R/qtl Workshop

Karl Broman

Biostatistics and Medical Informatics
University of Wisconsin – Madison

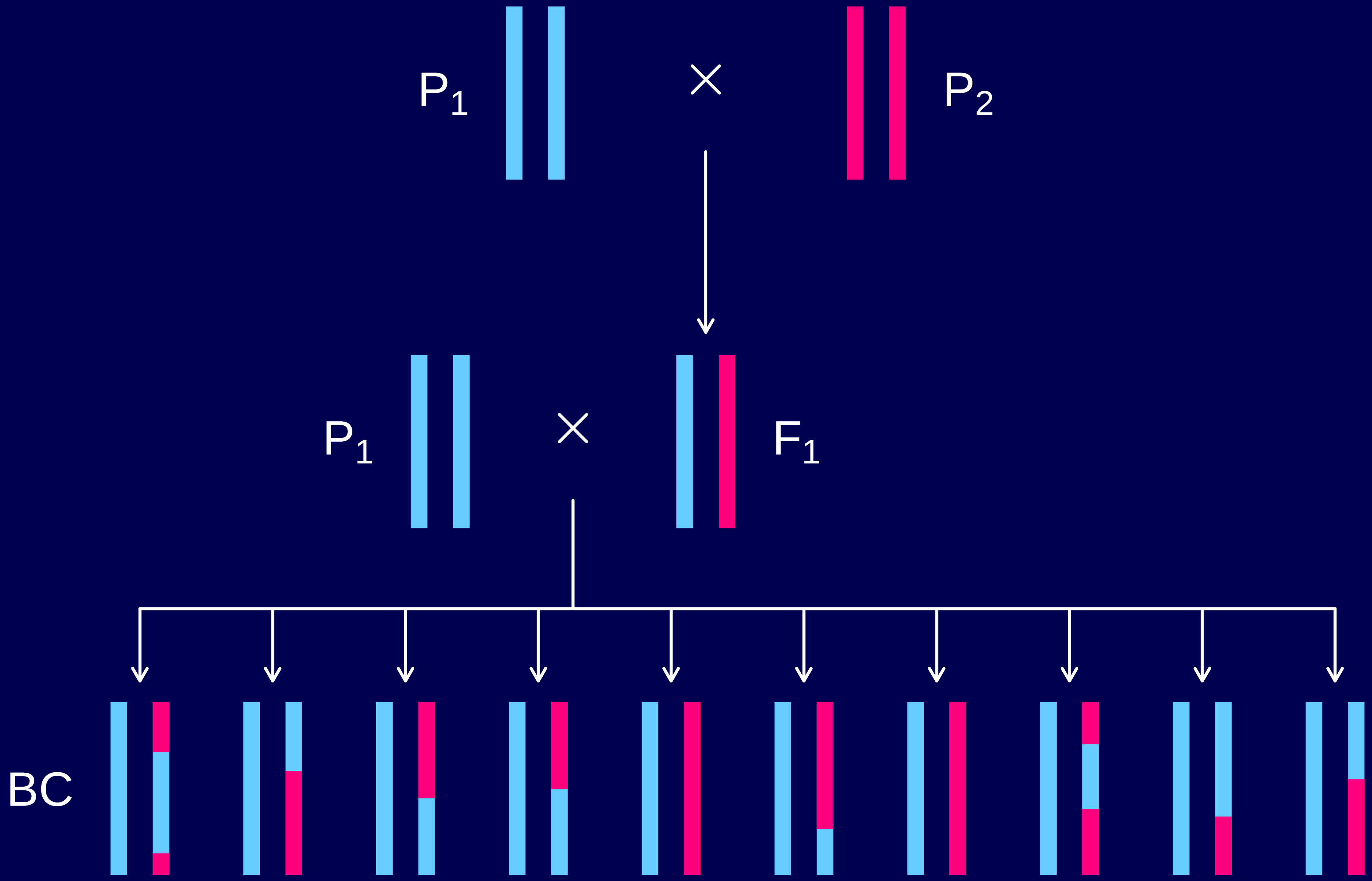
rqtl.org

kbroman.org

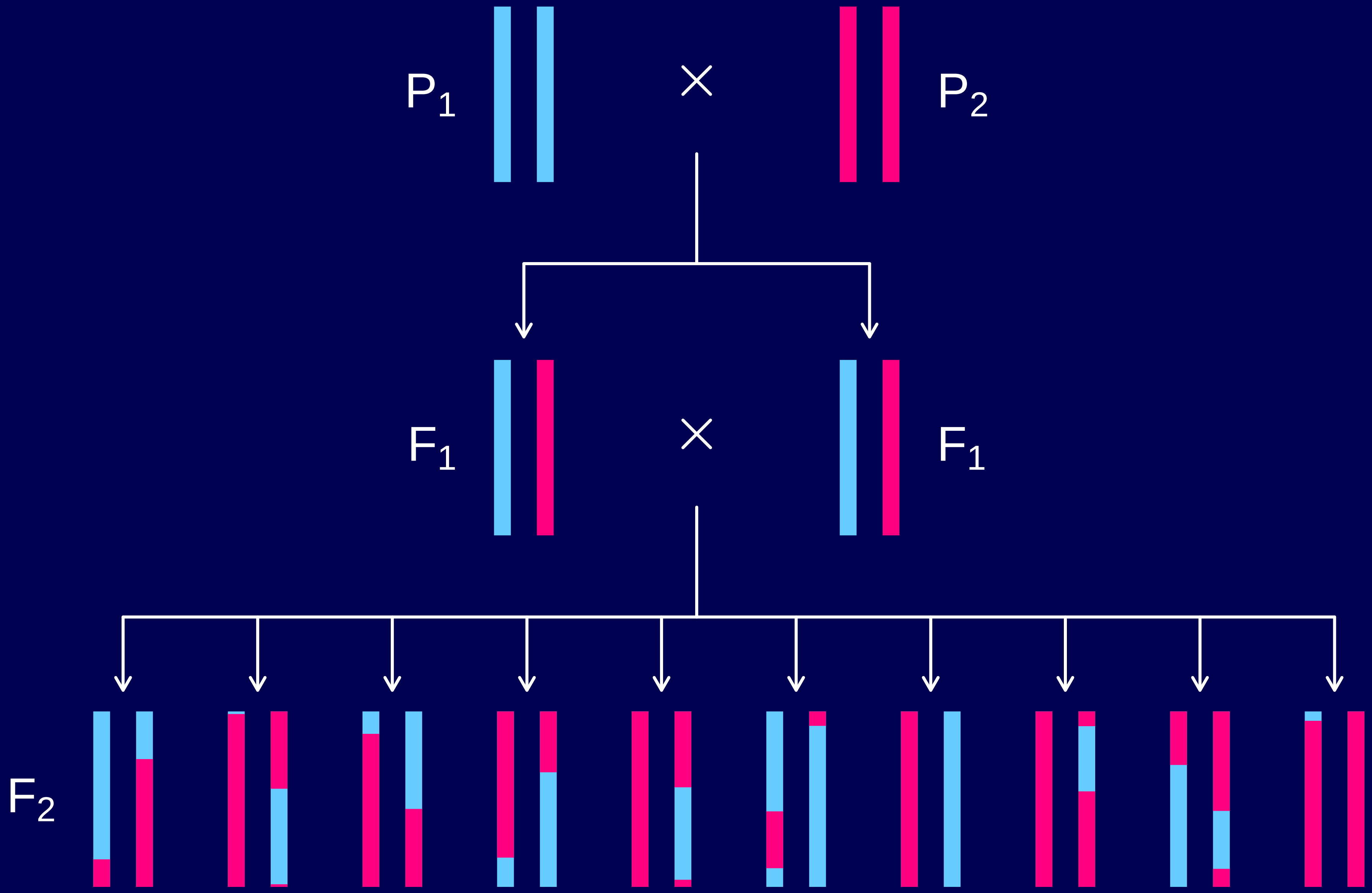
github.com/kbroman

@kwbroman

Backcross



Intercross



Goals

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

- Interval estimates of QTL location

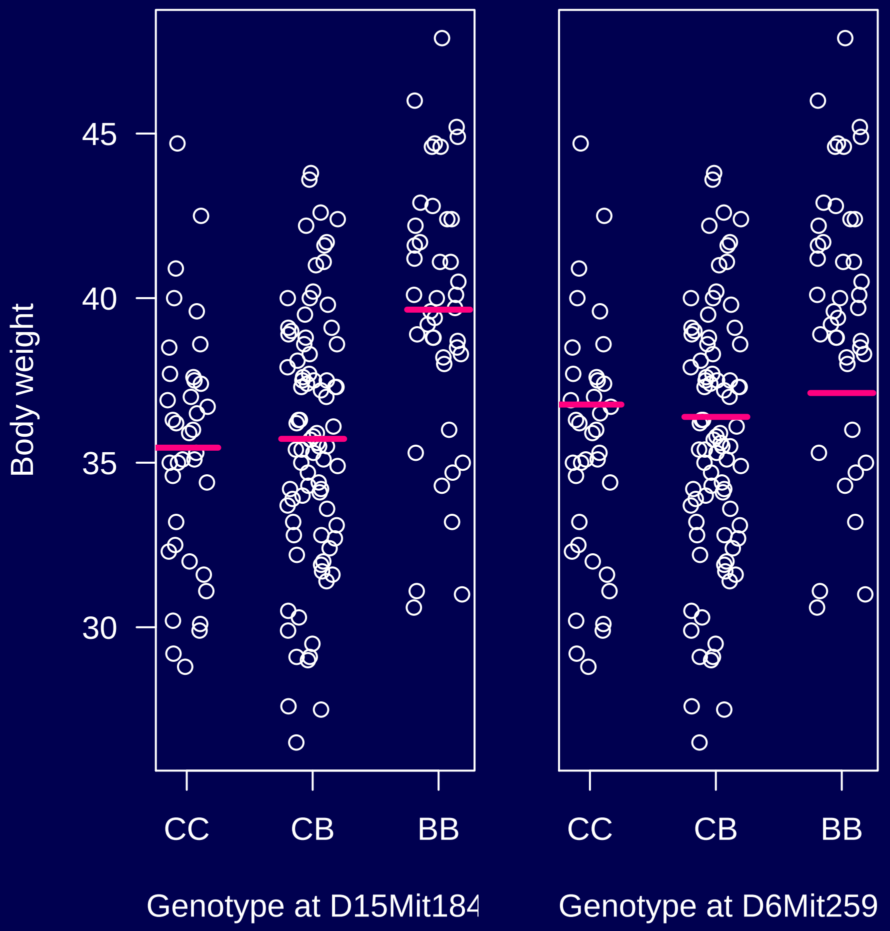
- Estimated QTL effects

→ R

- R, RStudio, and R/qtl
- `read.cross()`
- `summary()`, `plot()`
- `nind()`, `nmar()`, `totmar()`, `nchr()`, `nphe()`

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.

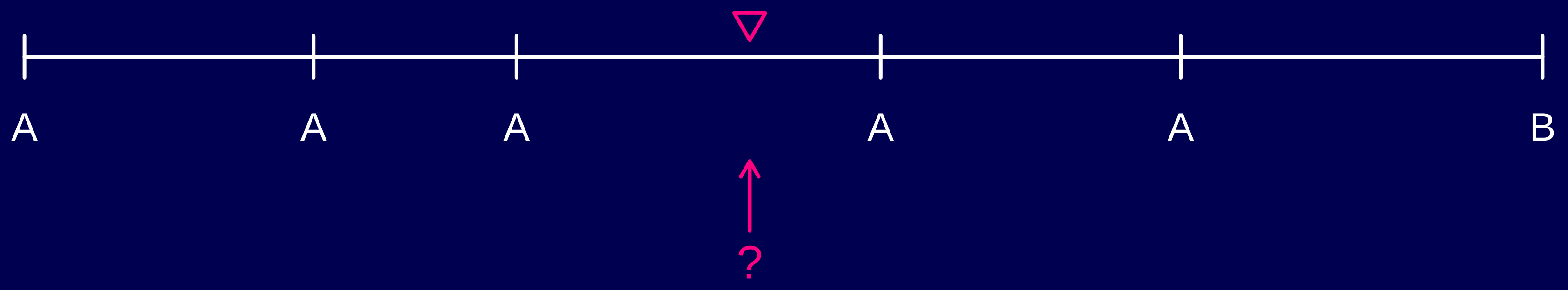
- Each position in the genome, one at a time, is posited as the putative QTL.

- Let $\mathsf{q} = $ the unobserved QTL genotype
  Assume $\mathsf{y}|\mathsf{q} \sim \mathsf{N}(\mu_{\mathsf{q}}, \sigma)$

- We don't know q, but we can calculate $\Pr(\mathsf{q} \mid \text{marker data})$

- Estimate μ_{q}, σ by *maximum likelihood* using an iterative EM algorithm

Genotype probabilities



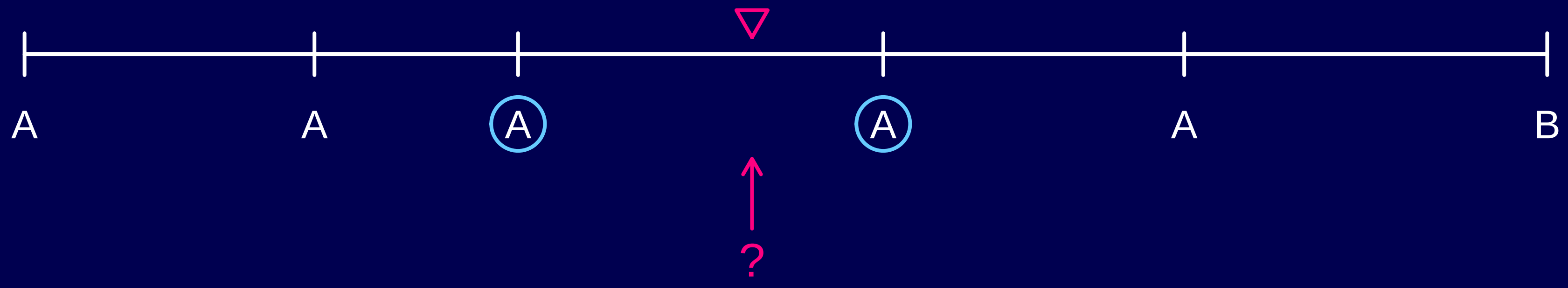
Calculate $\Pr(\mathbf{q} \mid \text{marker data})$, assuming

- 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



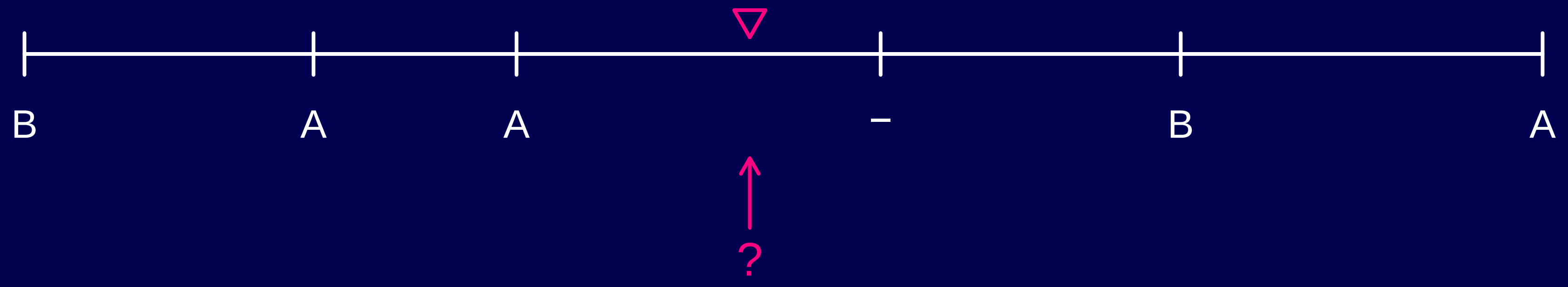
Calculate $\Pr(\mathbf{q} \mid \text{marker data})$, assuming

- 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



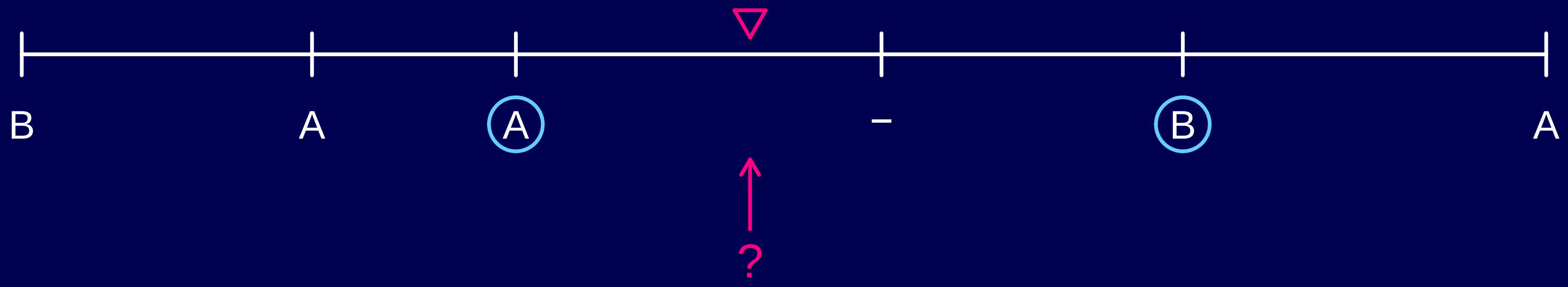
Calculate $\Pr(\mathbf{q} \mid \text{marker data})$, assuming

- 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



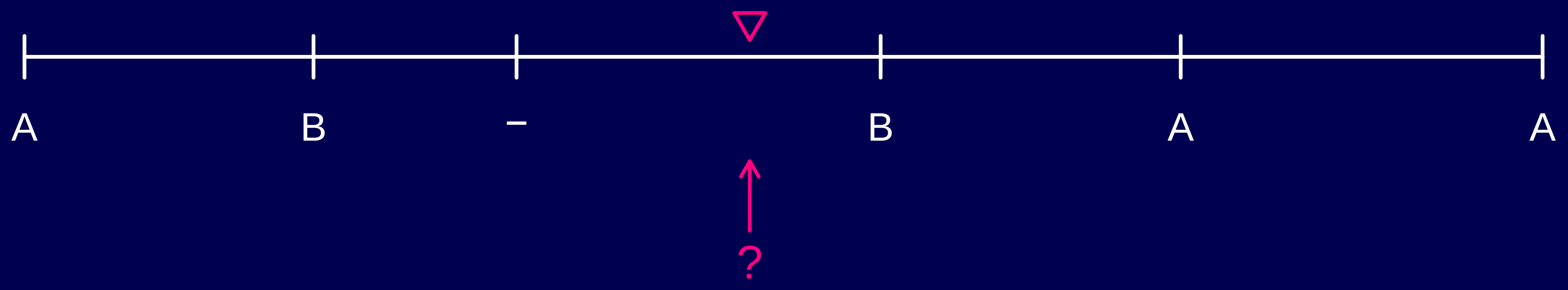
Calculate $\Pr(\mathbf{q} \mid \text{marker data})$, assuming

- 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



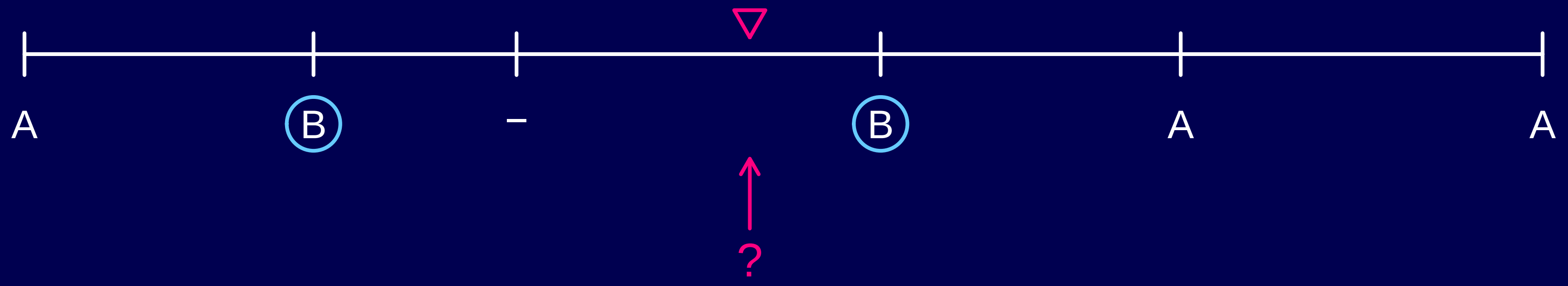
Calculate $\mathrm{Pr}(\mathbf{q} \mid \text{marker data})$, assuming

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

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).

→ R

- `calc.genoprob()`
- `scanone()`
- `iplotScanone()` from R/qtlcharts

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?

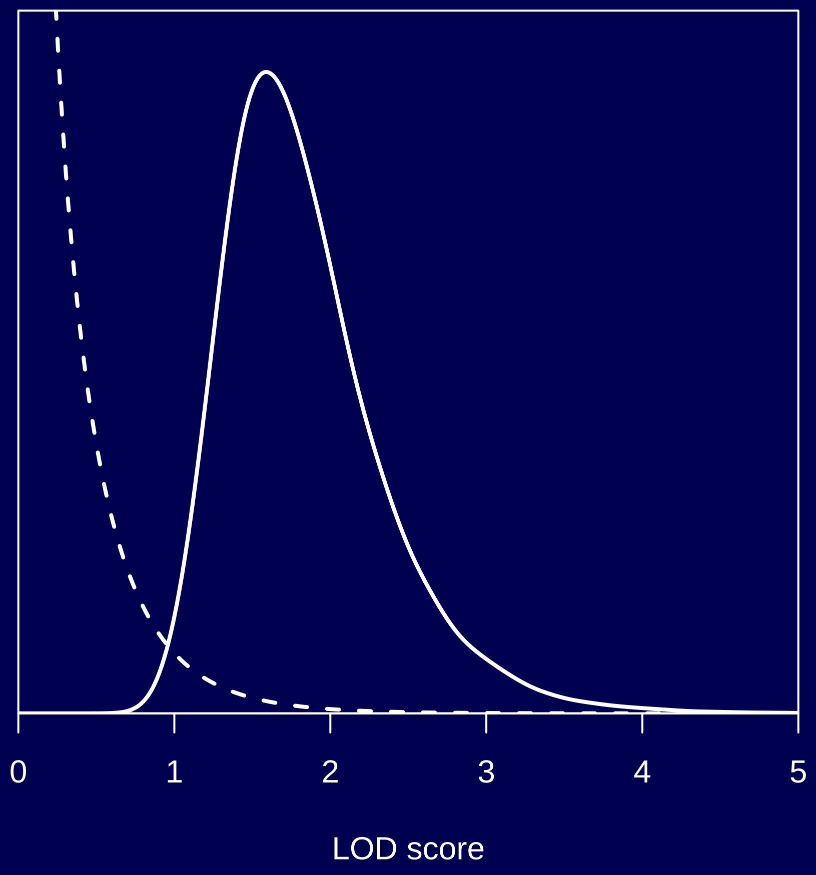
LOD threshold = 95 %ile of distr’n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:

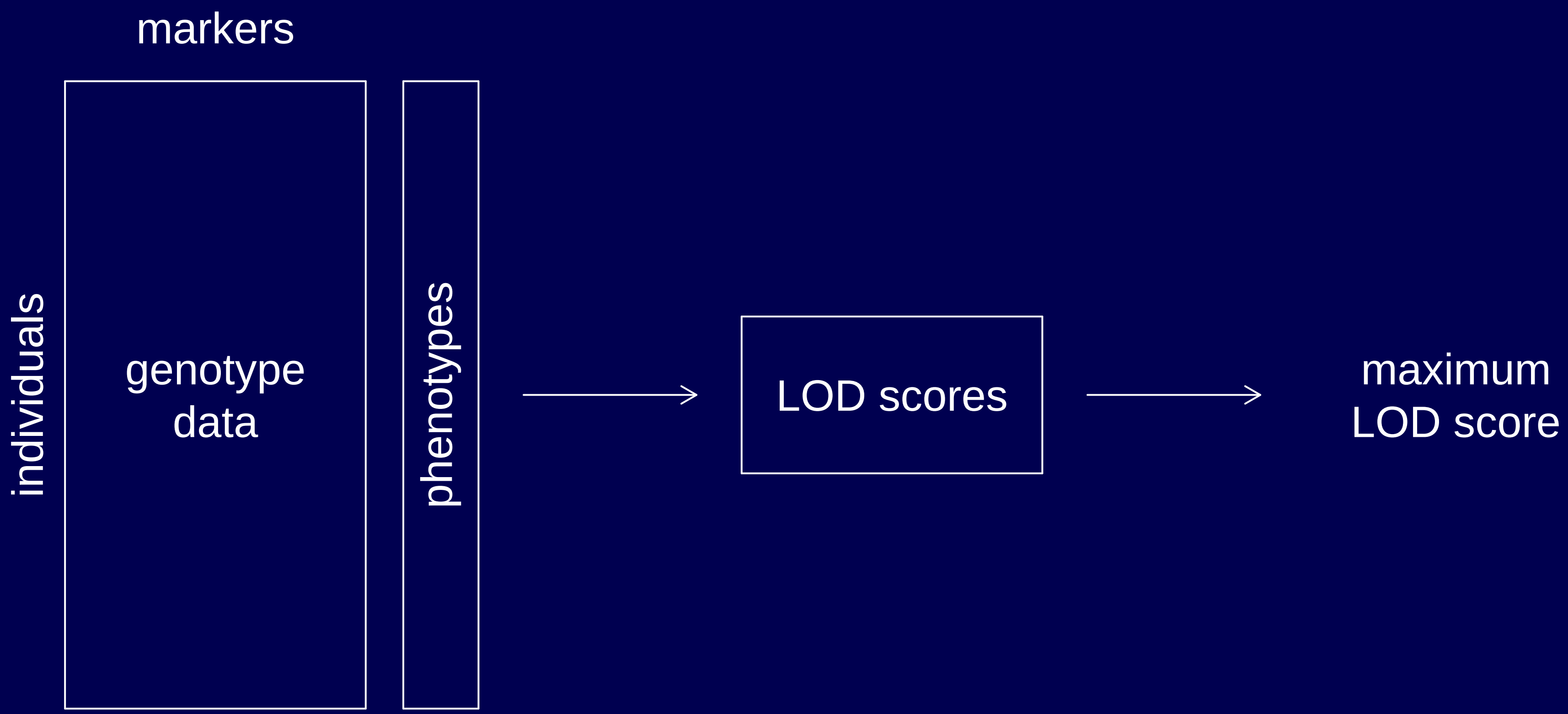
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)

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

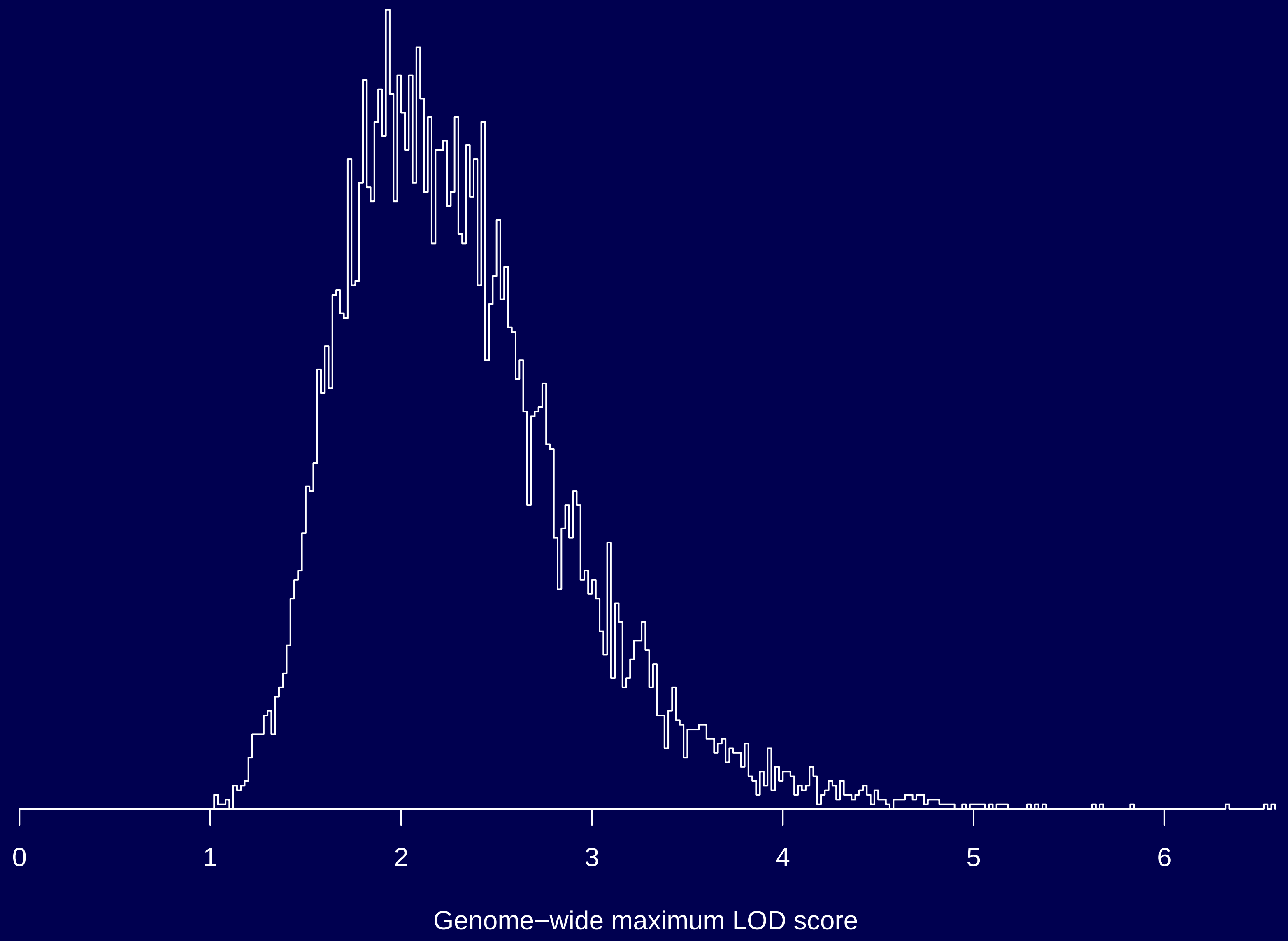
genotype data

phenotypes

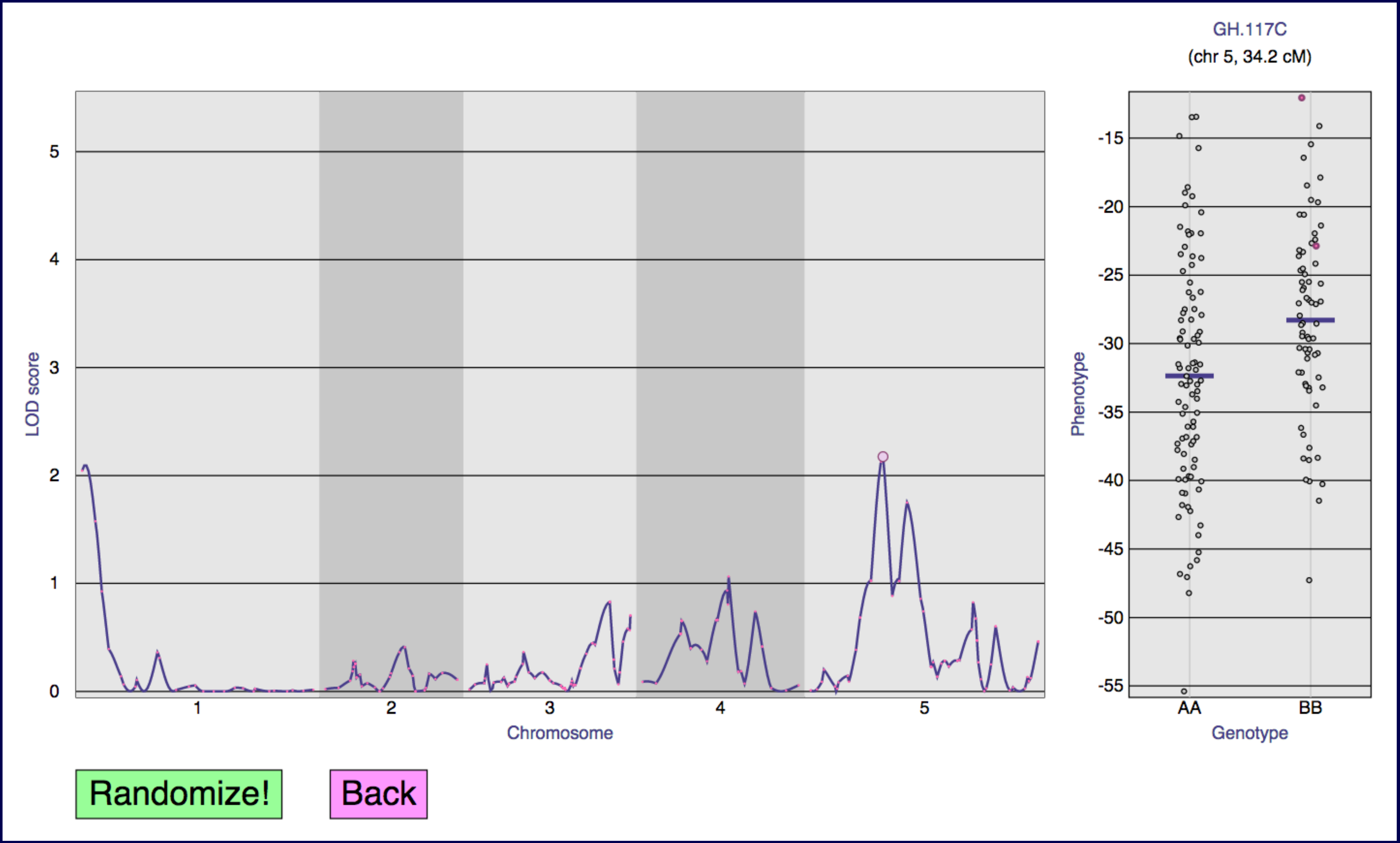
LOD scores

maximum LOD score

Permutation results



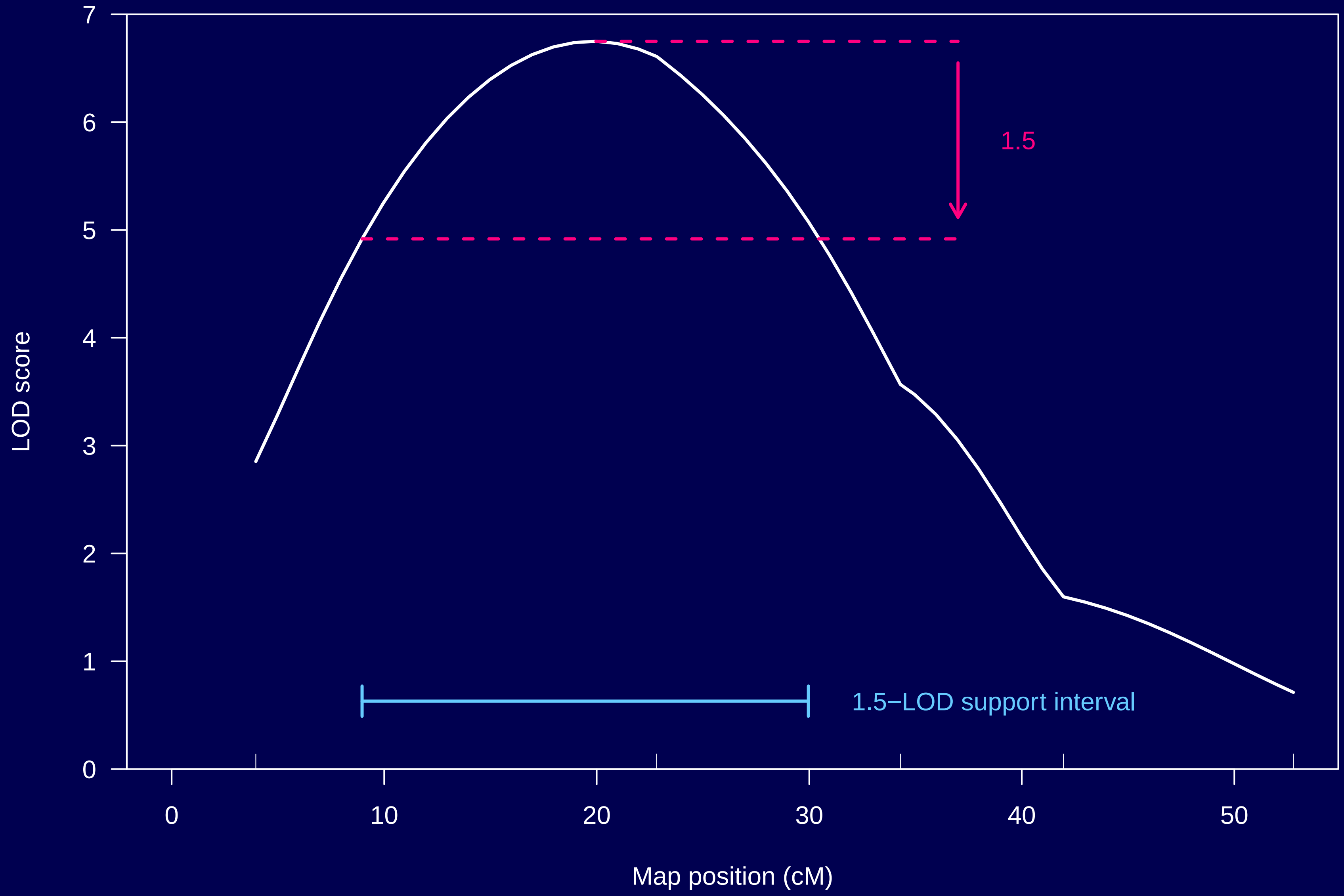
Interactive plot



→ R

- `scanone()` for permutations

LOD support intervals



→ R

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

Haley-Knott regression

A quick approximation to Interval Mapping.

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

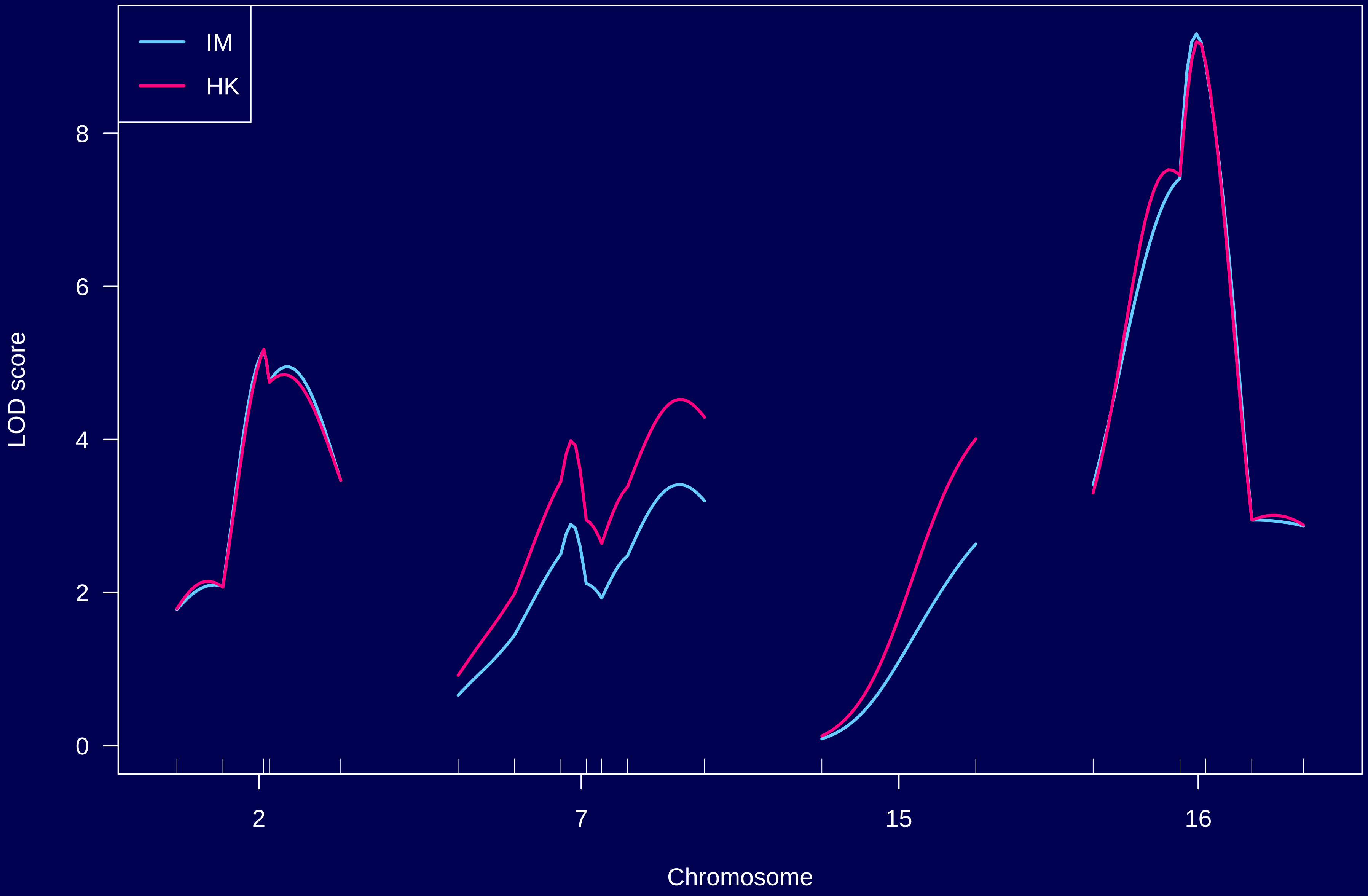
Regress y on p_i, pretending the residual variation is normally distributed (with constant variance).

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

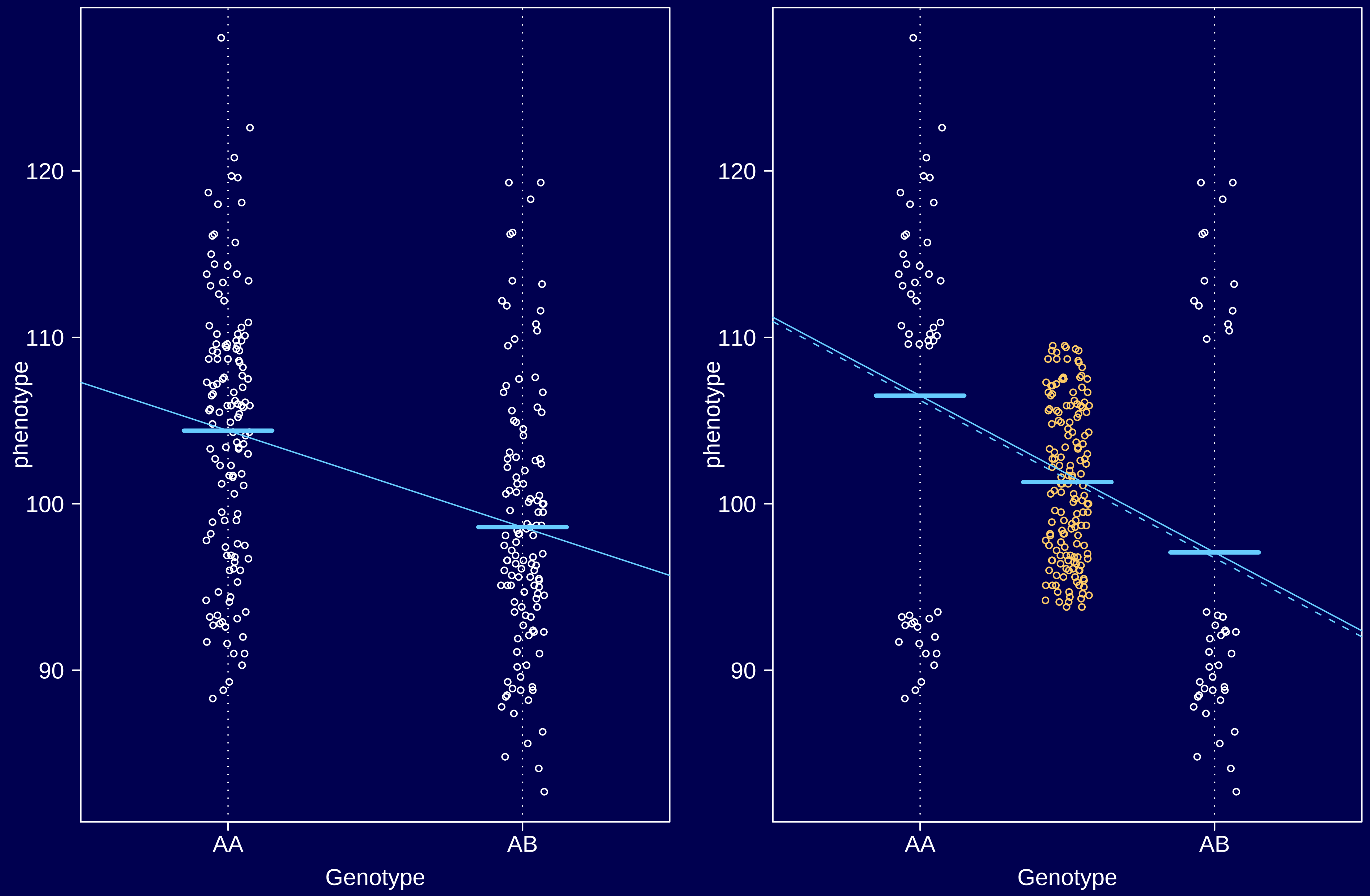
→ R

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

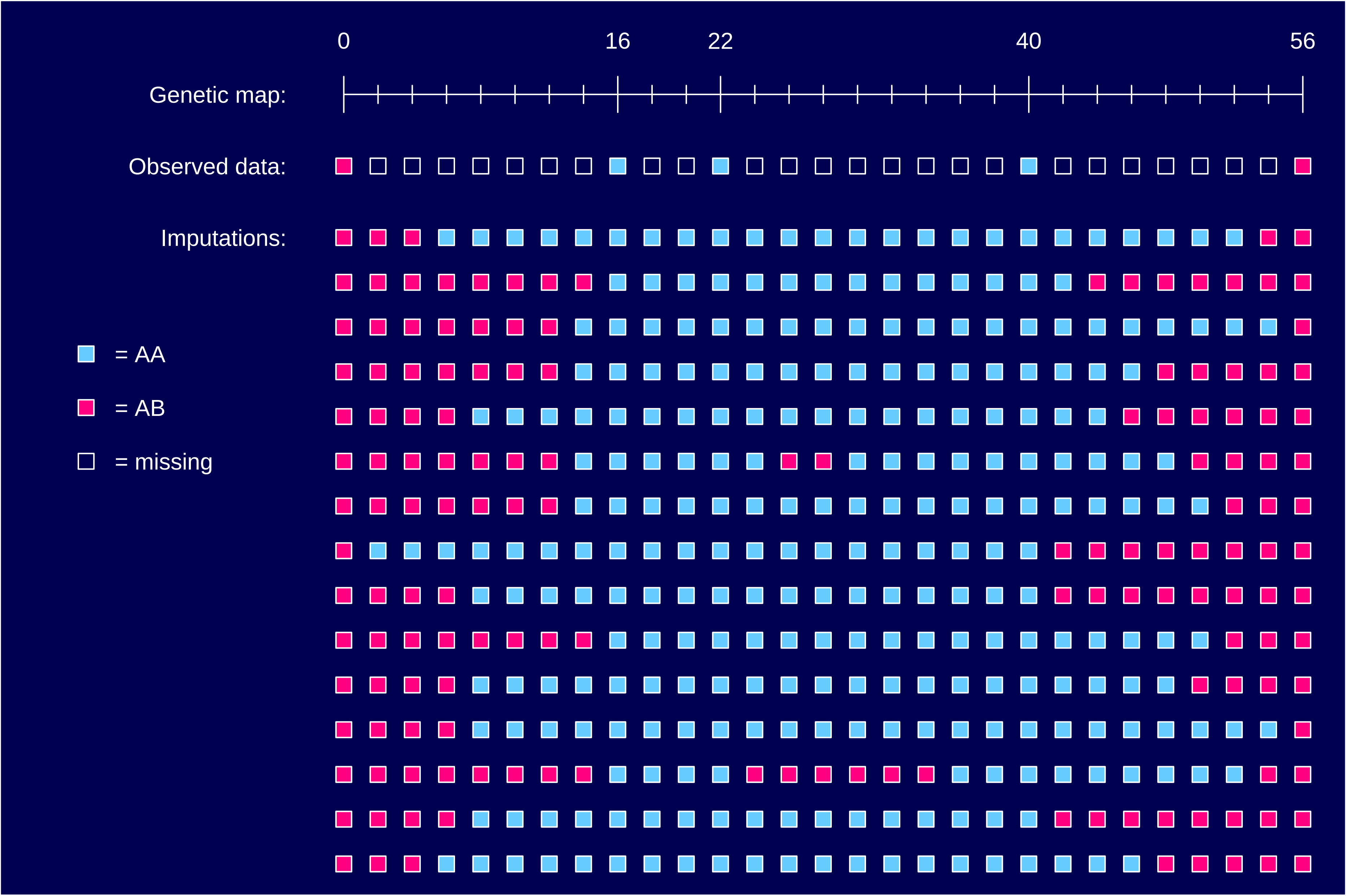
Haley-Knott results



H-K with selective genotyping



Multiple imputation



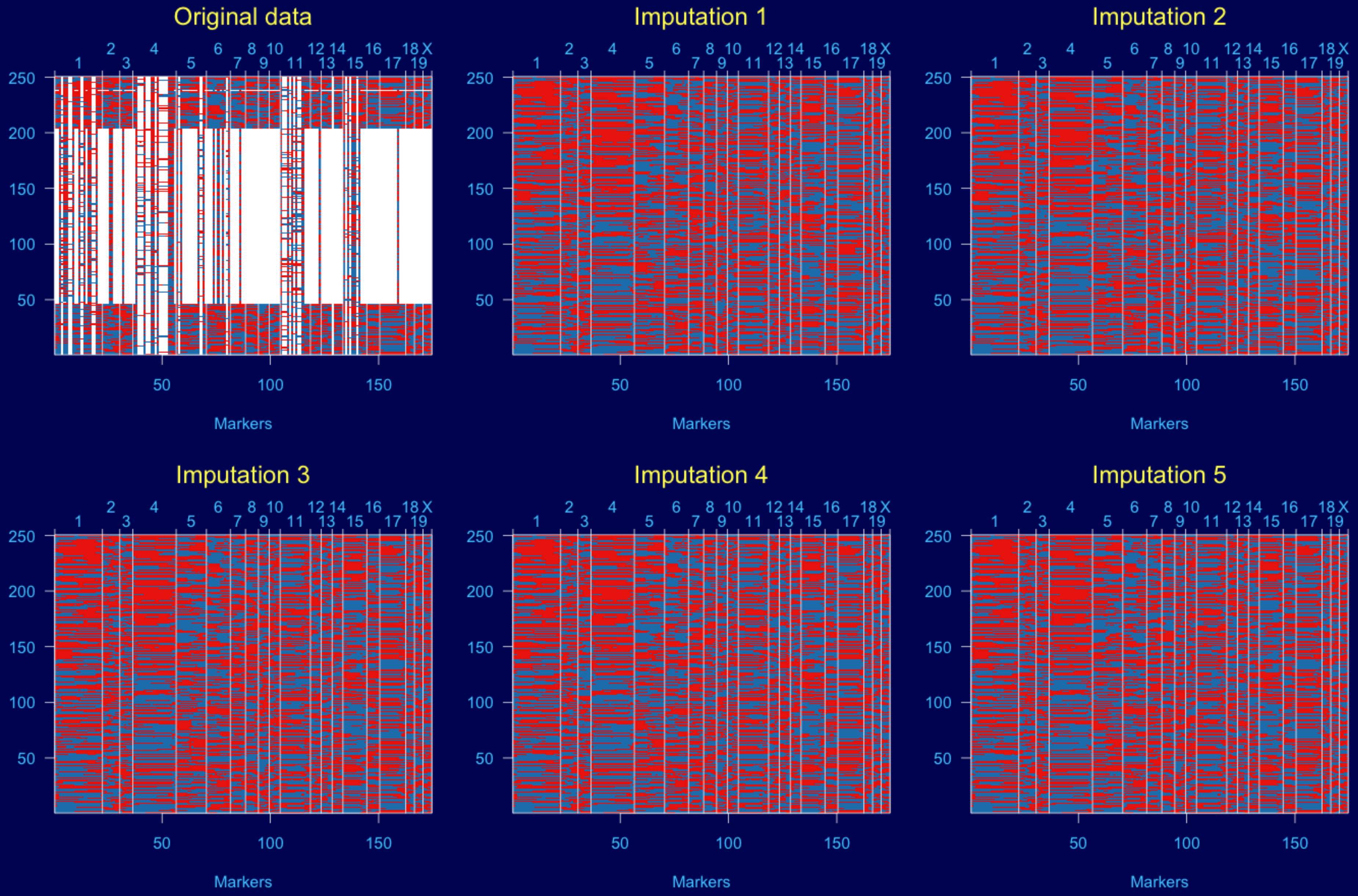
🟦 = AA

🟥 = AB

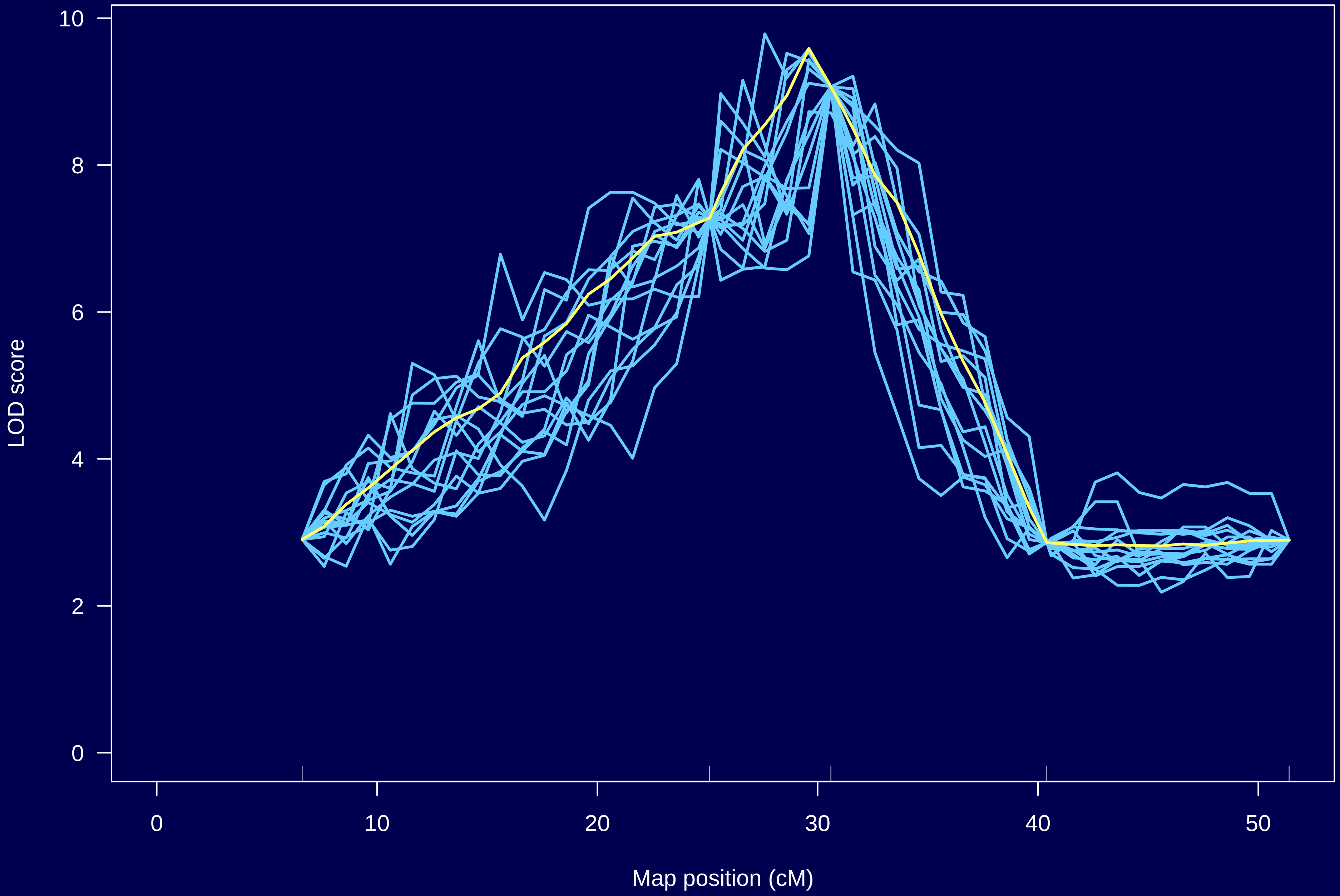
□ = missing

	0								16			22									40								56
Observed data:	🟥	□	□	□	□	□	□	□	🟦	□	□	🟦	□	□	□	□	□	□	□	□	🟦	□	□	□	□	□	□	□	🟥
Imputations:	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥
	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥
	🟥	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	🟥	🟥	🟥	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟥	🟥	🟥

Multiple imputations



Imputation LOD curves



→ R

- `sim.geno()`

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

Summary comparison

Approach	Speed	Extensibility	Stability	Missing data	Parallelization
HK	++	+	+	−	++
EM	+	—	—	+	—
Imputation	—	+	+	+	+

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 two-part model in Broman 2003)

→ R

- `nqrank()`
- `scanone()` with `model="binary"` or `model="np"`

Covariates

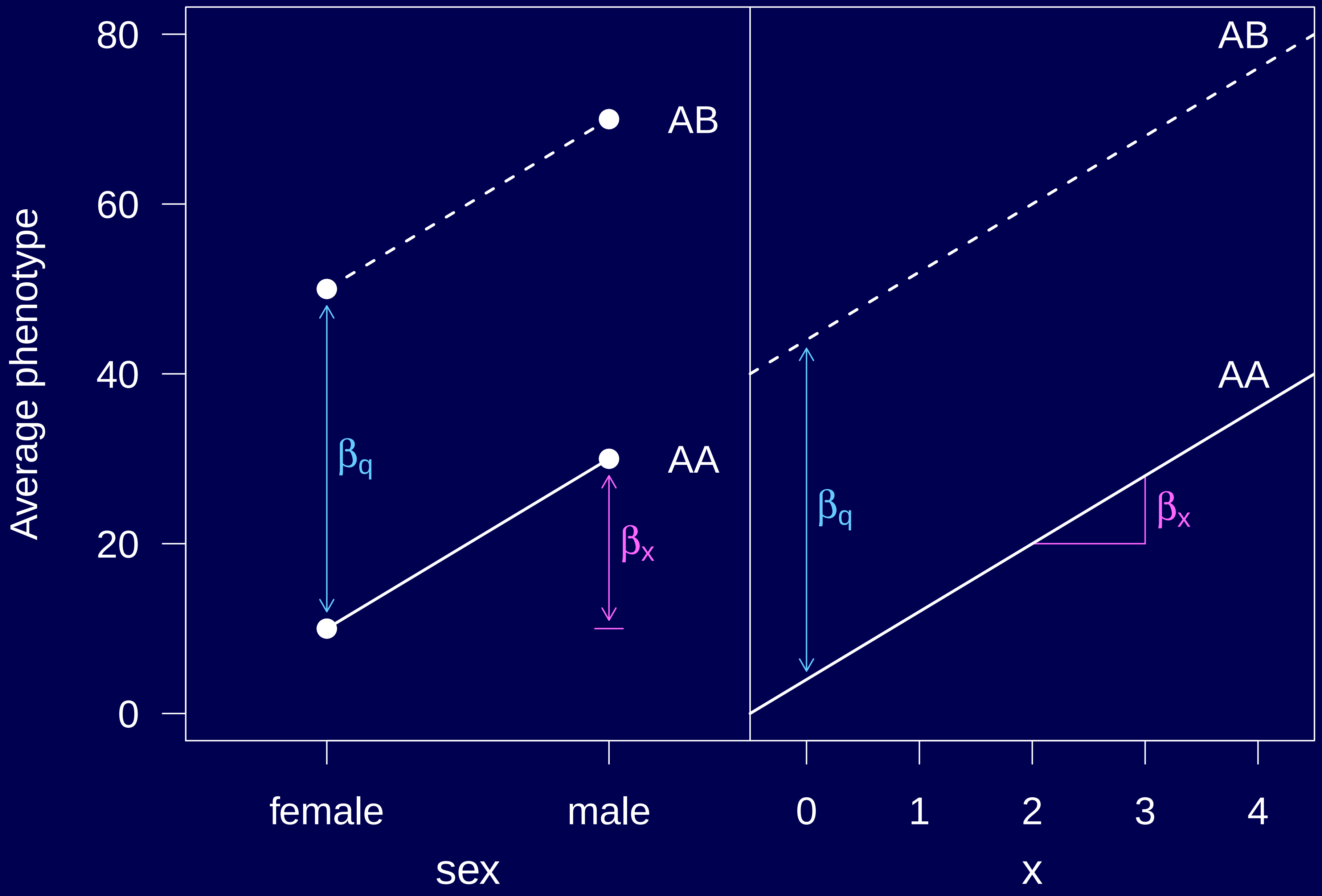
- Examples: treatment, sex, age, weight
- Control residual variation → increase power
- Look for QTL × covariate interactions

Additive covariate

$$\text{H}_0 : y = \mu + \beta_x x + \epsilon$$
$$\text{H}_a : y = \mu + \beta_x x + \beta_q q + \epsilon$$

- If covariate has strong effect on the phenotype, accounting for it can give improved power to detect QTL.
- In permutations, keep phenotype and covariate together
- Use care when the covariate is another phenotype

Additive covariate

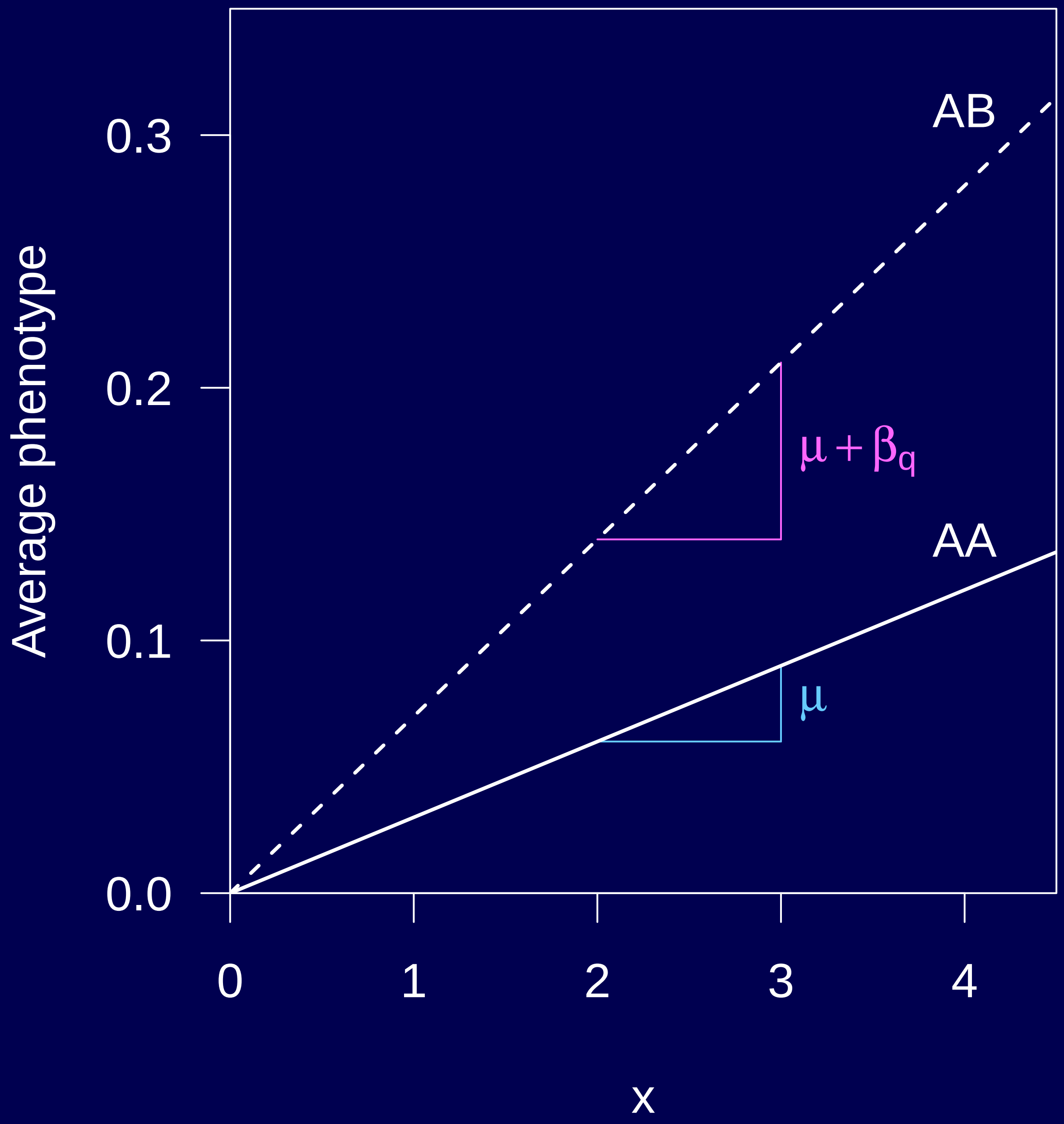


Adjust then scan?

- Consider adjusted phenotype $y' = y/x$
- The QTL model is $(y/x) = \mu + \beta_q q + \epsilon$
- Equivalently

$$y = \begin{cases} \mu\ x + \epsilon' & \text{if } q = 0 \\ (\mu + \beta_q)x + \epsilon' & \text{if } q = 1 \end{cases}$$

Adjust then scan?



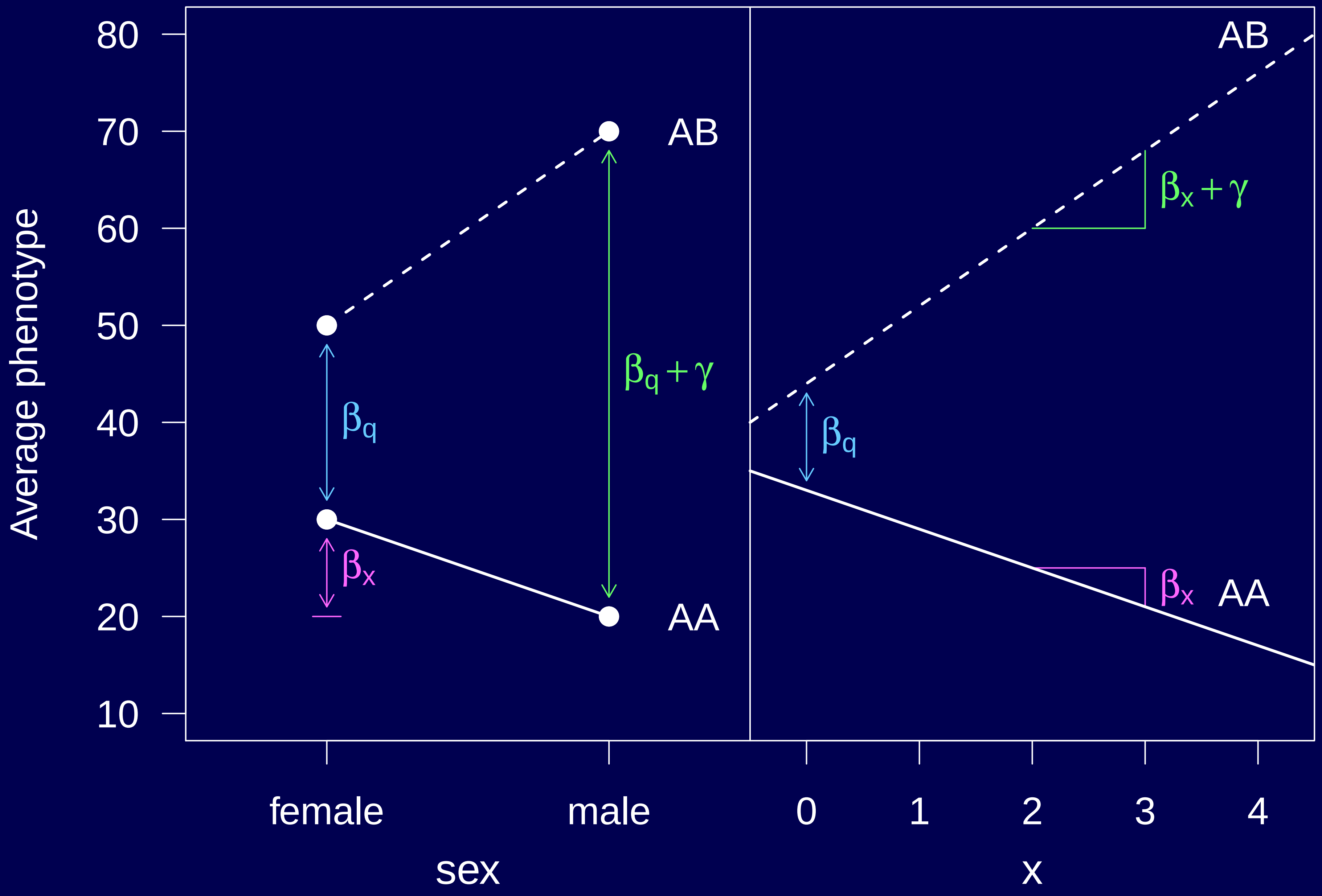
Interactive covariate

$$\begin{aligned} \text{H}_0 : y &= \mu + \beta_x x + \epsilon \\ \text{H}_a : y &= \mu + \beta_x x + \beta_q q + \epsilon \\ \text{H}_i : y &= \mu + \beta_x x + \beta_q q + \gamma x q + \epsilon \end{aligned}$$

Can consider 3 LOD scores:

- LOD$_a$ comparing H$_a$ and H$_0$
- LOD$_f$ comparing H$_i$ and H$_0$
- LOD$_i$ comparing H$_i$ and H$_a$

Interactive covariate

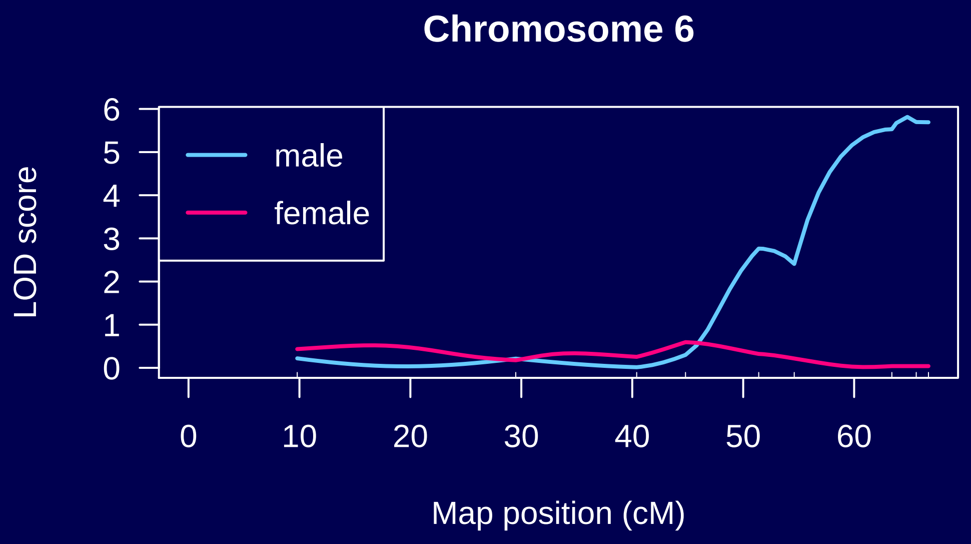


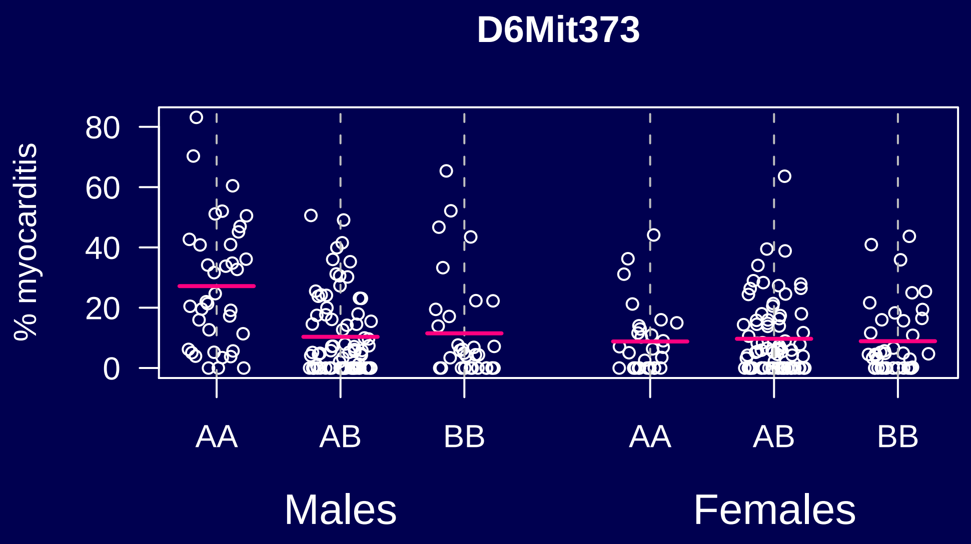
→ R

- `scanone()` with `addcovar` and `intcovar`
- `set.seed()` to do permutations

Split on sex?

- Informative, understandable
- But tempting to falsely conclude "sex-specific QTL"
- Absence of evidence is not *evidence of absence*.
- Use explicit test of QTL × sex interaction

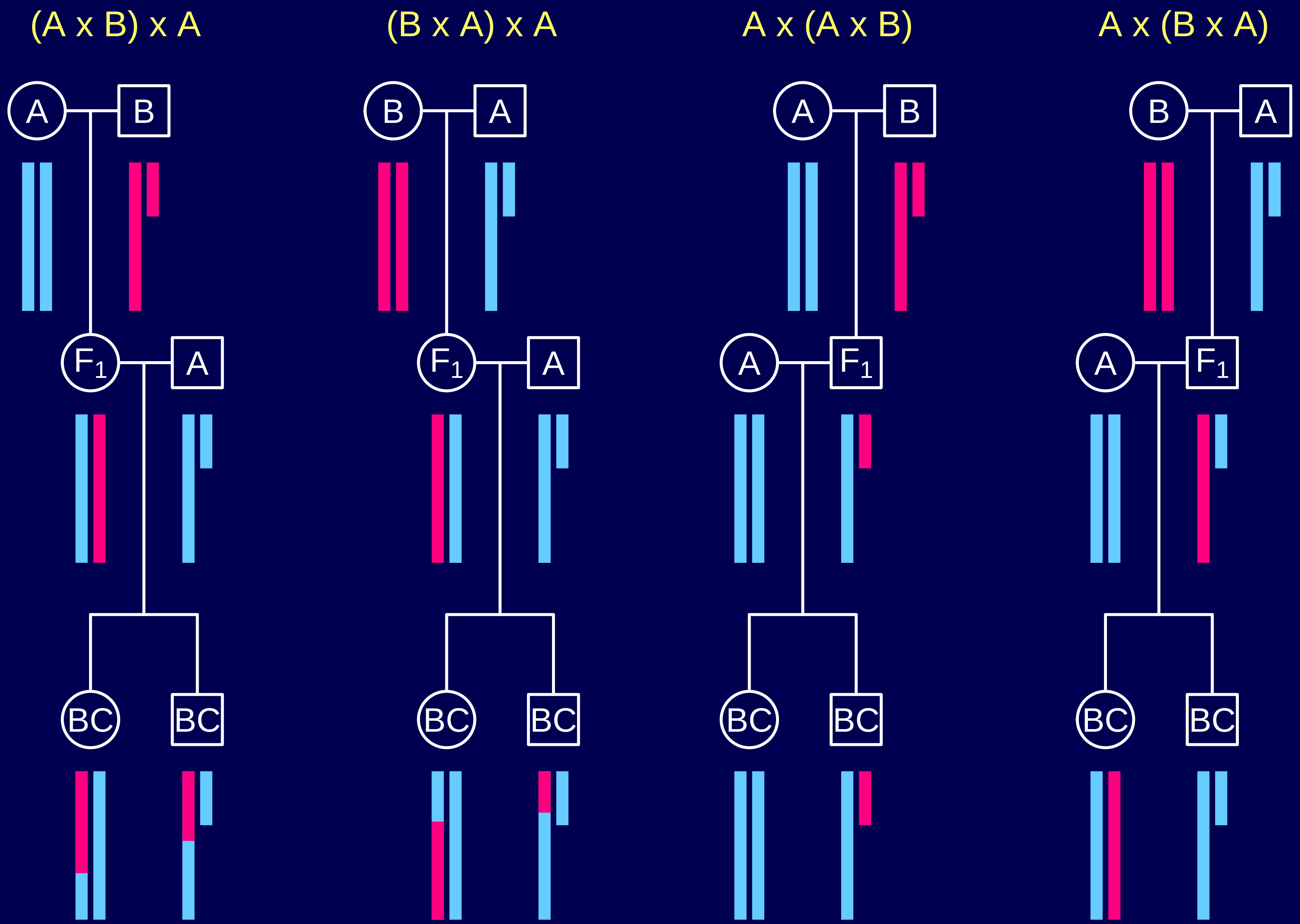




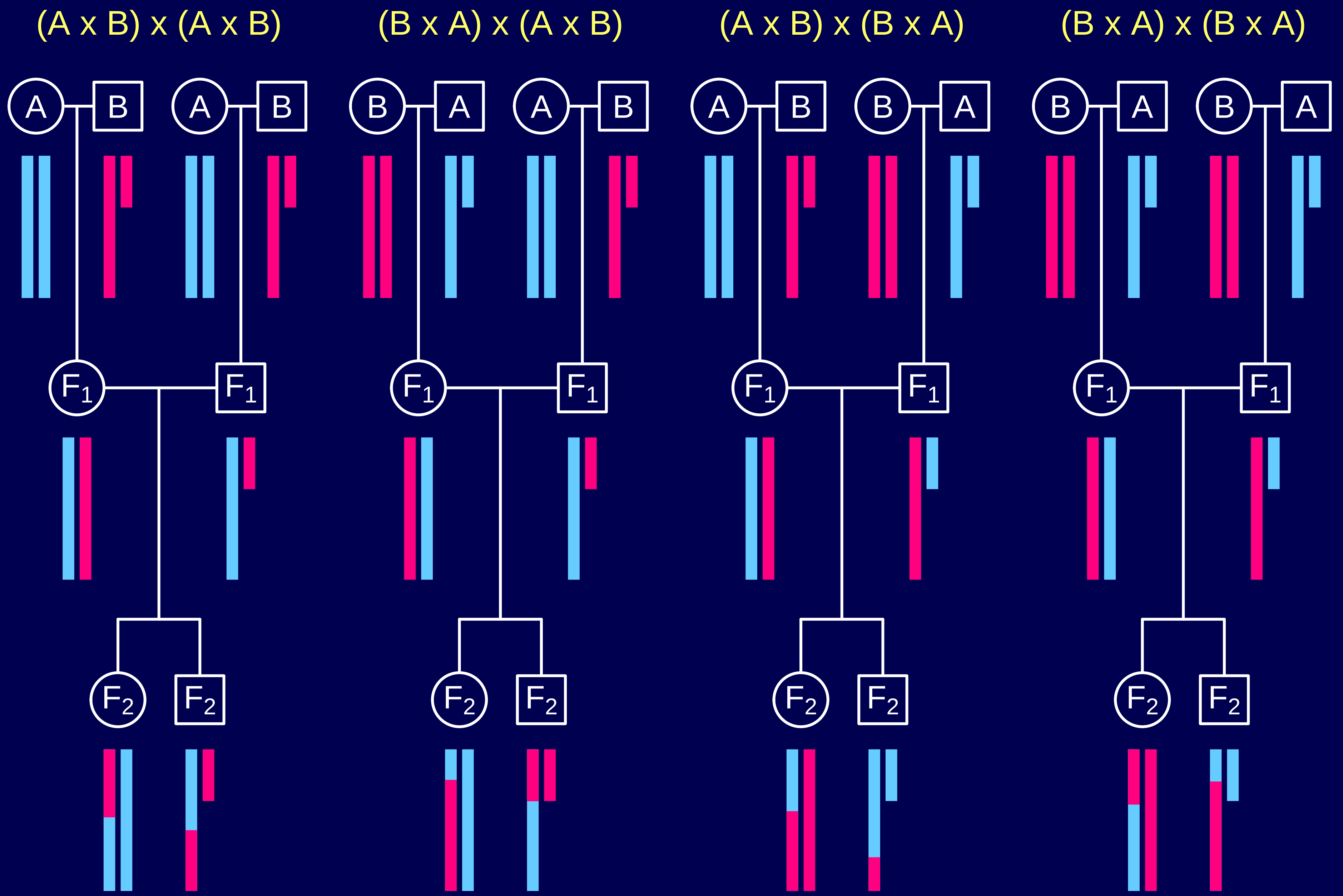
→ R

- `subset()` to split on sex

X chr in backcross



X chr in intercross



Example

Intercross: both dir, both sexes

♀	forward	AA or AB		
♀	reverse		AB or BB	
♂	forward			AY or BY
♂	reverse			AY or BY

→ R

- `scanone()` permutations with `perm.Xsp=TRUE`

Data diagnostics

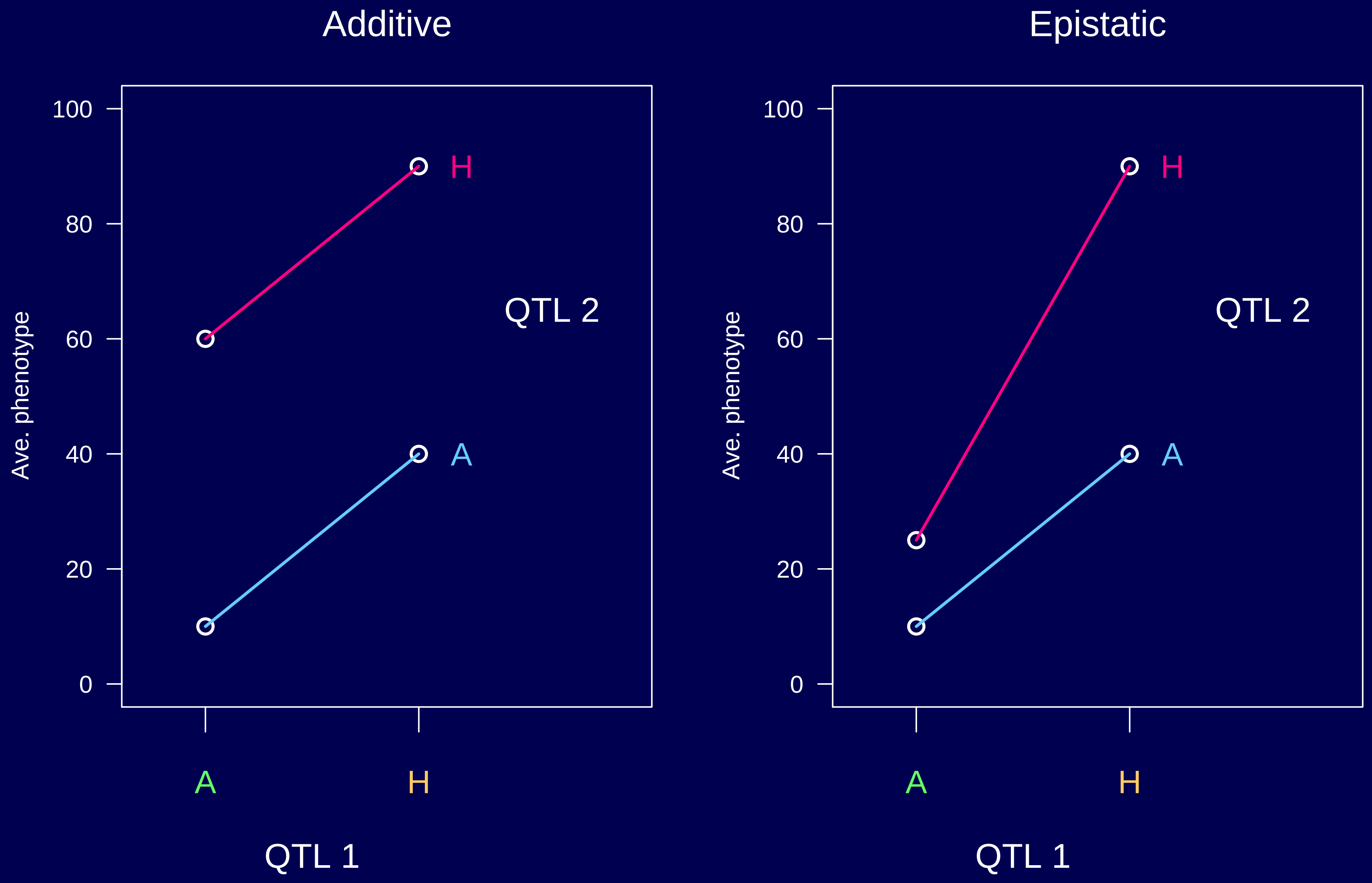
- Plot phenotypes
- Look for sample duplicates
- Look for excessive missing data
- Investigate segregation distortion
- Verify genetic maps/marker positions
- Look for genotyping errors
- Look at counts of crossovers

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

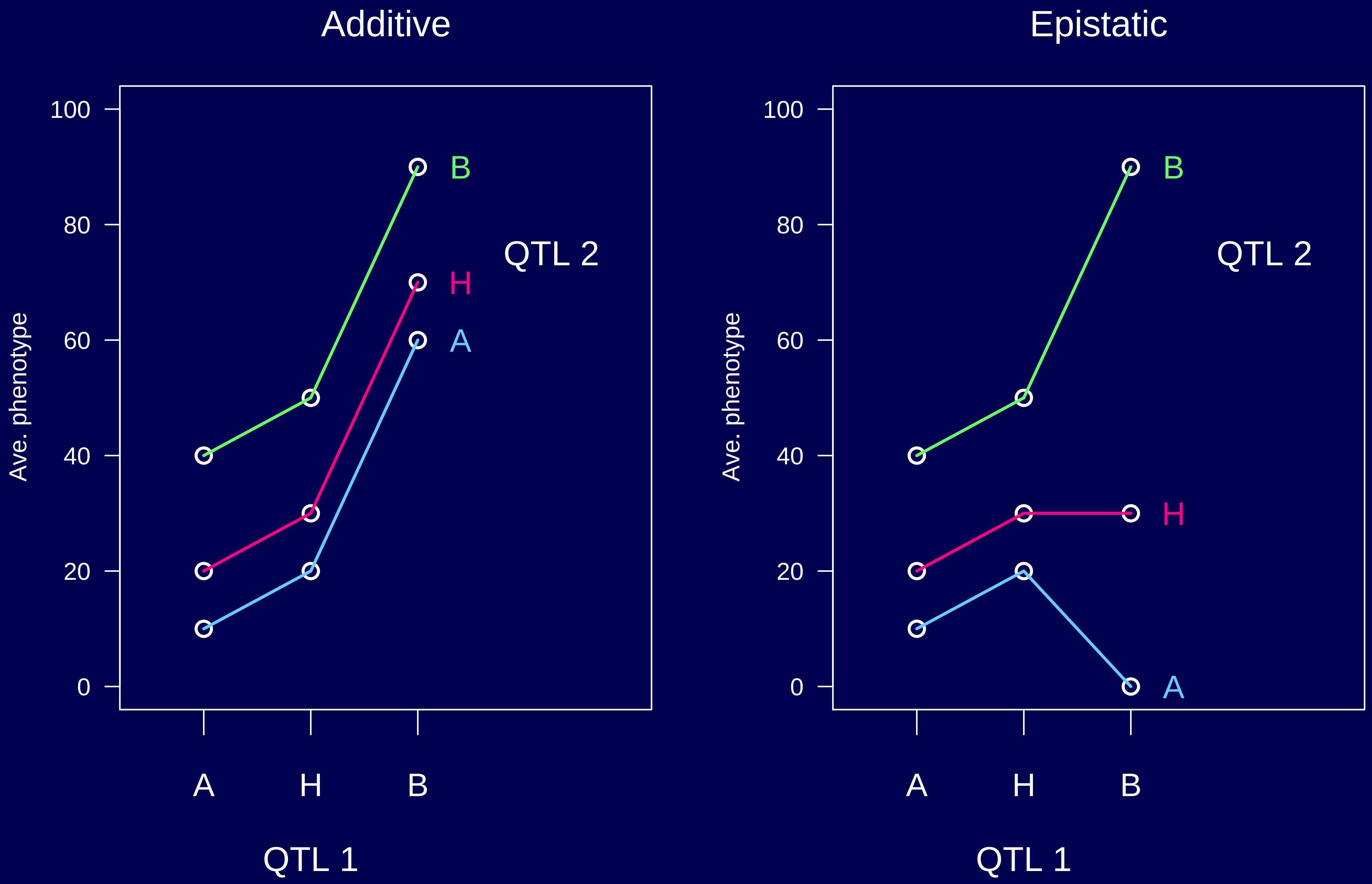
Modeling multiple QTL

- Reduce residual variation ⟶ increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)

Epistasis in BC



Epistasis in F_2



Additive

Epistatic

2-dim, 2-QTL scan

For all pairs of positions, fit the following models:

$$H_f : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \gamma q_1 q_2 + \epsilon$$

$$H_a : y = \mu + \beta_1 q_1 + \beta_2 q_2 + +\epsilon$$

$$H_1 : y = \mu + \beta_1 q_1 + \epsilon$$

$$H_0 : y = \mu + \epsilon$$

$\log_{10}$ likelihoods:

$$l_f(s,t) \qquad l_a(s,t) \qquad l_1(s) \qquad l_0$$

2-dim, 2-QTL scan

LOD scores:

$$\begin{aligned} \mathsf{LOD}_f(s,t) &= l_f(s,t) - l_0 \\ \mathsf{LOD}_a(s,t) &= l_a(s,t) - l_0 \\ \mathsf{LOD}_i(s,t) &= l_f(s,t) - l_a(s,t) \\ \mathsf{LOD}_1(s) &= l_1(s) - l_0 \end{aligned}$$

Summaries

Consider each pair of chromosomes, (j, k),
and let $c(s)$ denote the chromosome for position s.

$$\begin{aligned}
M_f(j,k) &= \max_{c(s)=j, c(t)=k} LOD_f(s,t) \\
M_a(j,k) &= \max_{c(s)=j, c(t)=k} LOD_a(s,t) \\
M_1(j,k) &= \max_{c(s)=j \text{ or } k} LOD_1(s)
\end{aligned}$$

$$\begin{aligned}
M_i(j,k) &= M_f(j,k) - M_a(j,k) \\
M_{fv1}(j,k) &= M_f(j,k) - M_1(j,k) \\
M_{av1}(j,k) &= M_a(j,k) - M_1(j,k)
\end{aligned}$$

→ R

- `scantwo()`
- `iplotScantwo()` in R/qtlcharts

Hypothesis testing?

- In the past, QTL mapping has been regarded as a task of hypothesis testing.

  Is this a QTL?

  Much of the focus has been on adjusting for test multiplicity.

- It is better to view the problem as one of model selection.

  What set of QTL are well supported?

  Is there evidence for QTL-QTL interactions?

  Model = a defined set of QTL and QTL-QTL interactions
  (and possibly covariates and QTL-covariate interactions).

Model selection

- Class of models
  - Additive models
  - + pairwise interactions
  - + higher-order interactions
  - Regression trees

- Model fit
  - Maximum likelihood
  - Haley-Knott regression
  - extended Haley-Knott
  - Multiple imputation
  - MCMC

- Model comparison
  - Estimated prediction error
  - AIC, BIC, penalized likelihood
  - Bayes

- Model search
  - Forward selection
  - Backward elimination
  - Stepwise selection
  - Randomized algorithms

Target

- Selection of a model includes two types of errors:
  - Miss important terms (QTLs or interactions)
  - Include extraneous terms

- Unlike in hypothesis testing, we can make both errors at the same time.

- Identify as many correct terms as possible, while controlling the rate of inclusion of extraneous terms.

What is special here?

- Goal: identify the major players
- A continuum of ordinal-valued covariates (the genetic loci)
- Association among the covariates
  - Loci on different chromosomes are independent
  - Along chromosome, a very simple (and known) correlation structure

Exploratory methods

- Condition on a large-effect QTL
  - Reduce residual variation
  - Conditional LOD score:

$$\text{LOD}(q_2 \mid q_1) = \log_{10}\left\{\frac{\Pr(\text{data} \mid q_1, q_2)}{\Pr(\text{data} \mid q_1)}\right\}$$

- Piece together the putative QTL from the 1d and 2d scans
  - Omit loci that no longer look interesting (drop-one-at-a-time analysis)
  - Study potential interactions among the identified loci
  - Scan for additional loci (perhaps allowing interactions), conditional on these

→ R

- `scanone()` with marker as additive covariate

- `makeqtl()`, `fitqtl()`, `addqtl()`, `refineqtl()`

Automation

- Assistance to non-specialists
- Understanding performance
- Many phenotypes

Additive QTL

$y = \mu + \sum \beta_j q_j + \epsilon$ which $\beta_j \neq 0$?

$\text{pLOD}(\gamma) = \text{LOD}(\gamma) - T\,|\gamma|$

Additive QTL

$\mathsf{y} = \mu + \sum \beta_{\mathsf{j}}\, \mathsf{q}_{\mathsf{j}} + \epsilon$ which $\beta_{\mathsf{j}} \neq 0$?

$\mathsf{pLOD}(\gamma) = \mathsf{LOD}(\gamma) - \mathsf{T}\,|\gamma|$

0 vs 1 QTL: $\mathsf{pLOD}(\emptyset) = 0$

$\mathsf{pLOD}(\{\lambda\}) = \mathsf{LOD}(\lambda) - \mathsf{T}$

Additive QTL

$$\mathrm{y} = \mu + \sum \beta_{\mathrm{j}}\, \mathrm{q}_{\mathrm{j}} + \epsilon \qquad \text{which } \beta_{\mathrm{j}} \neq 0?$$

$$\mathrm{pLOD}(\gamma) = \mathrm{LOD}(\gamma) - \mathrm{T}\,|\gamma|$$

For the mouse genome:

T = 2.69 (BC) or 3.52 (F_2)

→ R

- `stepwiseqtl()`
- `plotLodProfile()`

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.

- 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.

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.