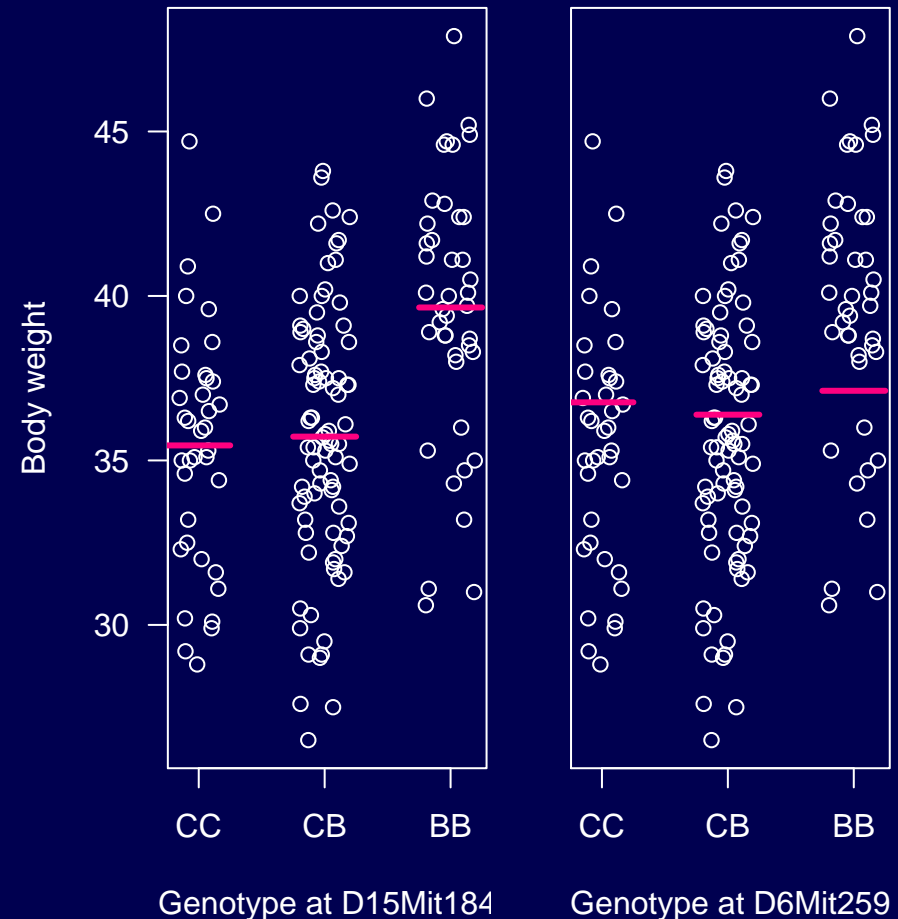
Markers \longleftrightarrow QTL

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

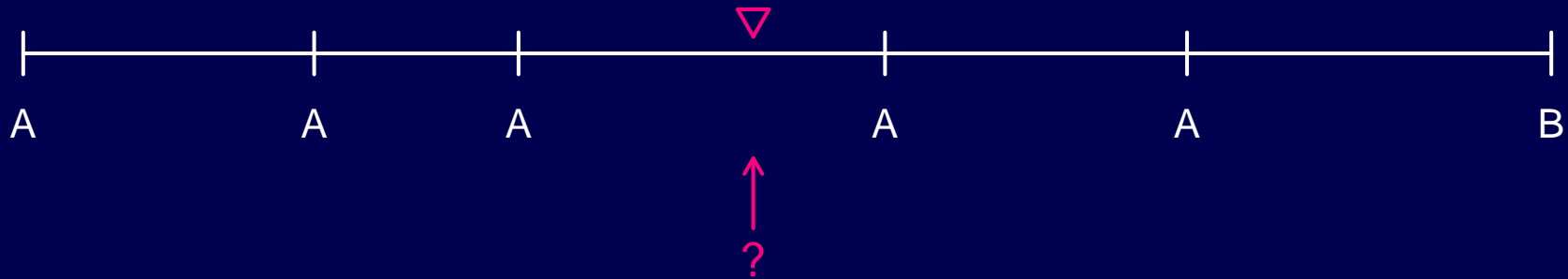
- Must exclude individuals with missing genotype data.
- Imperfect information about QTL location.
- Suffers in low density scans.
- Only considers one QTL at a time.

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 = the unobserved QTL genotype
Assume $y|q \sim N(\mu_q, \sigma)$
- We don't know q , but we can calculate $\Pr(q \mid \text{marker data})$
- Estimate μ_q, σ by *maximum likelihood* using an iterative EM algorithm

Genotype probabilities



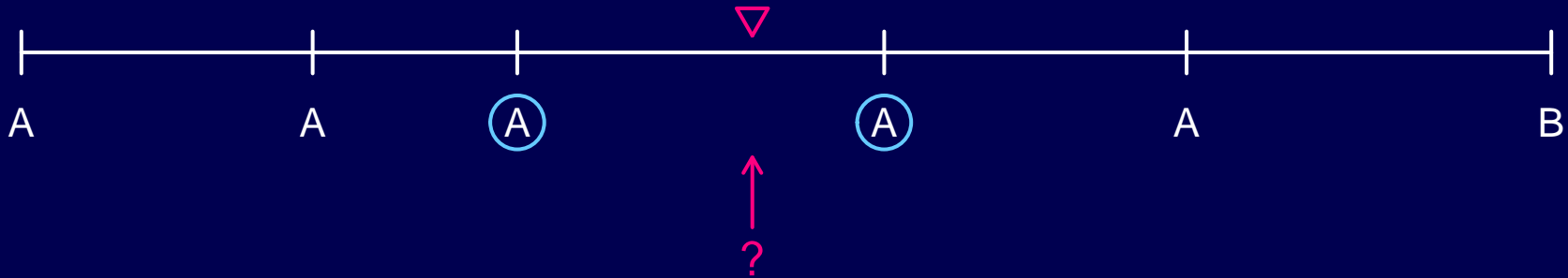
Calculate $\Pr(q \mid \text{marker data})$, assuming

- 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



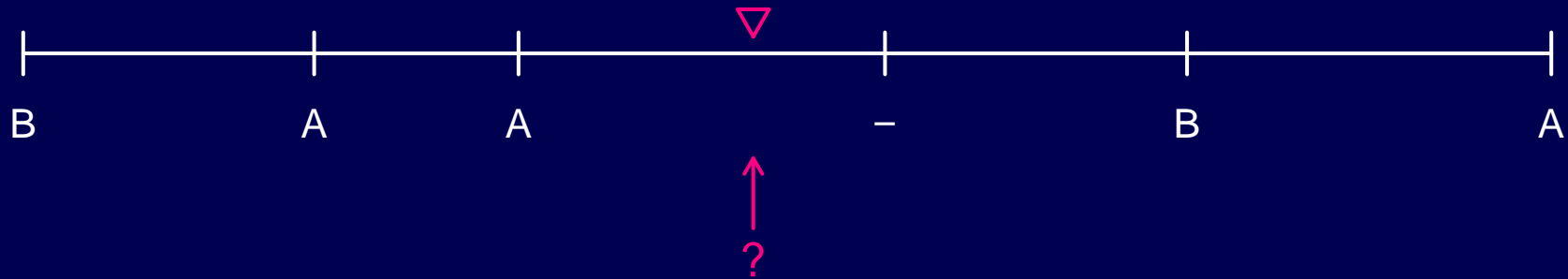
Calculate $\Pr(q \mid \text{marker data})$, assuming

- 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



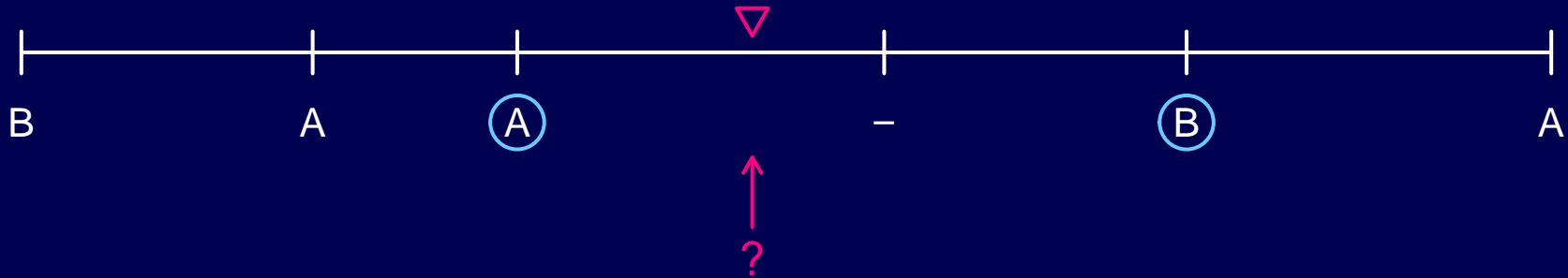
Calculate $\Pr(q \mid \text{marker data})$, assuming

- 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



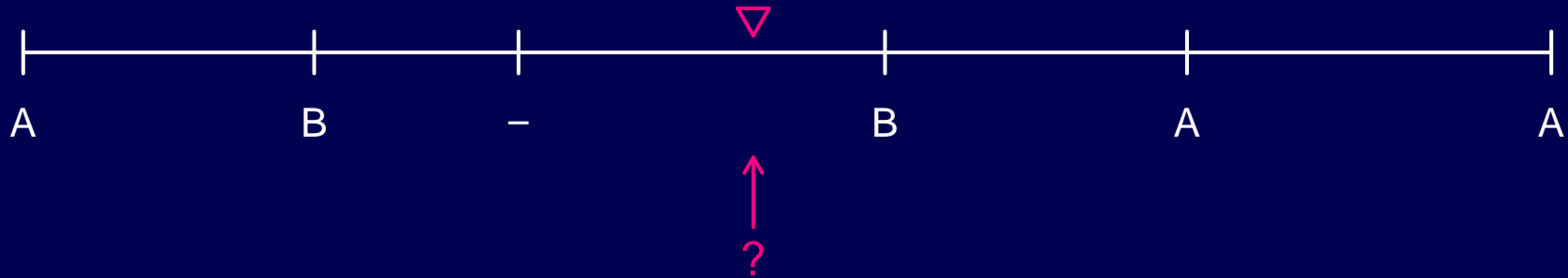
Calculate $\Pr(q \mid \text{marker data})$, assuming

- 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



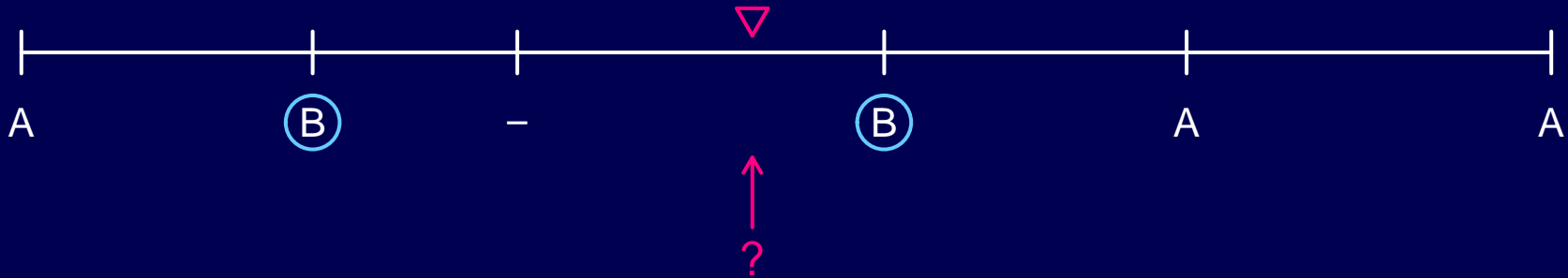
Calculate $\Pr(q \mid \text{marker data})$, assuming

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

EM algorithm

Dempster et al. (1977)

E step:

$$\begin{aligned}\text{Let } w_{ij}^{(k)} &= \Pr(q_i = j | y_i, \text{marker data}, \hat{\mu}_0^{(k-1)}, \hat{\mu}_1^{(k-1)}, \hat{\sigma}^{(k-1)}) \\ &= \frac{p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}{\sum_j p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}\end{aligned}$$

M step:

$$\begin{aligned}\text{Let } \hat{\mu}_j^{(k)} &= \sum_i y_i w_{ij}^{(k)} / \sum_i w_{ij}^{(k)} \\ \hat{\sigma}^{(k)} &= \sqrt{\sum_i \sum_j w_{ij}^{(k)} (y_i - \hat{\mu}_j^{(k)})^2 / n}\end{aligned}$$

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.

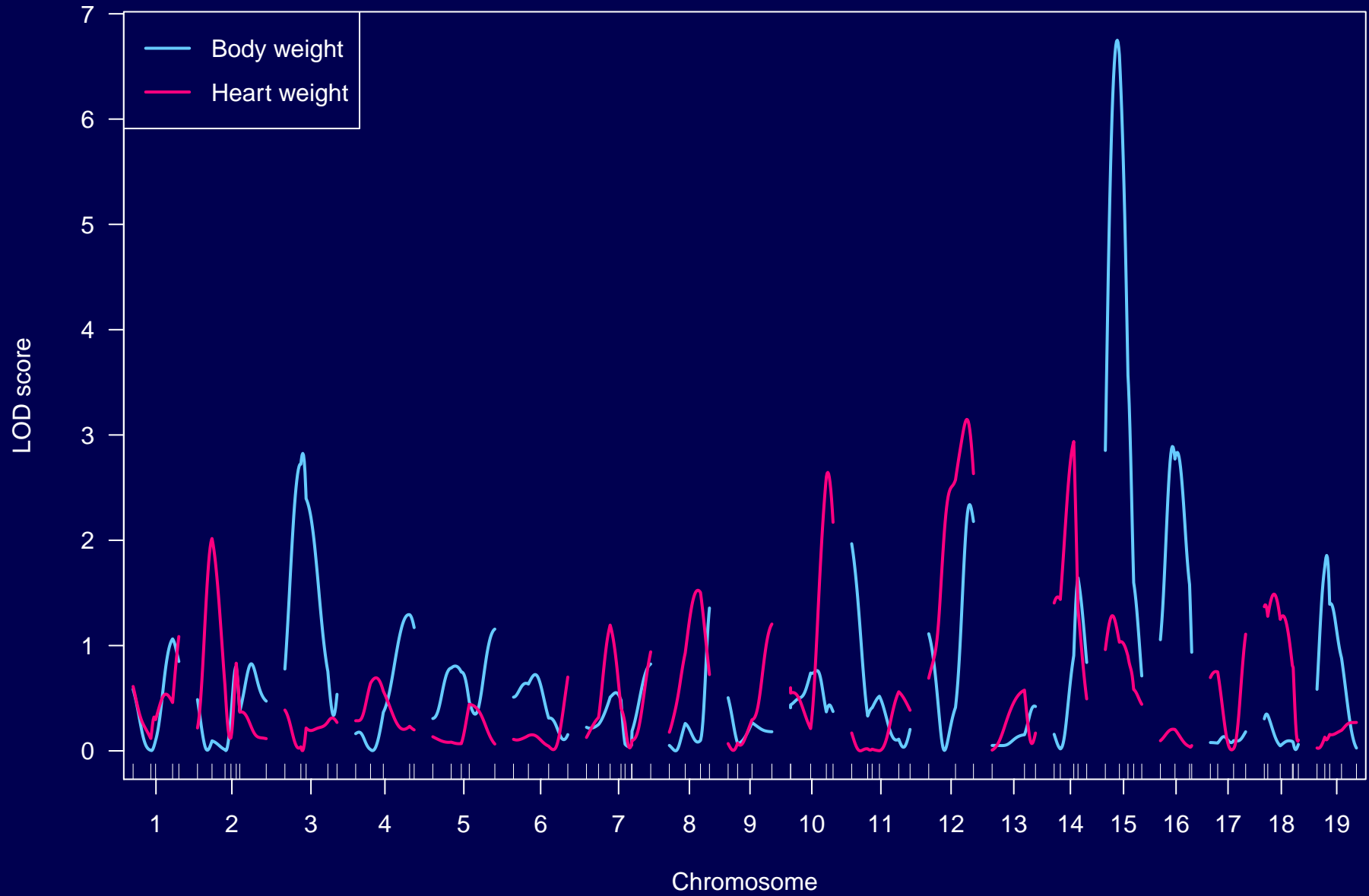
$\text{LOD}(\lambda) = \log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position λ versus that of no QTL

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

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

No QTL model: The phenotypes are independent and identically distributed (iid) $N(\mu, \sigma^2)$.

Results



→ R

- `read.cross()`
- `summary()`, `plot()`
- `nind()`, `nmar()`, `totmar()`, `nchr()`, `nphe()`
- `calc.genoprob()`
- `scanone()`

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?

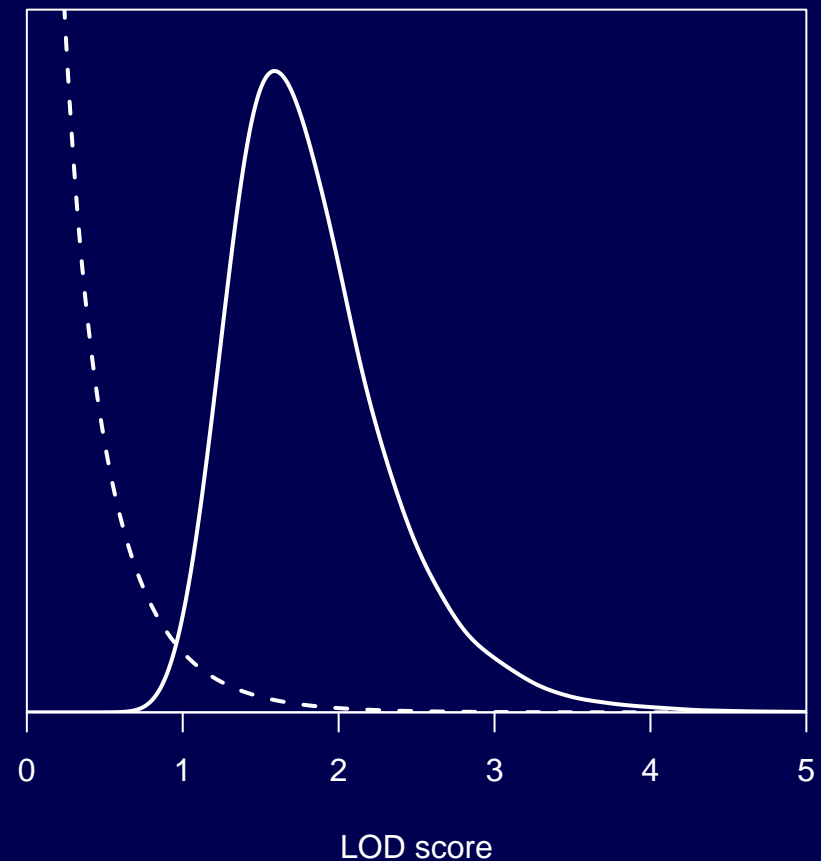
LOD threshold = 95 %ile of distr'n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:

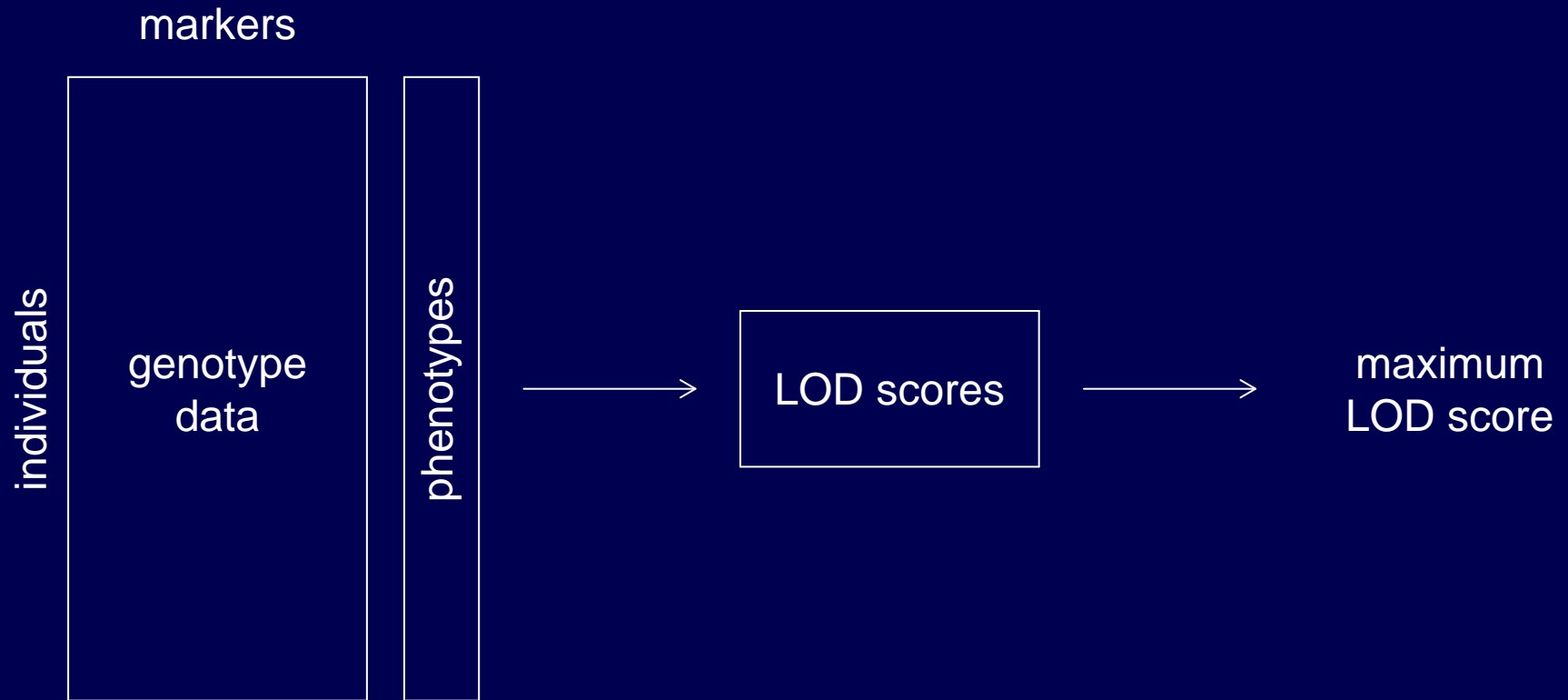
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)

Null distribution of the LOD score

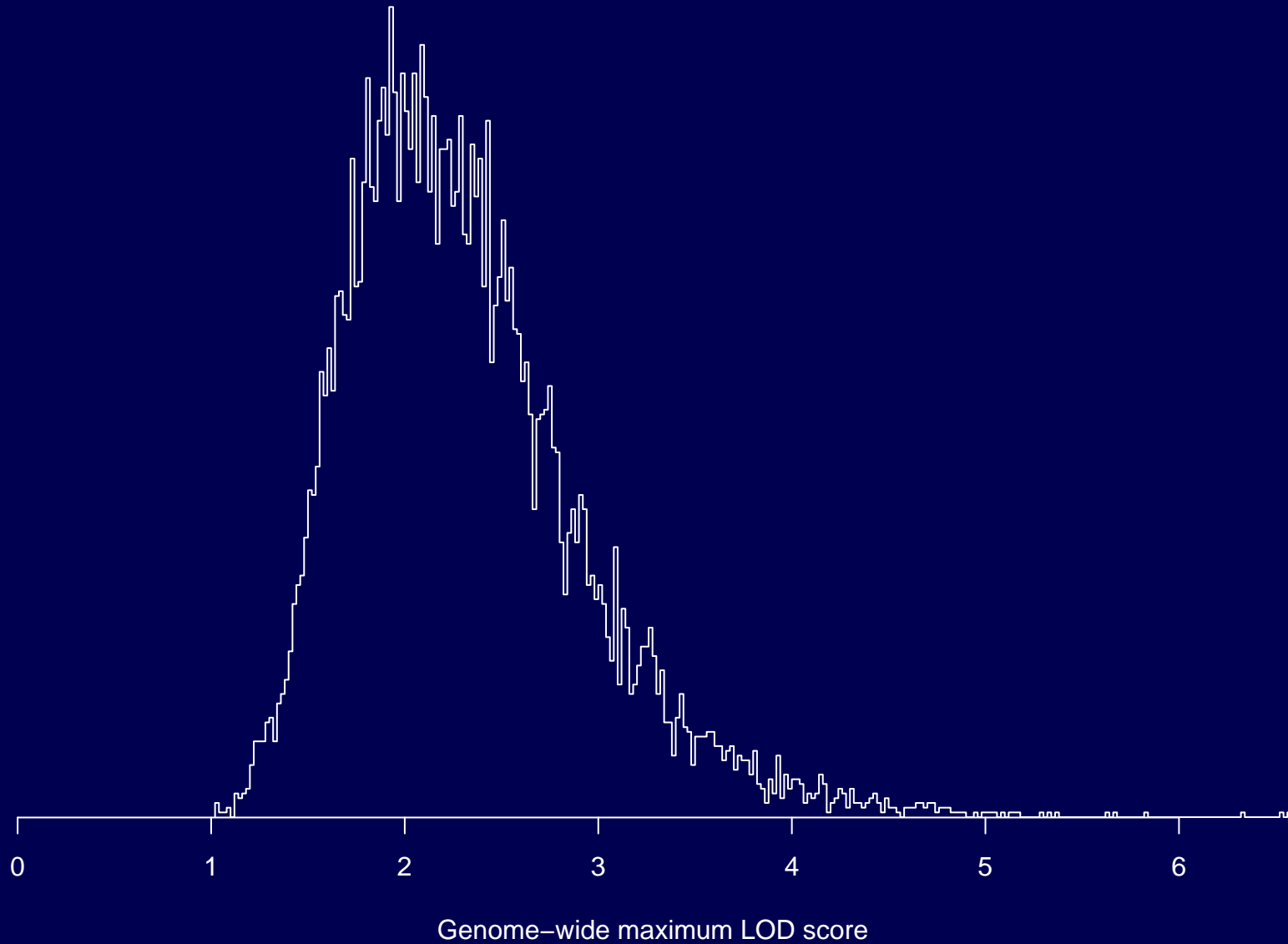
- Null distribution derived by computer simulation of backcross with genome of typical size.
- Dashed curve: distribution of LOD score at any one point.
- Solid curve: distribution of maximum LOD score, genome-wide.



Permutation test



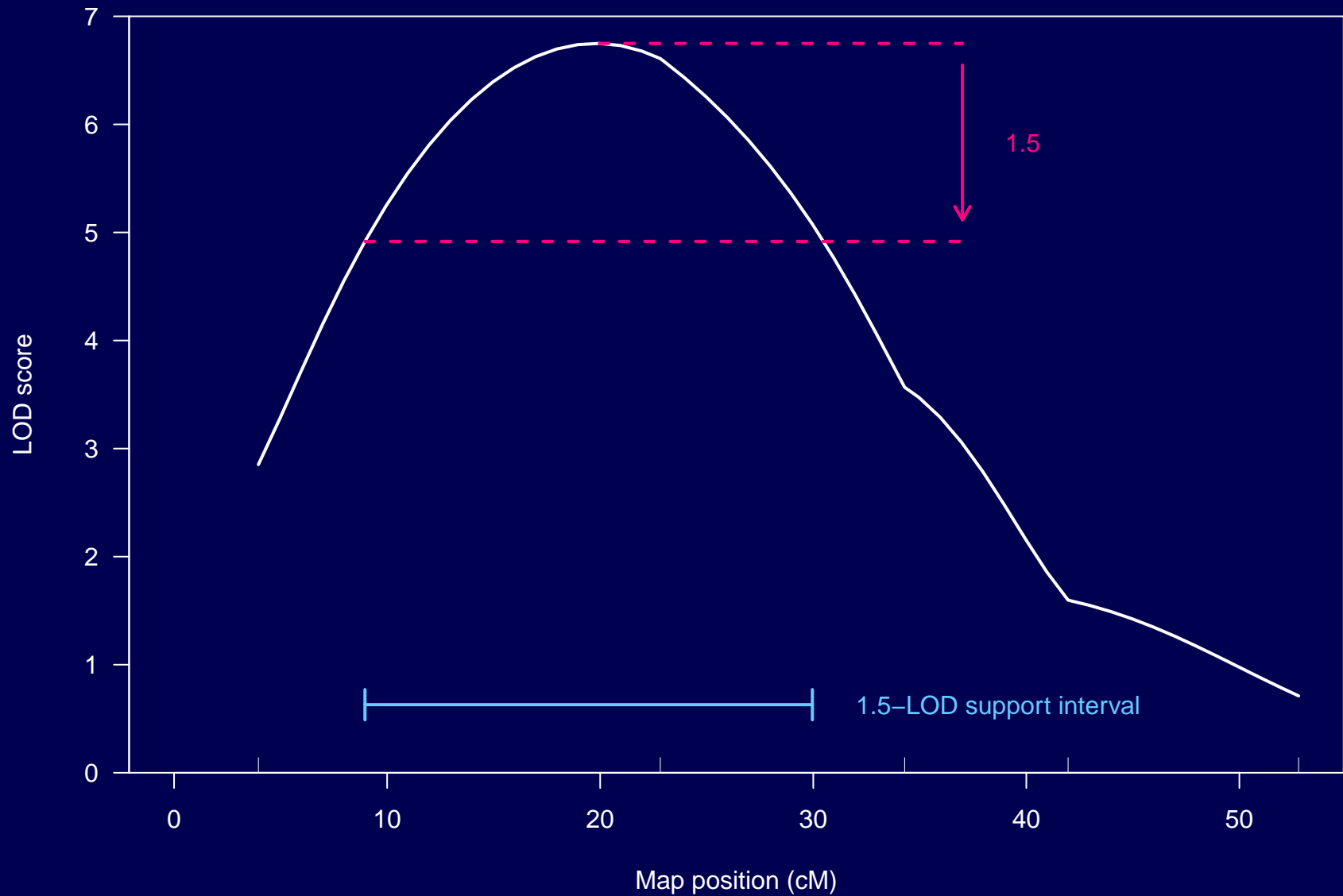
Permutation results



→ R

- `scanone()` for permutations

LOD support intervals



→ R

- `lodint()`
- `bayesint()`

Haley-Knott regression

A quick approximation to Interval Mapping.

$$E(y_i|q_i) = \mu_q$$

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

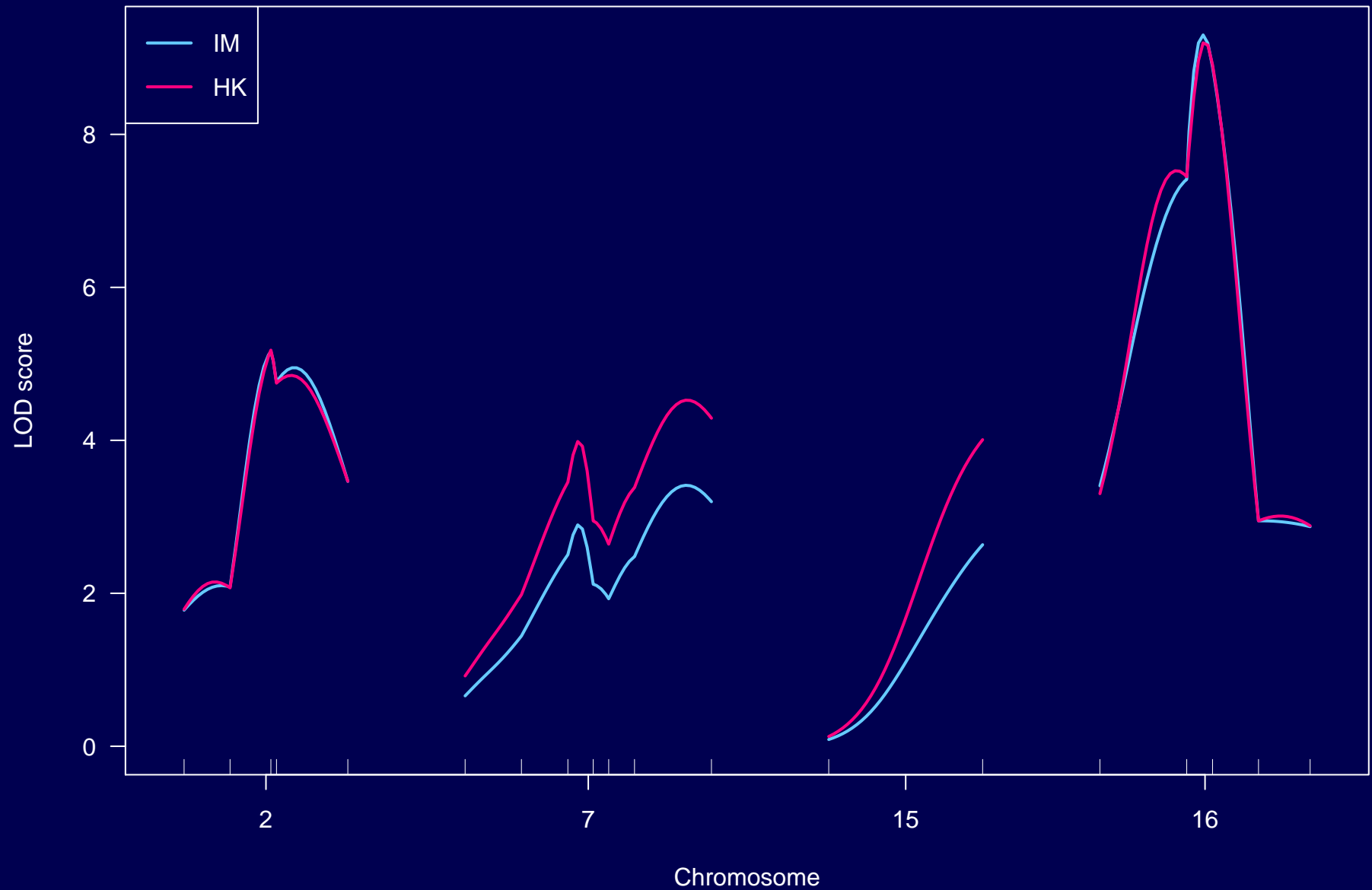
Regress y on p_i , pretending the residual variation is normally distributed (with constant variance).

$$\text{LOD} = \frac{n}{2} \log_{10} \left(\frac{\text{RSS}_0}{\text{RSS}_1} \right)$$

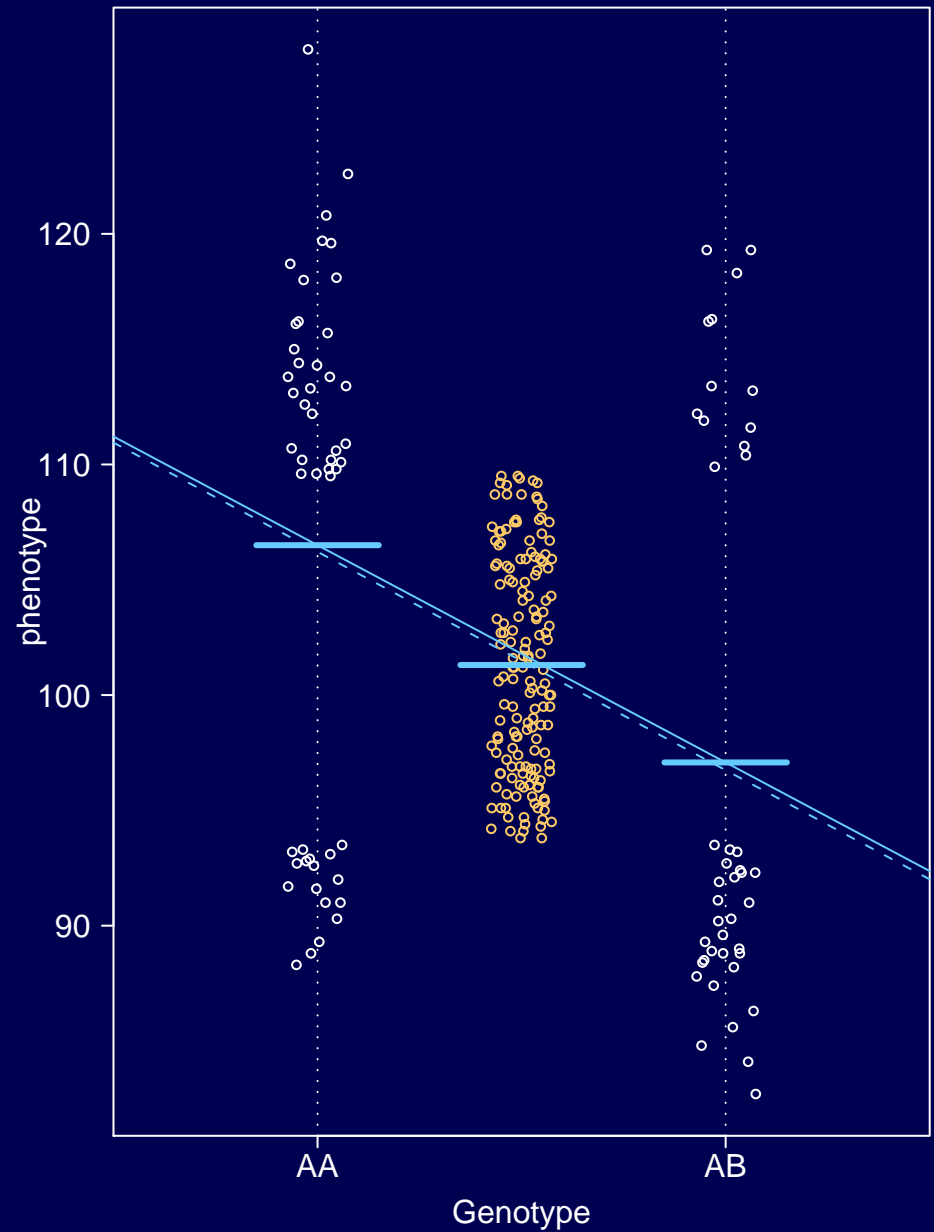
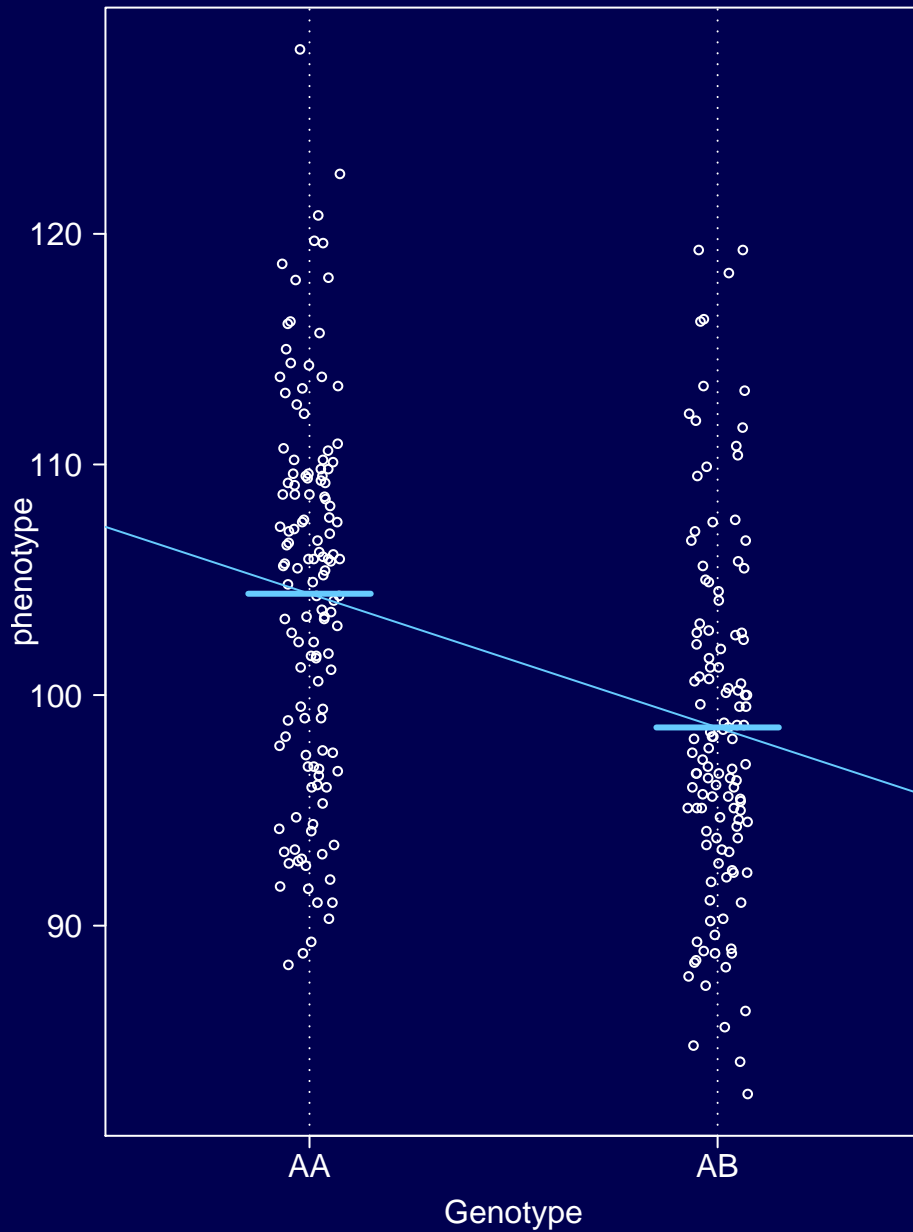
→ R

- `scanone()` with `method="hk"`

Haley-Knott results



H-K with selective genotyping



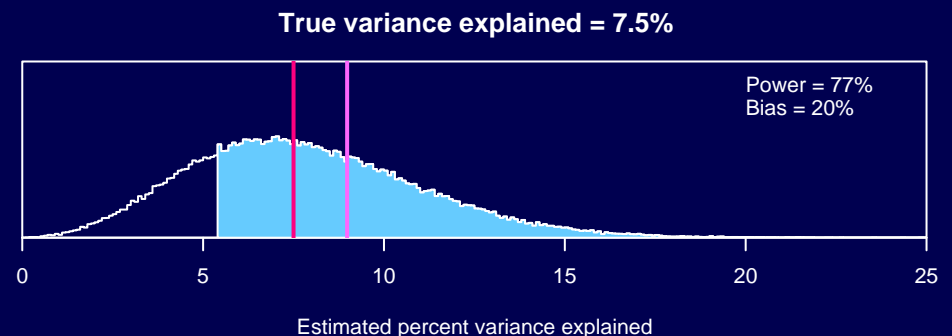
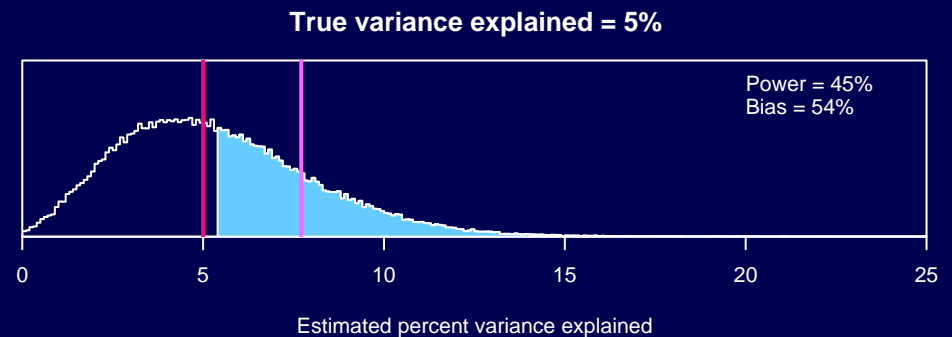
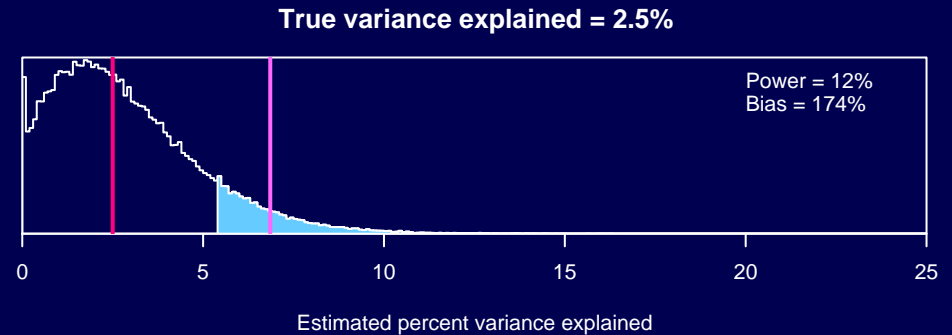
Data diagnostics

- 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

See Ch 3 in the R/qtl book, rqtl.org/book

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

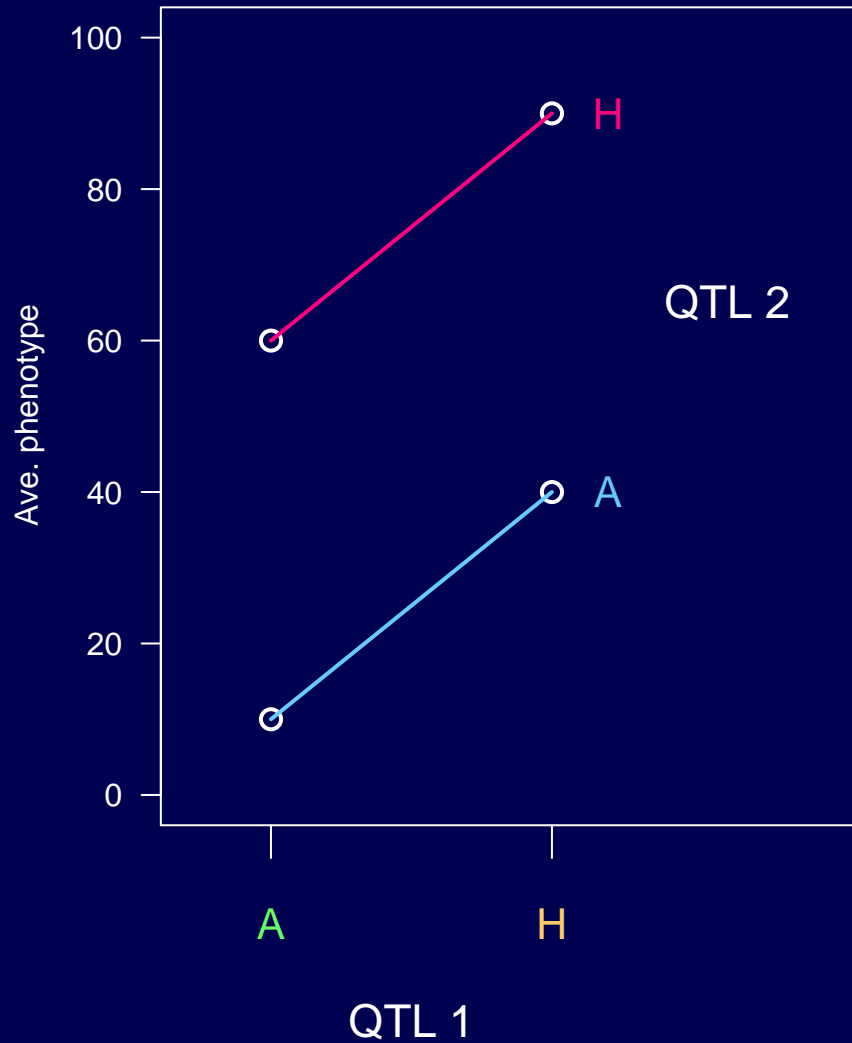
- Estimated % variance explained by identified QTLs
- Repeating an experiment
- Congenics (aka near isogenic lines)
- Marker-assisted selection

Modelling multiple QTL

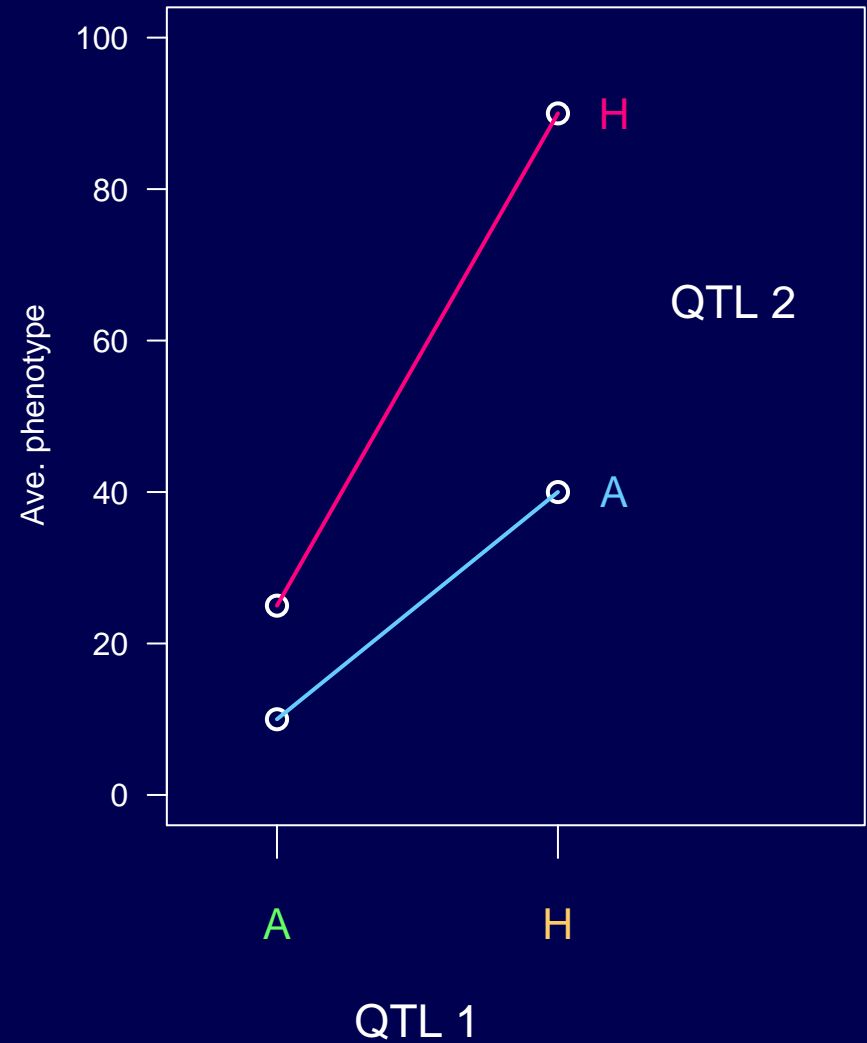
- Reduce residual variation \implies increased power
- Separate linked QTL
- Identify interactions among QTL

Epistasis in BC

Additive

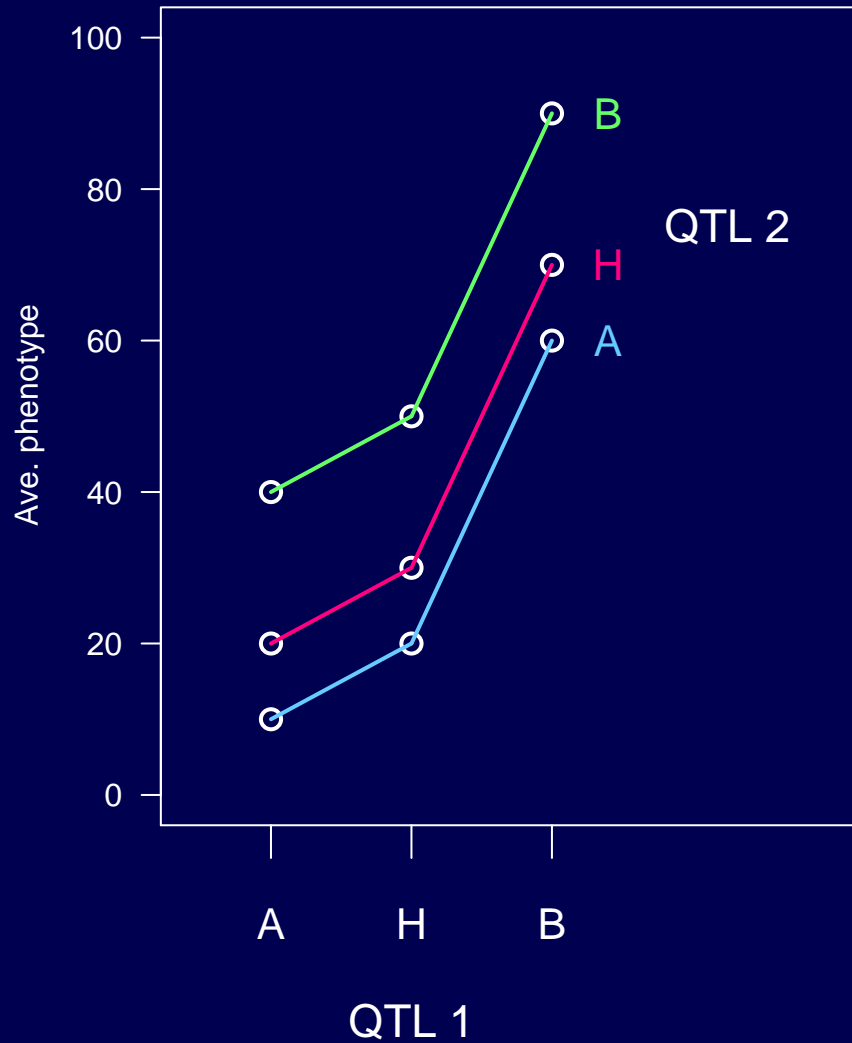


Epistatic

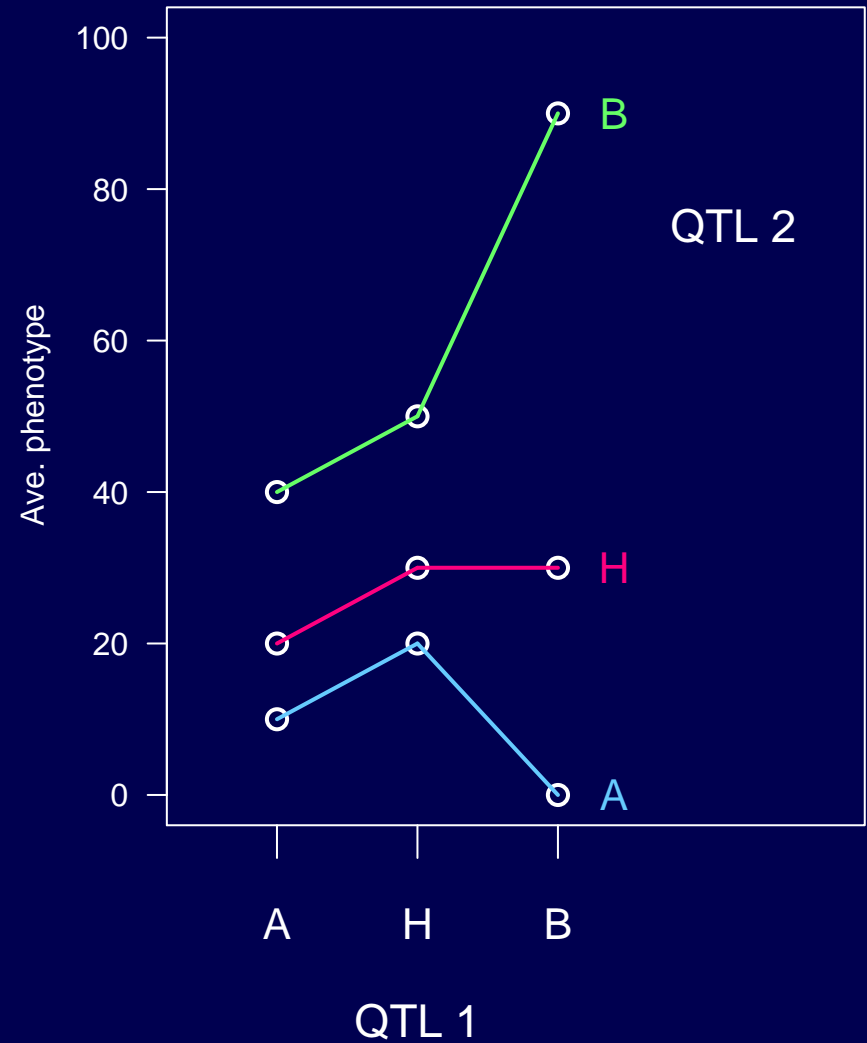


Epistasis in F_2

Additive



Epistatic



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.

References

- Beavis WD (1994). The power and deceit of QTL experiments: Lessons from comparative QTL studies. In DB Wilkinson, (ed) 49th Ann Corn Sorghum Res Conf, pp 252–268. Amer Seed Trade Asso, Washington, DC.
Discusses selection bias in estimated QTL effects.
- Broman KW (2003) Mapping quantitative trait loci in the case of a spike in the phenotype distribution. Genetics 163:1169–1175
Two-part model; also discusses binary traits and non-parametric QTL mapping.
- 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
- Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159: 371–387
Multiple imputation
- Solberg LC, et al. (2004) Sex- and line-specific lineage inheritance of depression-like behavior in the rat. Mamm Genome 15:648–662
Additive and interactive covariates.
- Broman KW et al (2006) The X chromosome in quantitative trait locus mapping. Genetics 174:2151–2158

References

- Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. *J Roy Stat Soc B* 64:641–656
Multiple-QTL model selection with additive QTL.
- Manichaikul A, Moon JY, Sen Ś, Yandell BS, Broman KW (2009) A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics* 181:1077–1086
Also account for epistasis.