Quantitative Trait Loci

Brian S. Yandell, UW-Madison

January 2017

evolution of QTL models

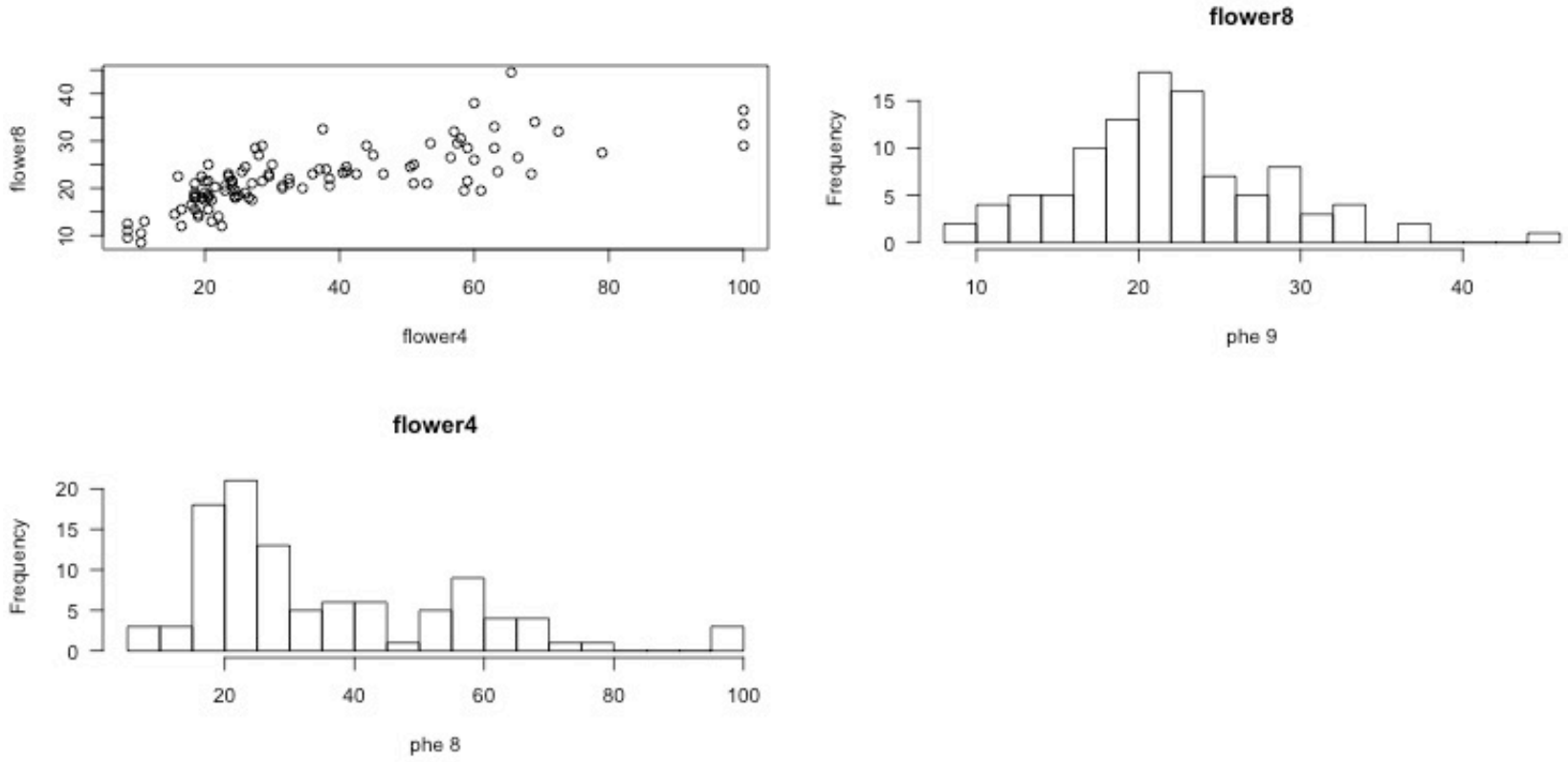
original ideas focused on rare & costly markers
models & methods refined as technology advanced

- **single marker regression**
- **QTL (quantitative trait loci)**
  - **single locus models: interval mapping for QTL**
  - QTL model search: QTLs & epistasis
- GWA (genome-wide association mapping)
  - adjust for population structure
  - capture "missing heritability"
  - genome-wide selection

strategy for QTL mapping

- Want to figure out what is going on
  - preliminary search: find important story
  - need strategies to uncover patterns
- Want to tell story in publication
- How to accomplish QTL mapping goal
  - organic search for patterns
  - organize methods as you go
  - document steps (so you can redo)

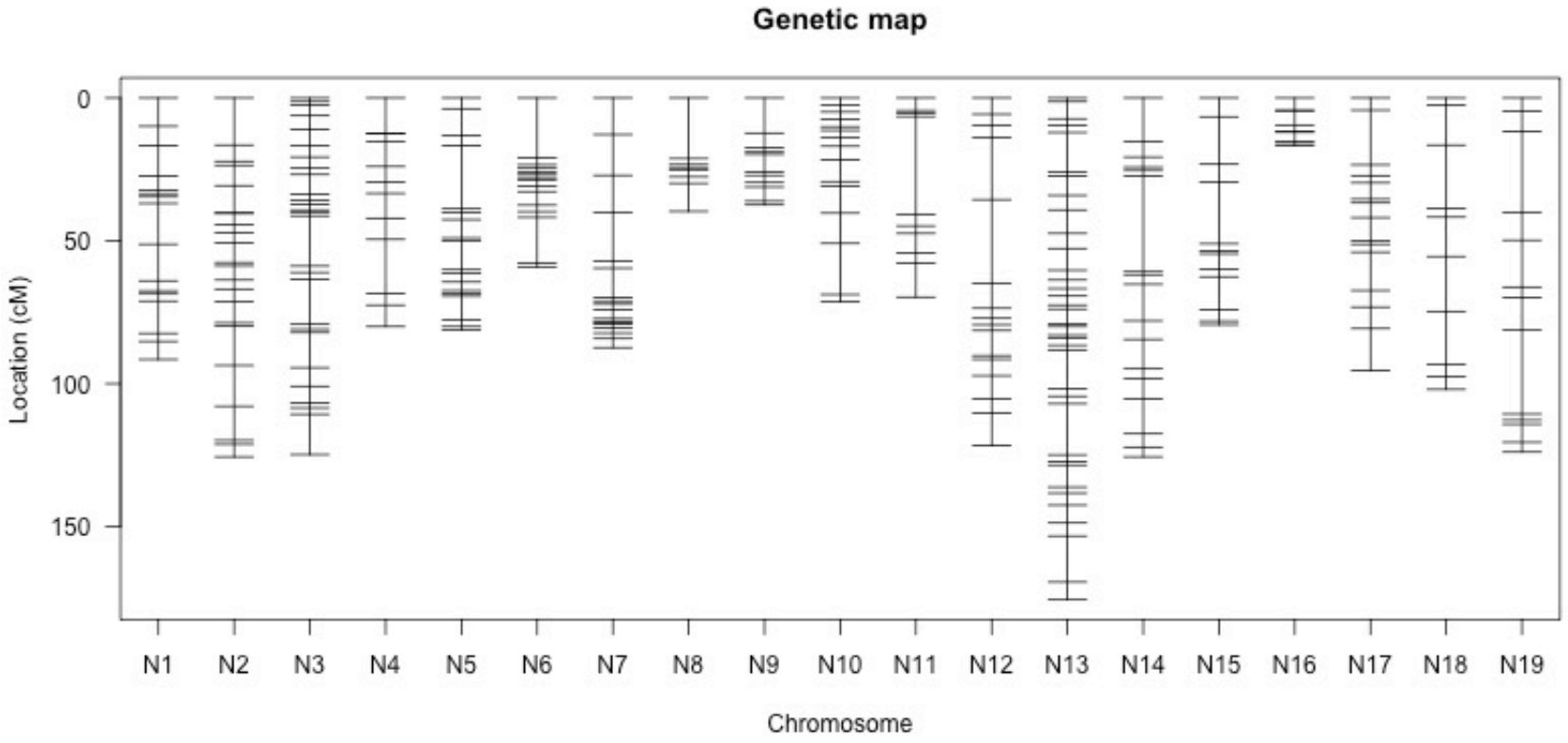
phenotype data: flowering time



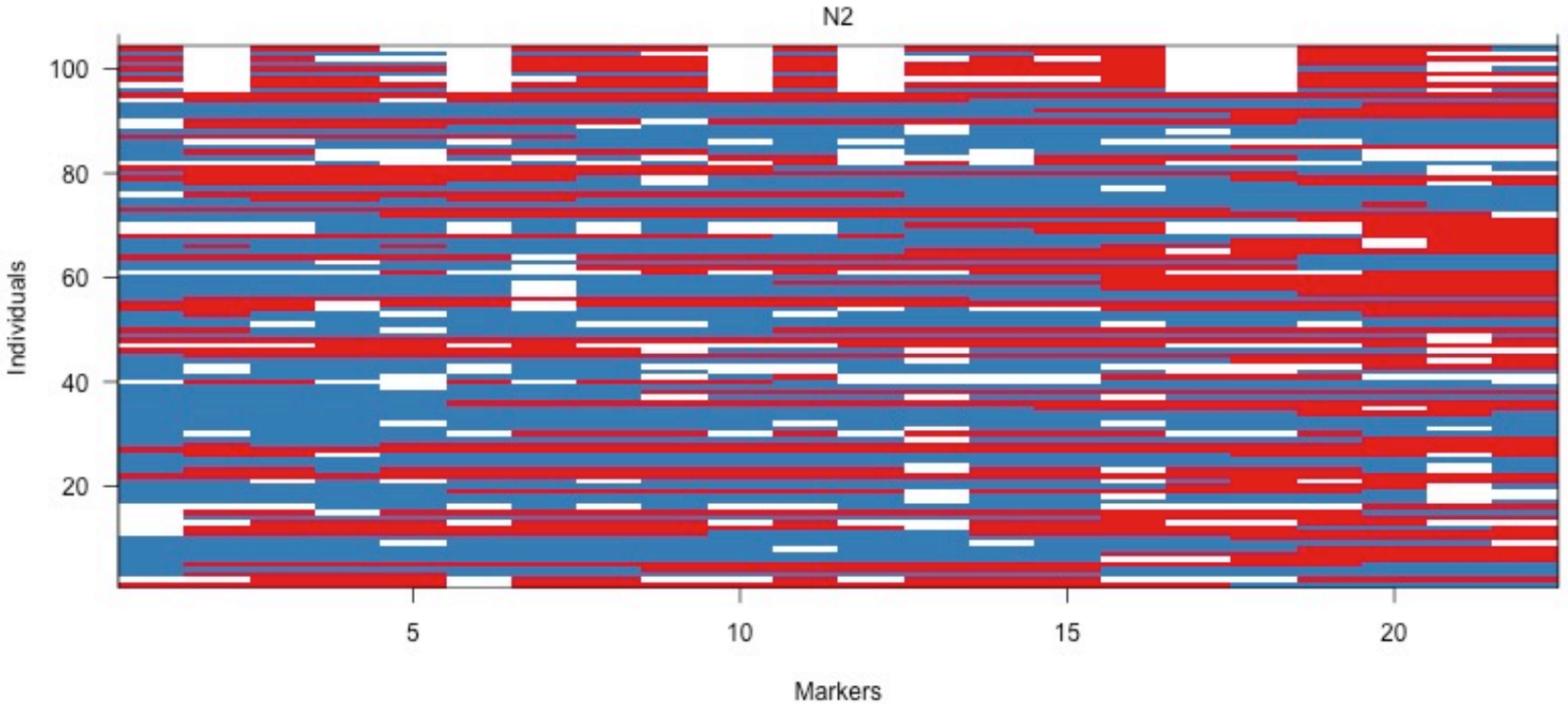
Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) Genetics

genotype data

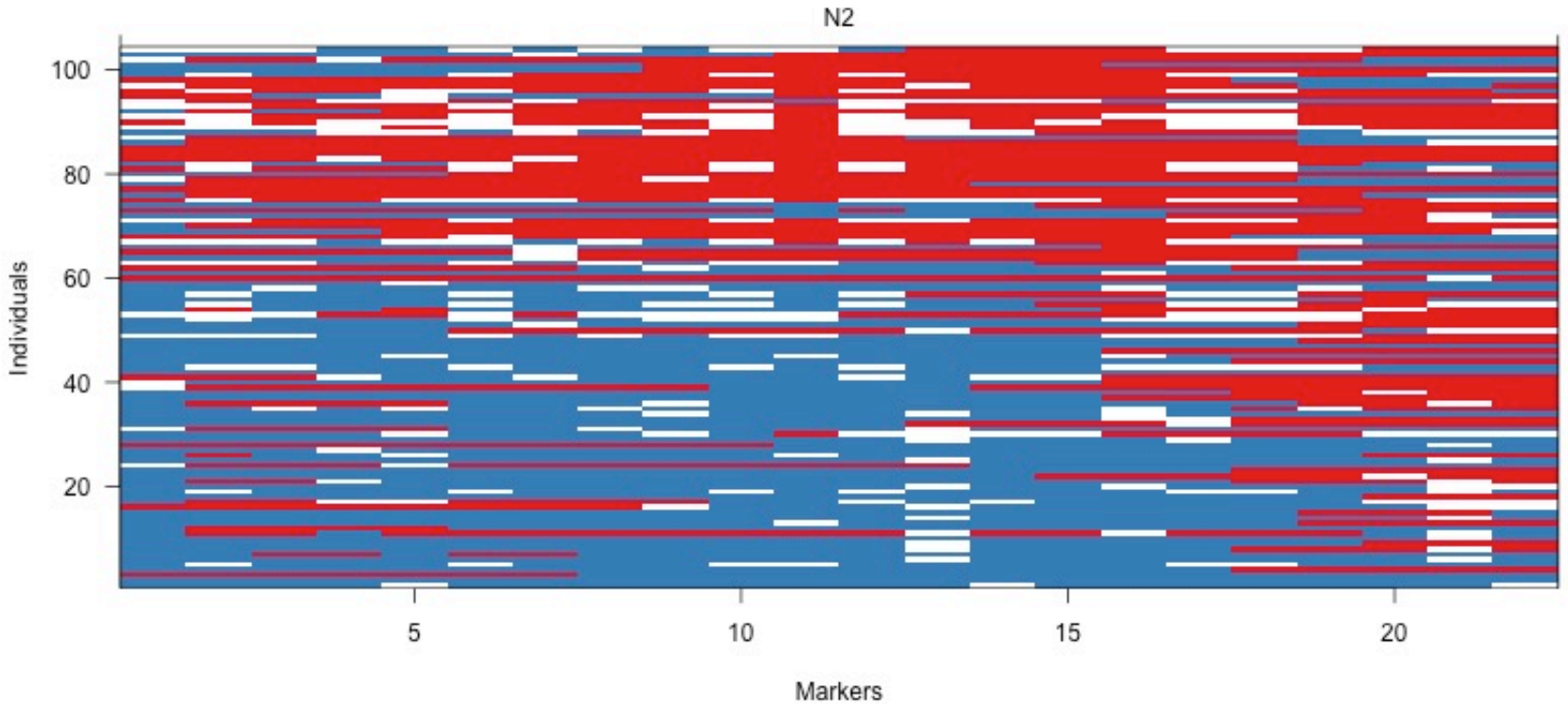
Genetic map for Osborn's *Brassica napus* study



genotypes on chr N2



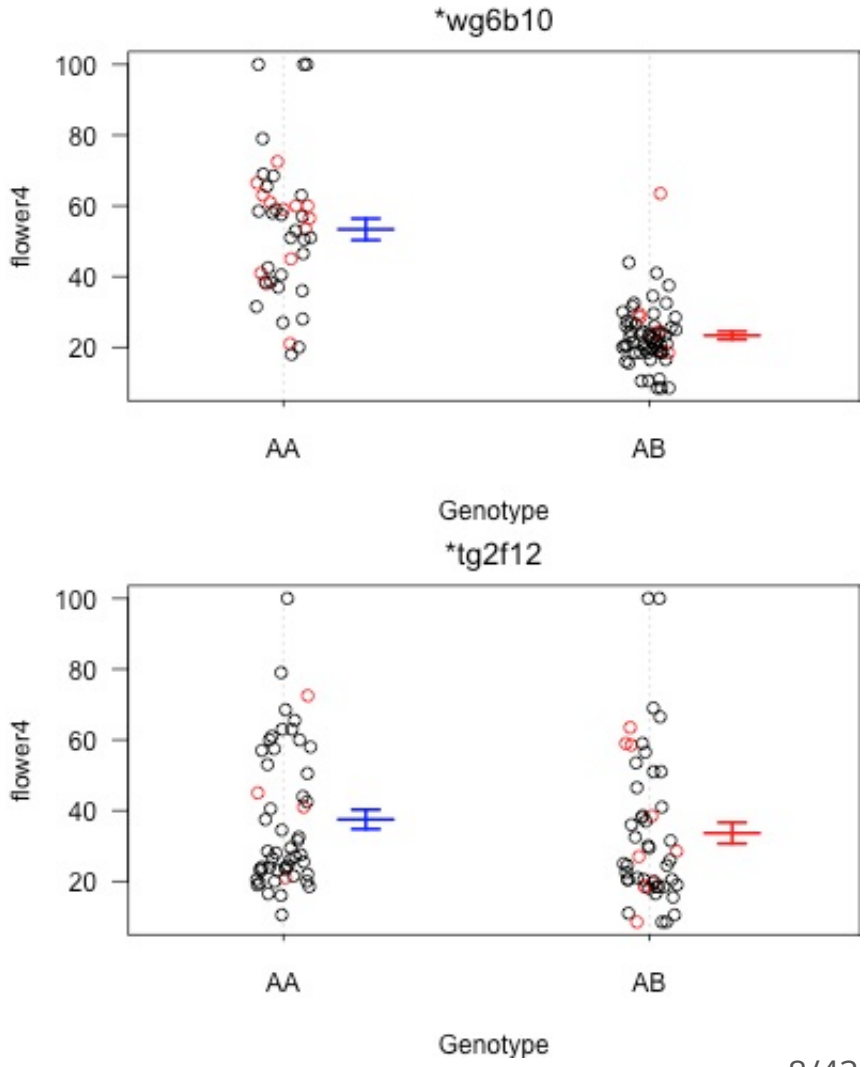
genotypes reordered by `flower4`



marker regression (BC or DH)

- Also known as ANOVA
- Split sample into groups
  - by genotype at marker
  - red = missing genotype
- Do a t-test or ANOVA
- Repeat for each marker

Soller *et al.* (1976)



marker regression model

$$y = \mu_m + e$$

- y = phenotypic trait
- m = marker genotype (0,1)
- μ_m = mean for genotype m
- e = error = unexplained variation

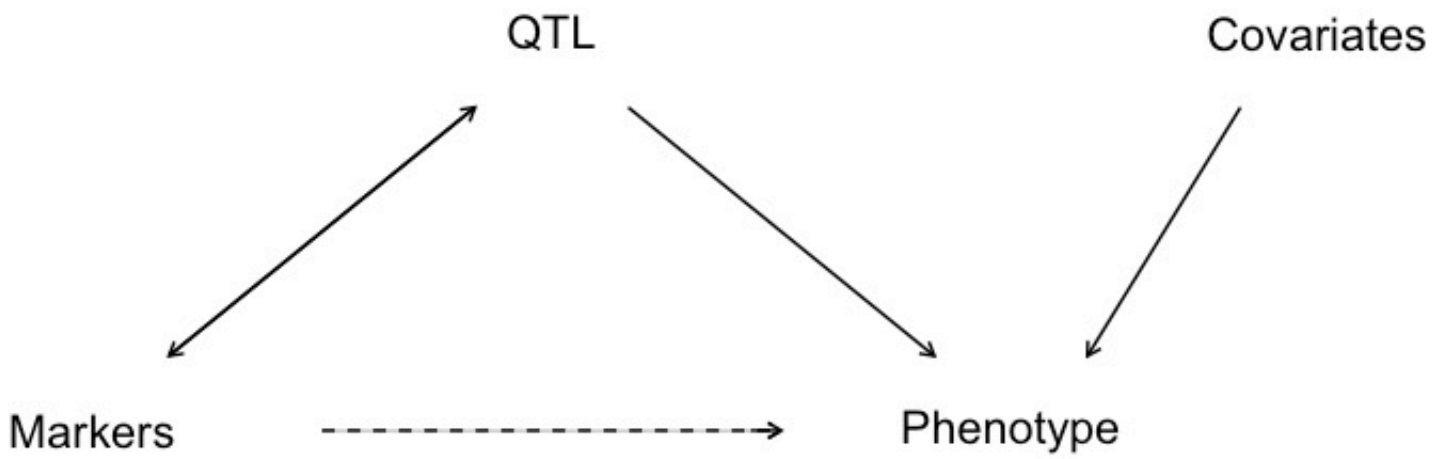
Marker regression:

- fit model for each marker across genome
- pick most significant marker

pros & cons of marker regression

- Advantages
  - simple; no need for genetic map
  - easy to add covariates
  - easily extended to more complex models
  - ignores marker position on genome
- Disadvantages
  - excludse individuals with missing genotype data
  - imperfect information about QTL location
  - suffers in low density scans
  - only considers one QTL at a time

statistical structure



- missing data problem: Markers $\longleftrightarrow$ QTL
- model selection problem: QTL, covariates $\longrightarrow$ phenotype

interval mapping (IM)

- Assume a single QTL model.
- posit each genome position λ, one at a time, as putative QTL
  - q = genotypes at locus λ

$$\texttt{pr}(y|q) : y = \mu_q + e$$

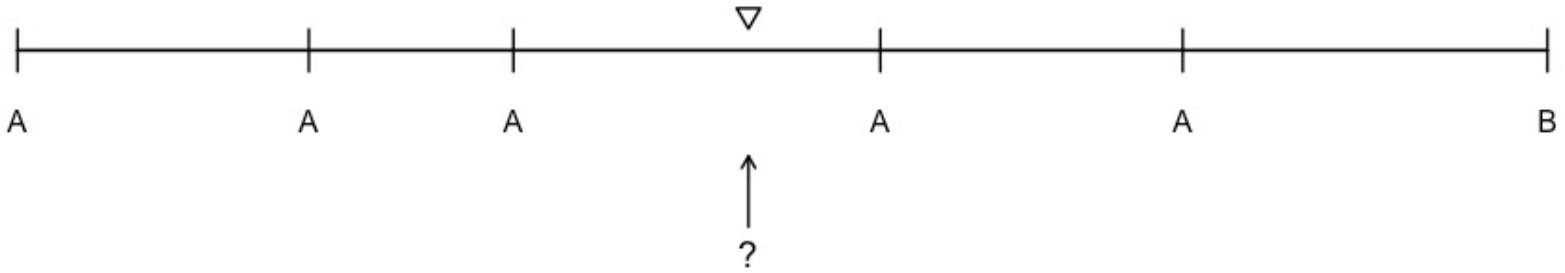
- mixing proportions over flanking markers

$$\texttt{pr}(q|m) : \texttt{table of proportions}$$

- model is mixture over possible QTL genotypes q
- mixture of normals

Lander & Botstein (1989) Genetics

genotype probabilities



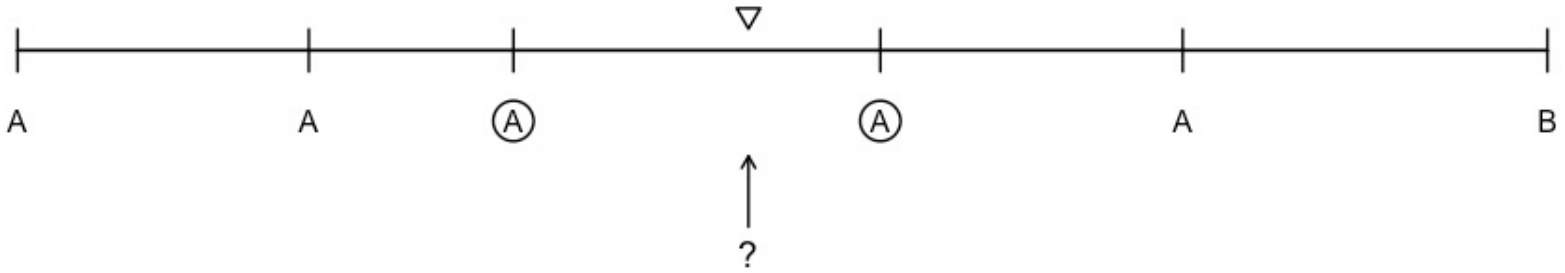
Calculate $pr(q|m)$ assuming

- no crossover interference
- no genotyping errors

Or use the hidden Markov model (HMM)} technology

- to allow for genotyping errors
- to incorporate dominant markers

genotype probabilities



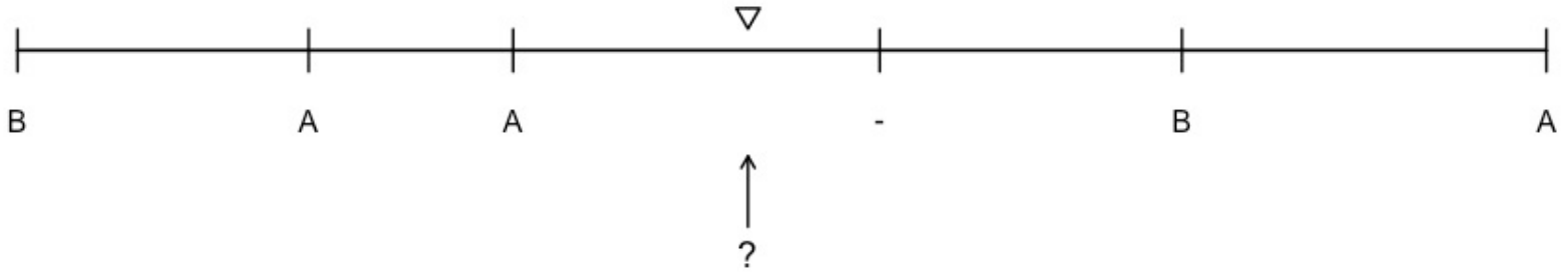
Calculate $pr(q|m)$ assuming

- no crossover interference
- no genotyping errors

Or use the hidden Markov model (HMM)} technology

- to allow for genotyping errors
- to incorporate dominant markers

genotype probabilities



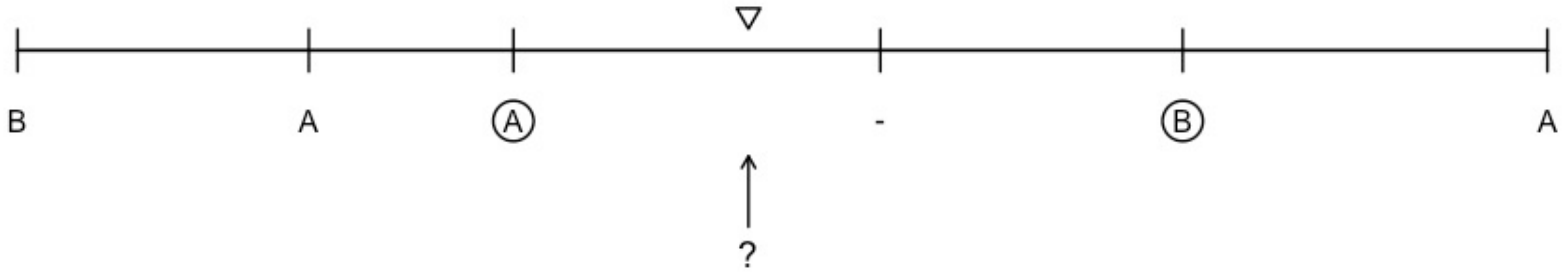
Calculate $pr(q|m)$ assuming

- no crossover interference
- no genotyping errors

Or use the hidden Markov model (HMM)} technology

- to allow for genotyping errors
- to incorporate dominant markers

genotype probabilities



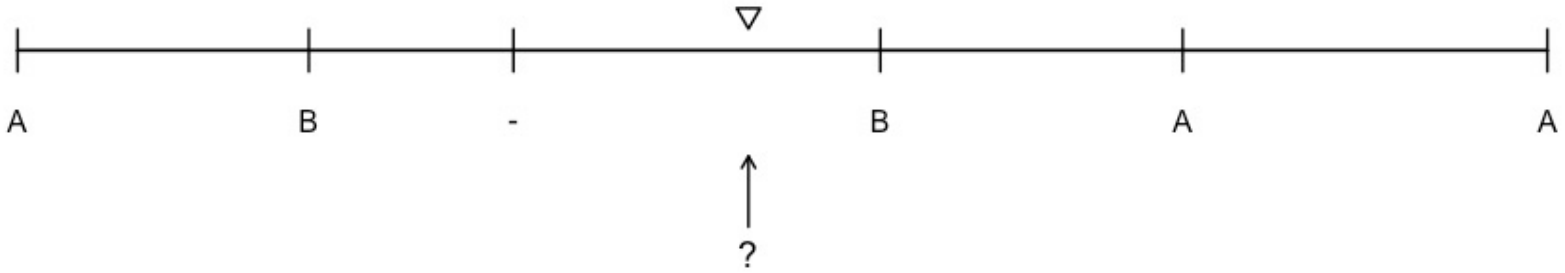
Calculate $pr(q|m)$ assuming

- no crossover interference
- no genotyping errors

Or use the hidden Markov model (HMM)} technology

- to allow for genotyping errors
- to incorporate dominant markers

genotype probabilities



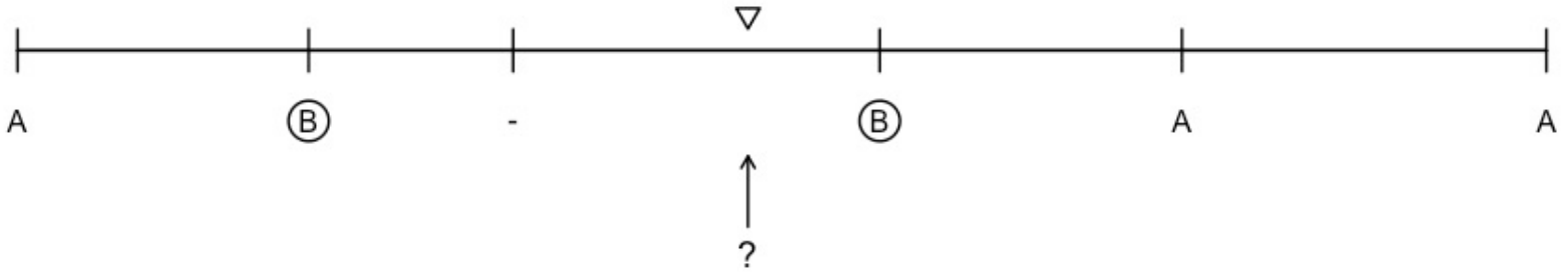
Calculate $pr(q|m)$ assuming

- no crossover interference
- no genotyping errors

Or use the hidden Markov model (HMM)} technology

- to allow for genotyping errors
- to incorporate dominant markers

genotype probabilities



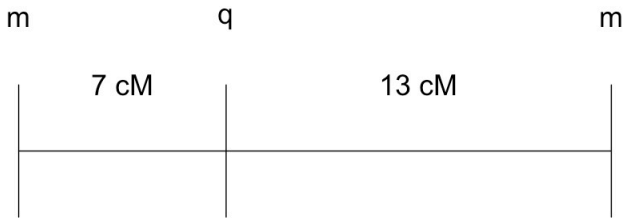
Calculate $pr(q|m)$ assuming

- no crossover interference
- no genotyping errors

Or use the hidden Markov model (HMM)} technology

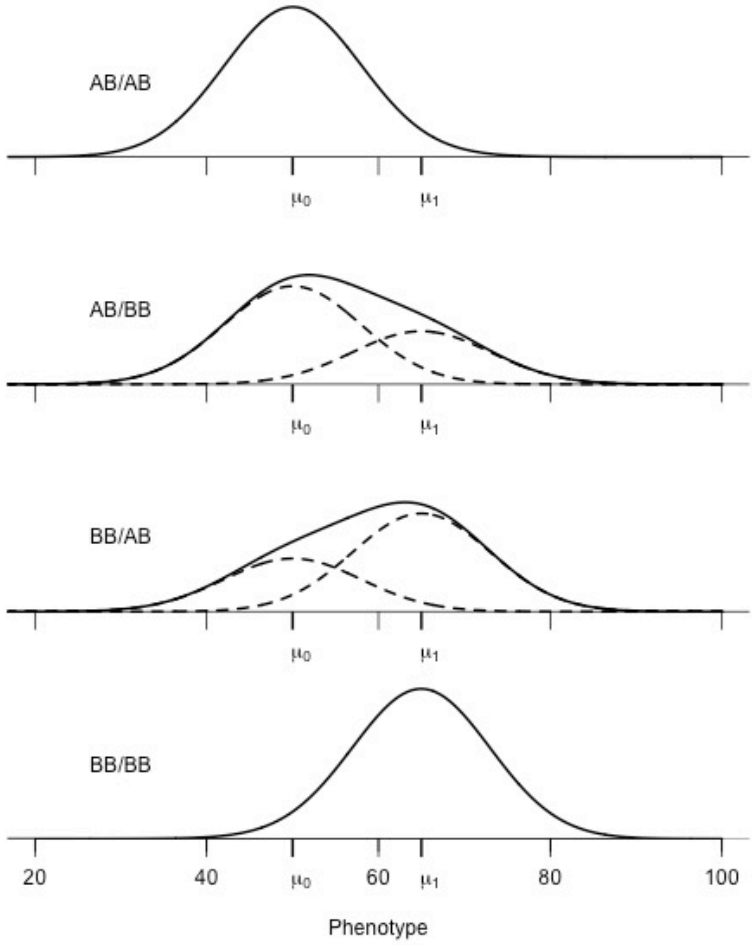
- to allow for genotyping errors
- to incorporate dominant markers

phenotype given unknown genotype



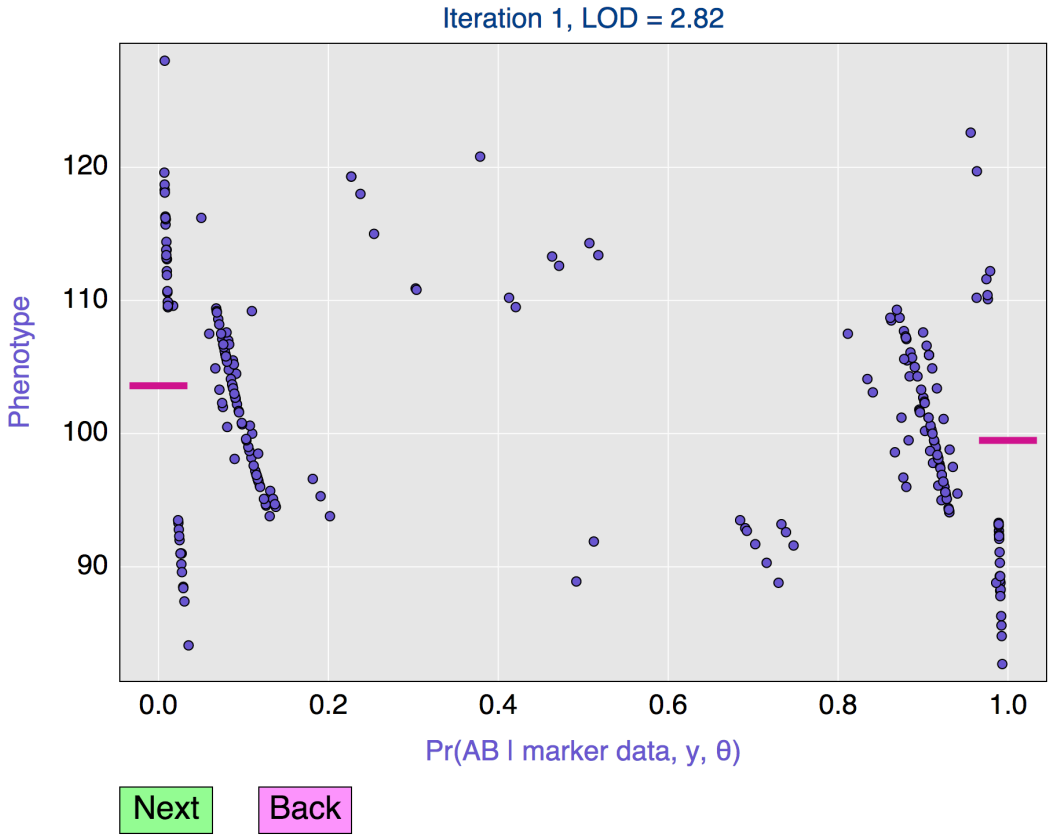
$$\mathrm{pr}(y|m) = \sum \mathrm{pr}(y|q)\mathrm{pr}(q|m)$$

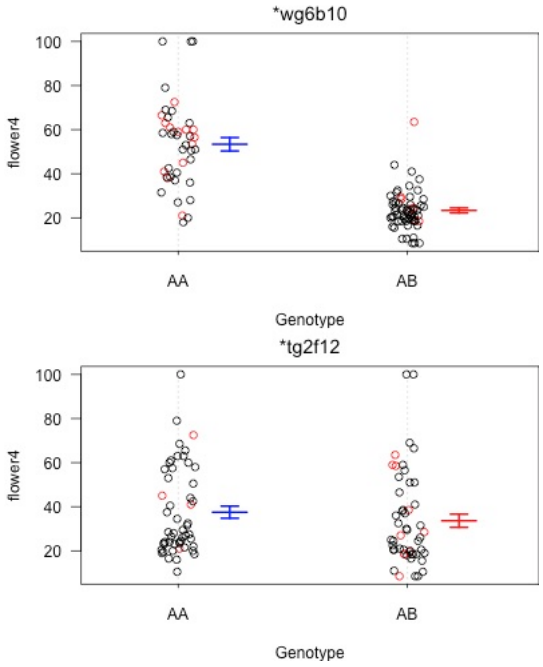
- 2 markers separated by 20 cM
  - QTL closer to left marker
- phenotype distribution
  - given marker genotypes
- mixture components
  - dashed curves



interval mapping idea

think marker regression with fuzzy groups





interval mapping (IM) details

QTL genotype given markers: $\mathtt{pr}(q|m)$

phenotype given QTL: $\mathtt{pr}(y|q) = \mathrm{N}(y|\mu_q, \sigma^2)$ (normal density)

$$\mathtt{pr}(y|m) = \sum_q \mathtt{pr}(y|q)\mathtt{pr}(q|m)$$

log likelihood over individuals:

$$l(\mu_0, \mu_1, \sigma) = \sum_i \log \mathtt{pr}(y_i|m_i)$$

find $\hat{\mu}_0$, $\hat{\mu}_1$, $\hat{\sigma}$ to maximize $l(\mu_0, \mu_1, \sigma)$ (MLEs)

EM algorithm (Dempster et al. 1977)

E step: (pseudo)weights for individual i, QTL genotype q

$$w_{iq} = \texttt{pr}(q|m_i, y_i, \hat{\mu}, \hat{\sigma}) = c_i * \texttt{pr}(q|m_i)\mathrm{N}(y_i|\hat{\mu}_q, \hat{\sigma})$$

c_i set so that $\sum_q w_{iq} = 1$

M step: (pseudo)values for QTL group means and variance

$$\hat{\mu}_q = \sum_i y_i w_{iq} / \sum_i w_{iq}$$

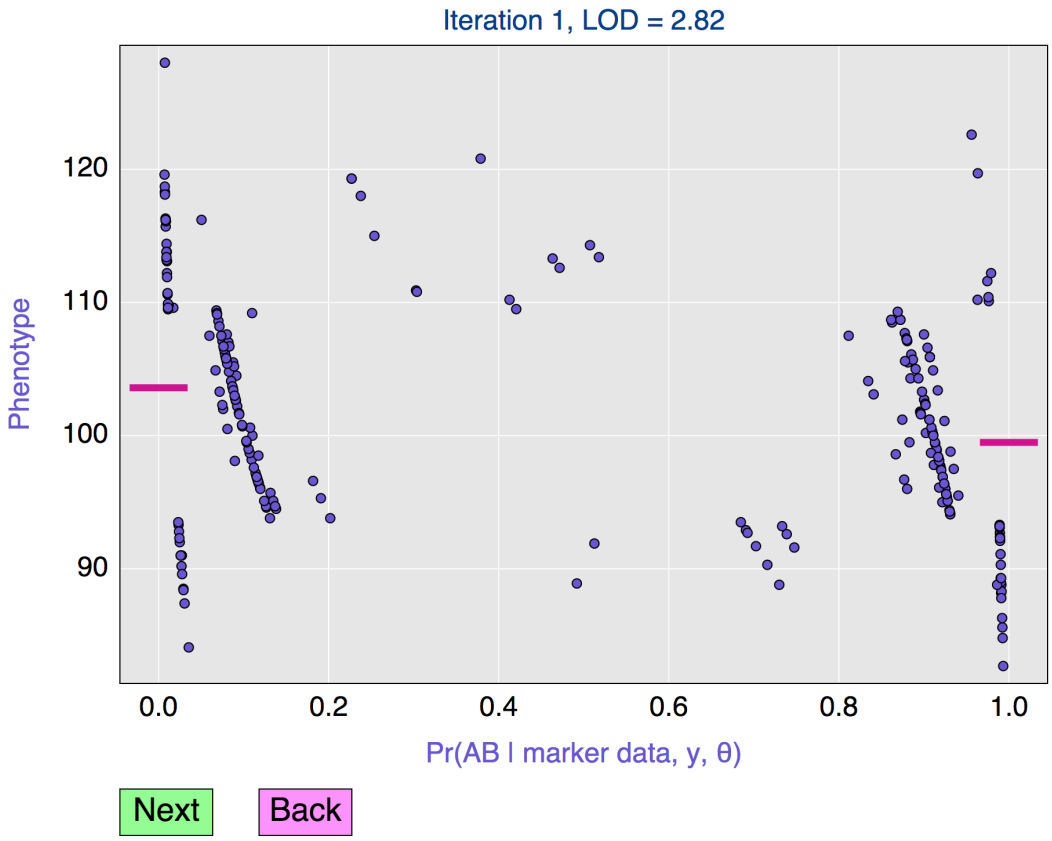
$$\hat{\sigma}^2 = \sum_i \sum_q w_{iq}(y_i - \hat{\mu}_q)^2 / n$$

EM algorithm: set $w_{iq} = \texttt{pr}(q|m_i)$; iterate E&M to converge

Haley-Knott regression

Idea: just run one iteration of EM algorithm

- becomes marker regression on genotype probabilities
- ignores mixture of normals issue
- now widely used for dense marker maps (high throughput)



Haley, Knott (1992

Martinez, Curnow (1992

LOD Scores

LOD score measures strength of evidence for QTL at locus λ
$\log_{10}$ likelihood ratio of models:

- model with QTL at λ (mean depends on QTL genotype q at λ)
- model with no QTL (common mean for all individuals)

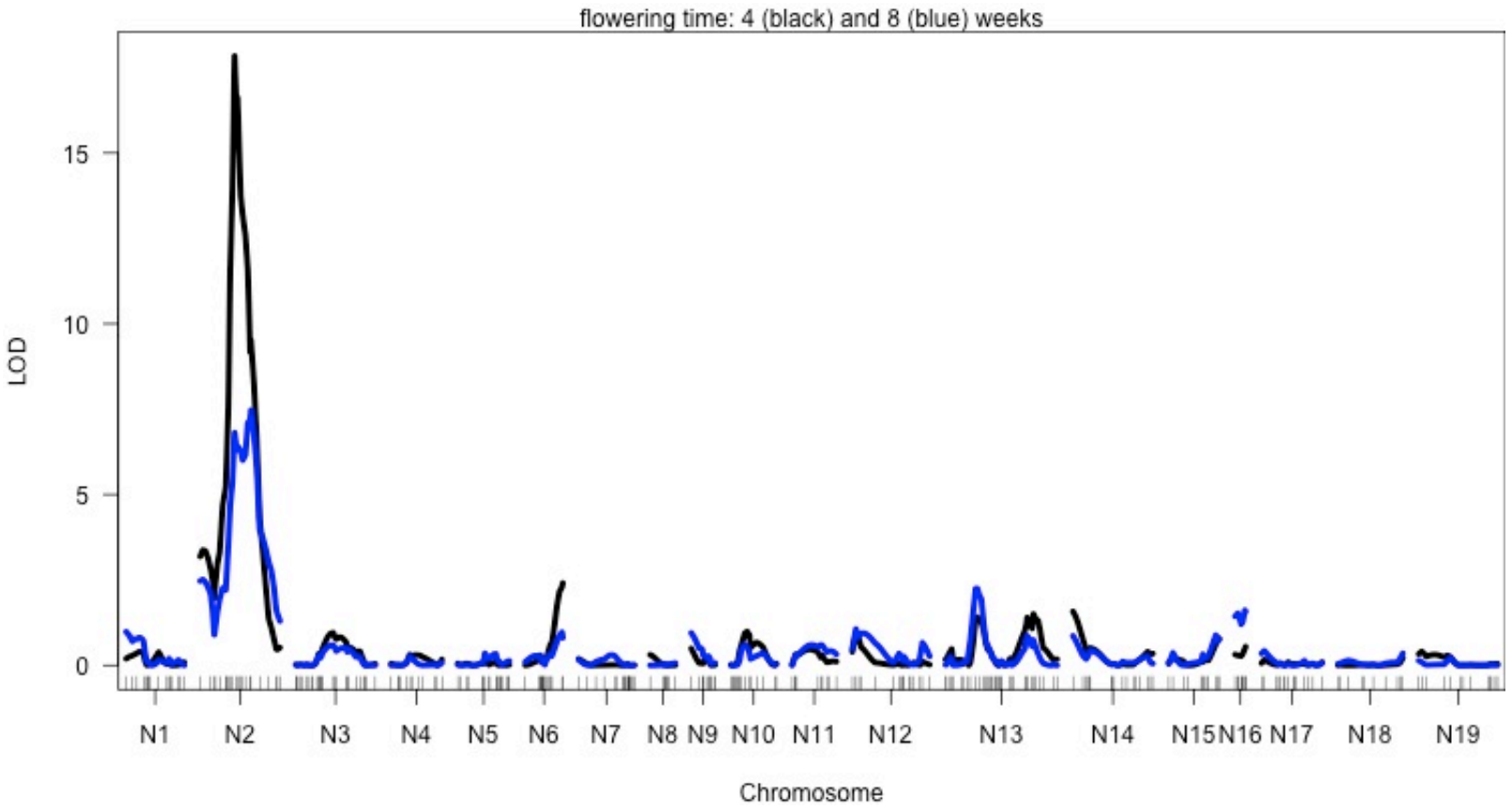
$$\texttt{lod}(\lambda) = [l(\hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda) - l(\hat{\mu}, \hat{\sigma})]/\log(10)$$

QTL model: means are MLEs $\hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}$ with QTL at λ

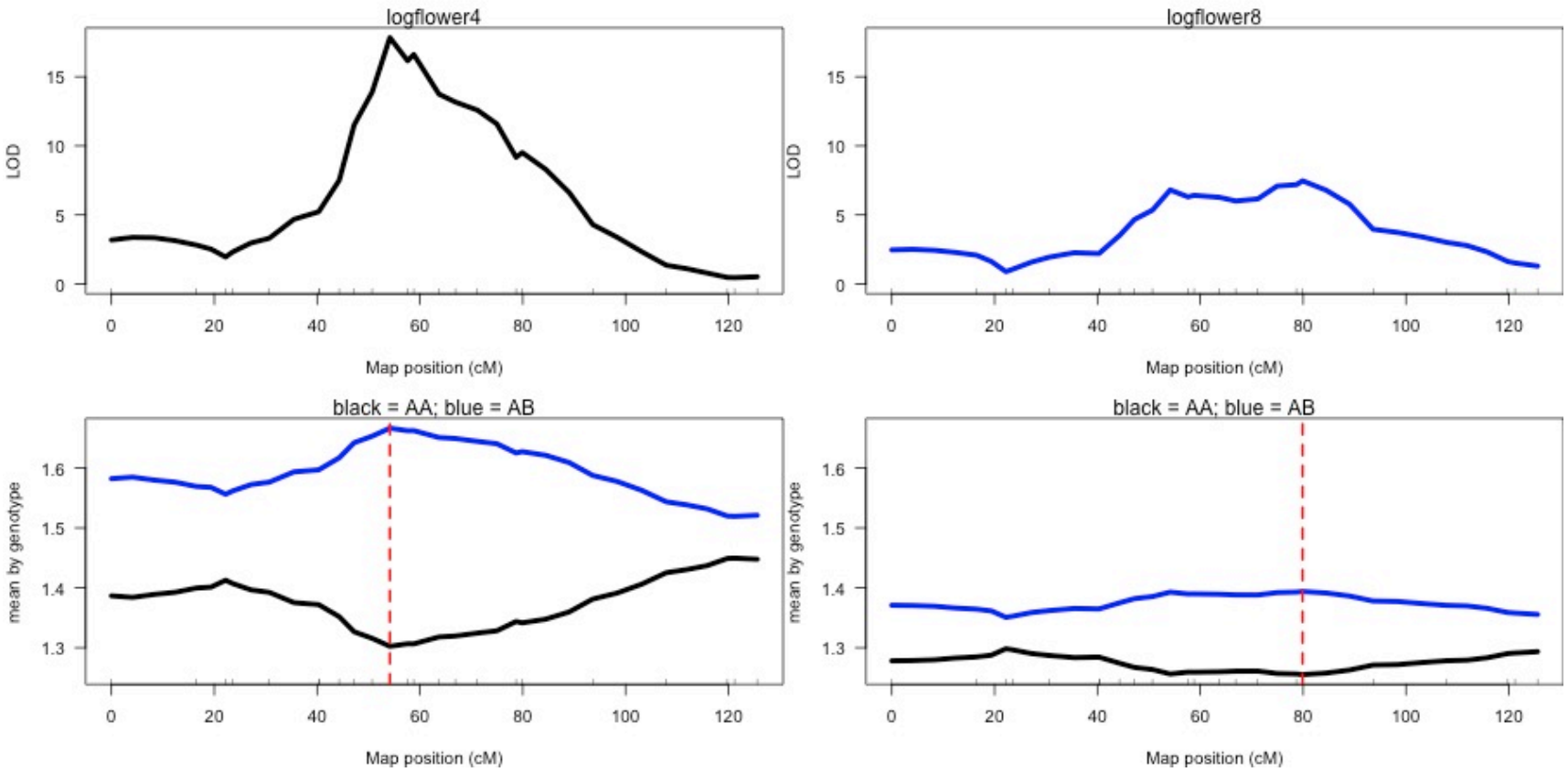
No QTL model: mean is unconditional MLE $\hat{\mu} = \bar{y}$

SD computed given model means: $\hat{\sigma}_\lambda, \hat{\sigma}$

LOD profile of flowering time

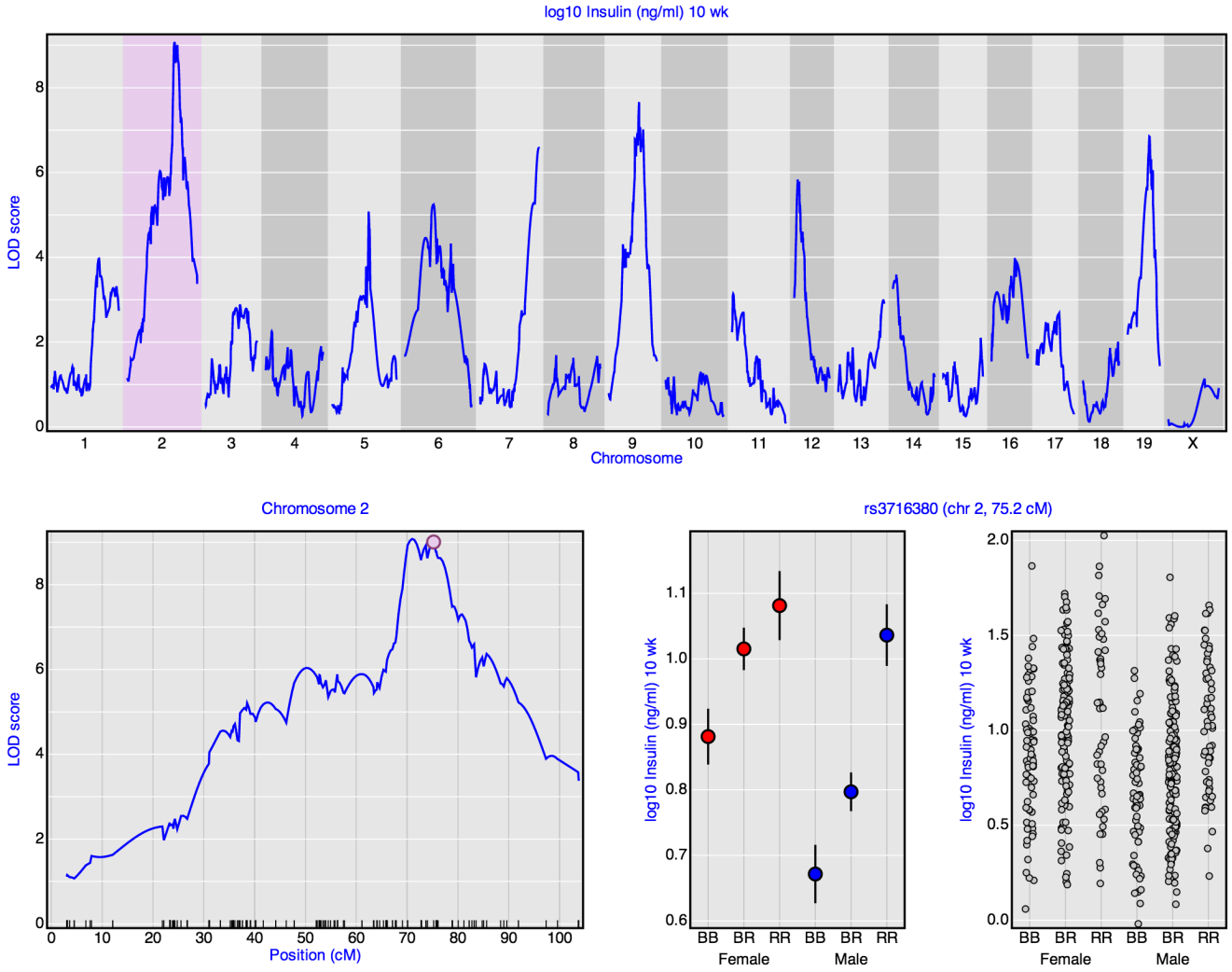


LOD profile for one chromosome



LOD and means by genotype scans on chr N2

Interactive LOD scan



pros and cons of IM

- Advantages
  - takes proper account of missing data
  - allows examination of positions between markers
  - gives improved estimates of QTL effects
  - provides pretty graphs (important!)
- Disadvantages
  - increased computation time
  - requires specialized software
  - difficult to generalize and extend
  - only one QTL at a time

LOD thresholds: how large is large?

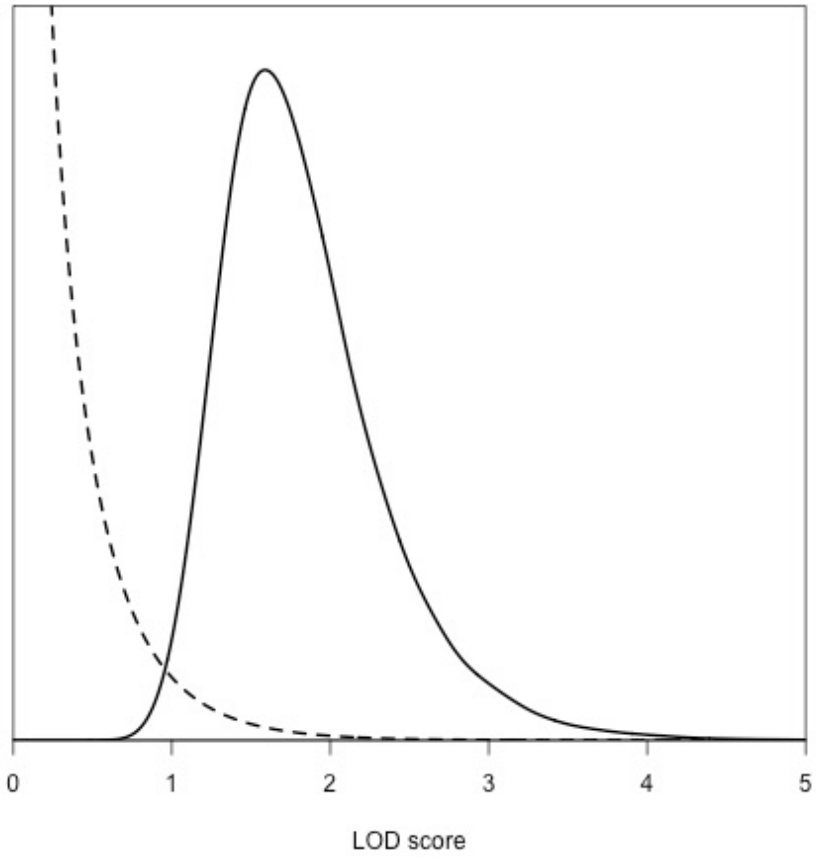
Large LOD scores = evidence for presence of a QTL
LOD threshold = 95 %ile of histogram of max LOD genome-wide (if there are no QTLs anywhere)

Derivation:

- Analytical calculations (Lander & Botstein 1989)
- Simulations (Lander & Botstein 1989)
- Permutation tests (Churchill & Doerge 1994)

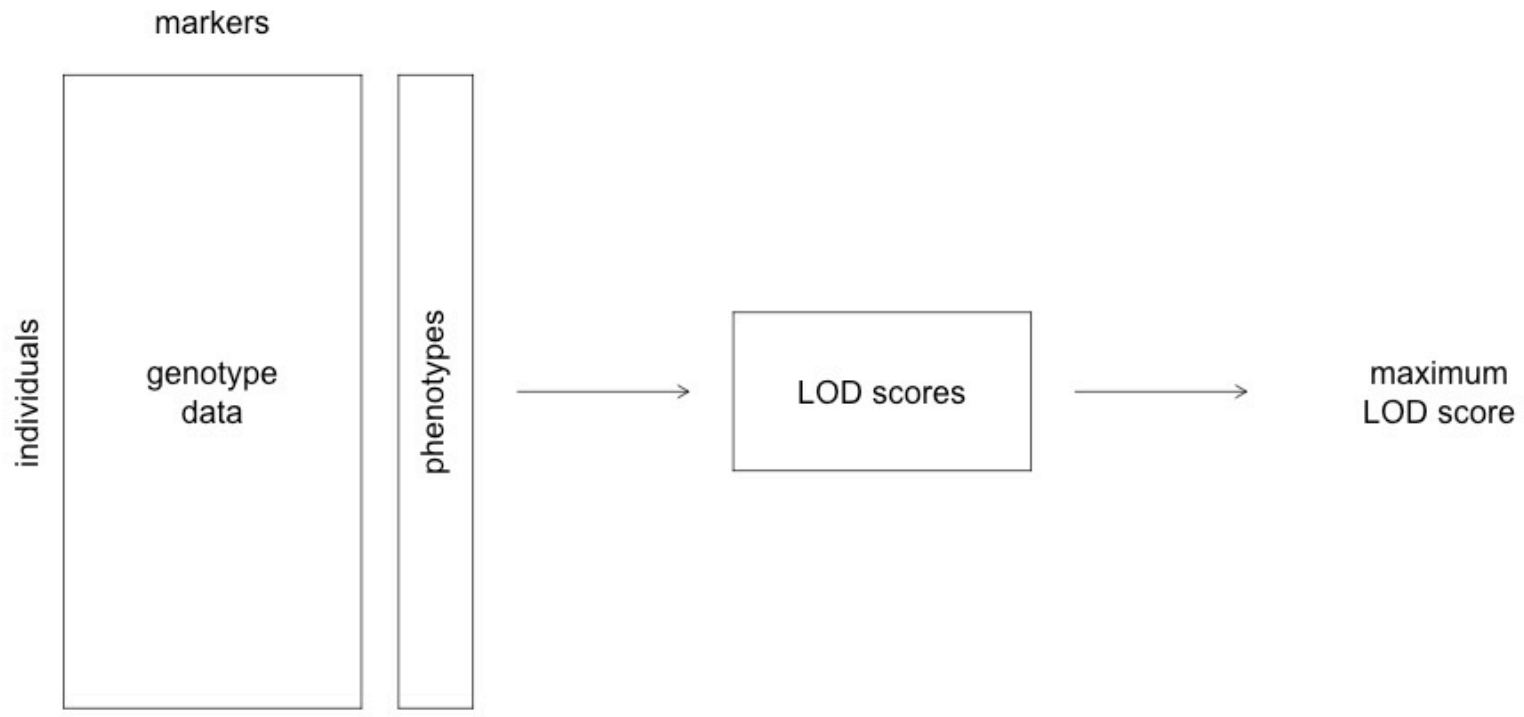
null distribution of the LOD score

- Null distribution from simulation
  - backcross with typical size genome
- Dashed curve:
  - LOD score histogram for any one point
- Solid curve:
  - max LOD histogram, genome-wide

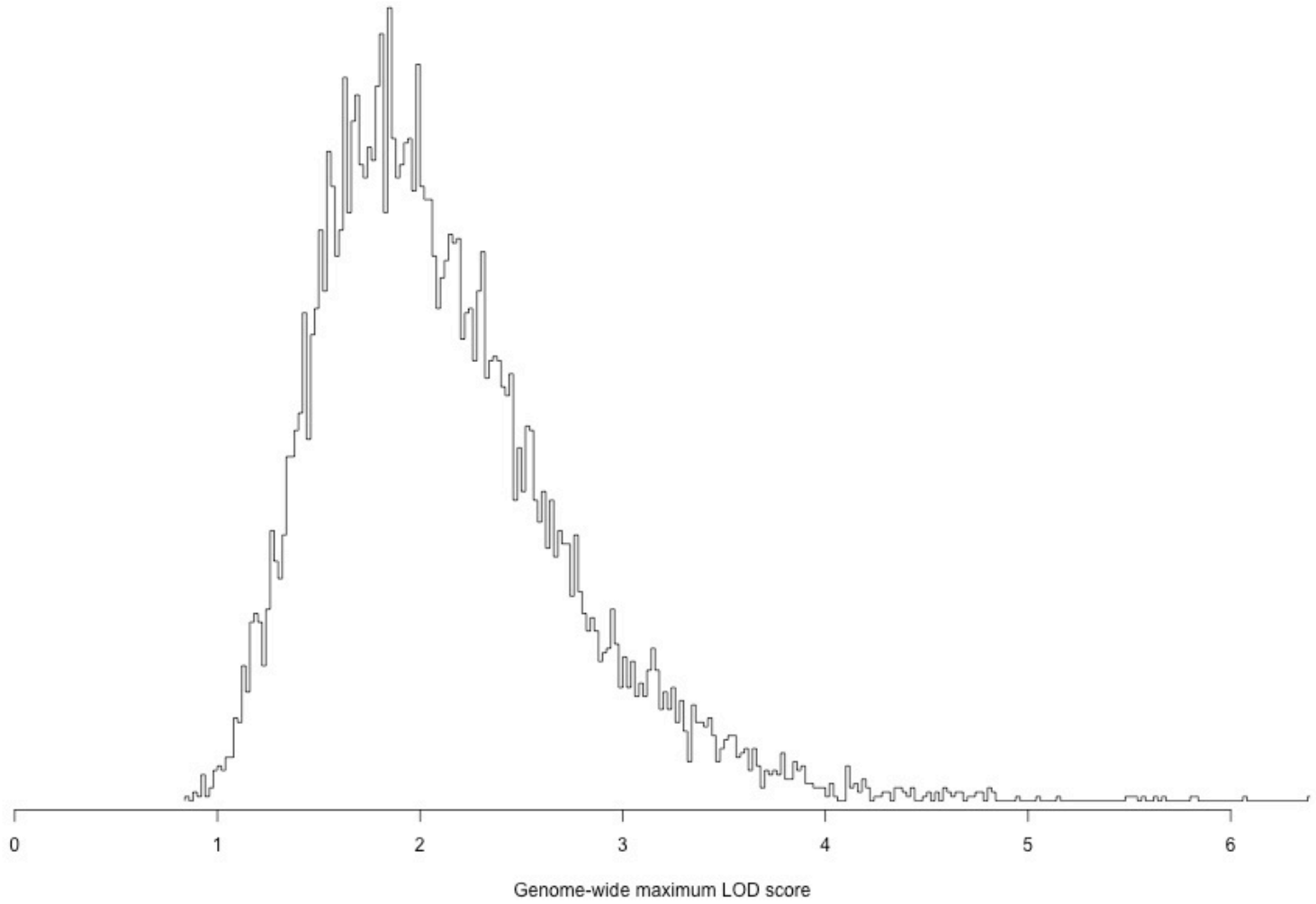


permutation test schematic

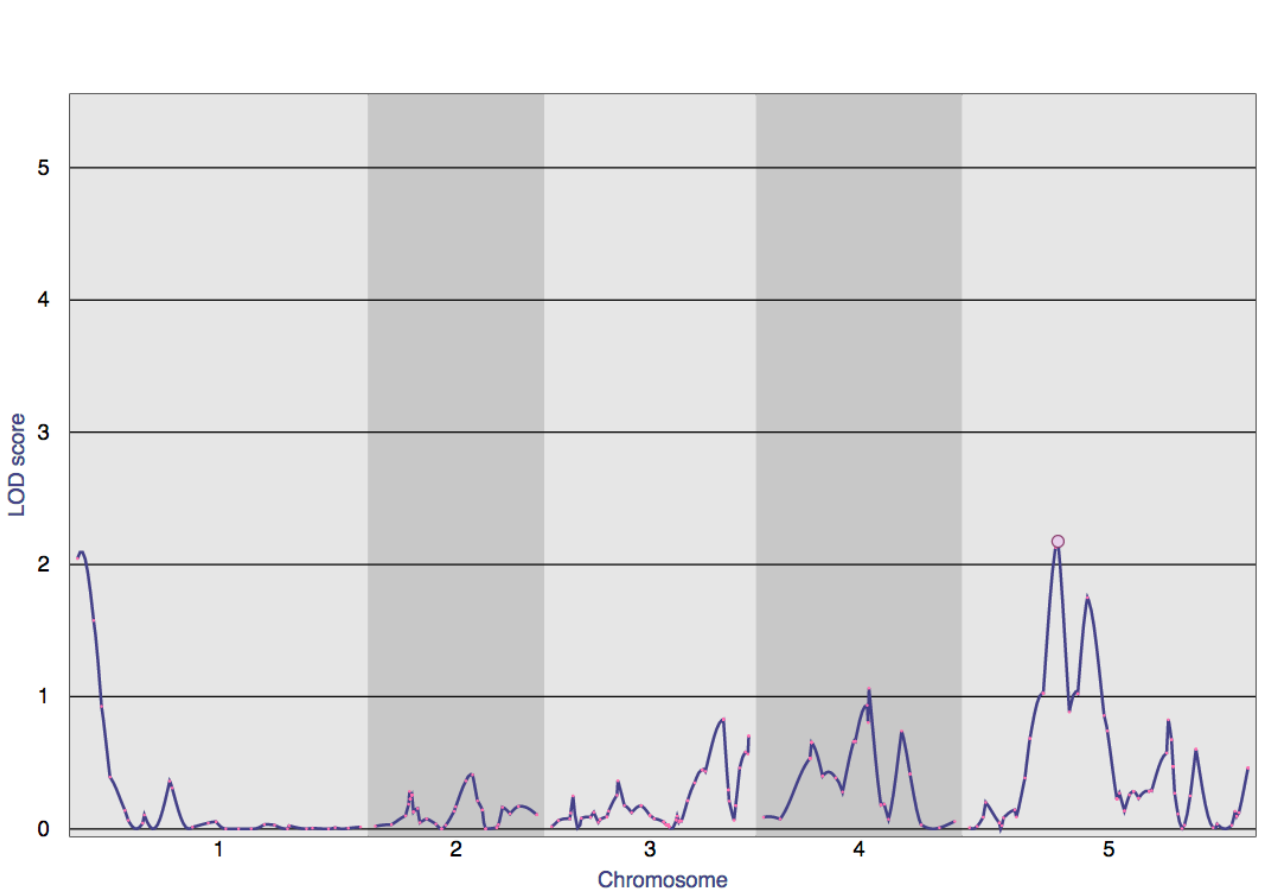
shuffle phenotypes independent of genotype data
repeat 10,000 times

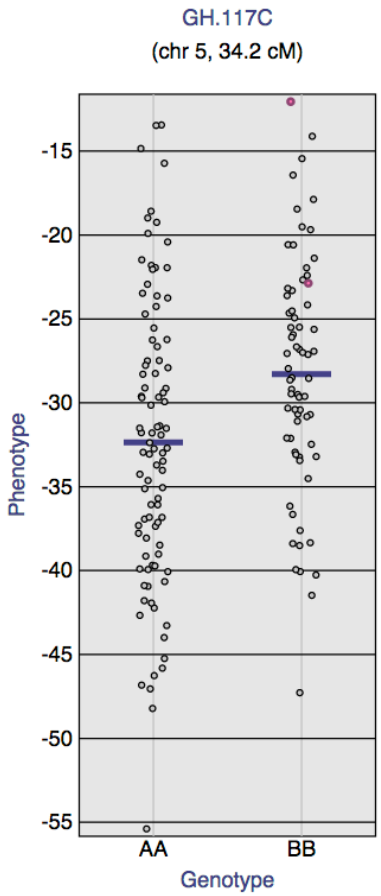


10,000 permutation results



interactive permutations

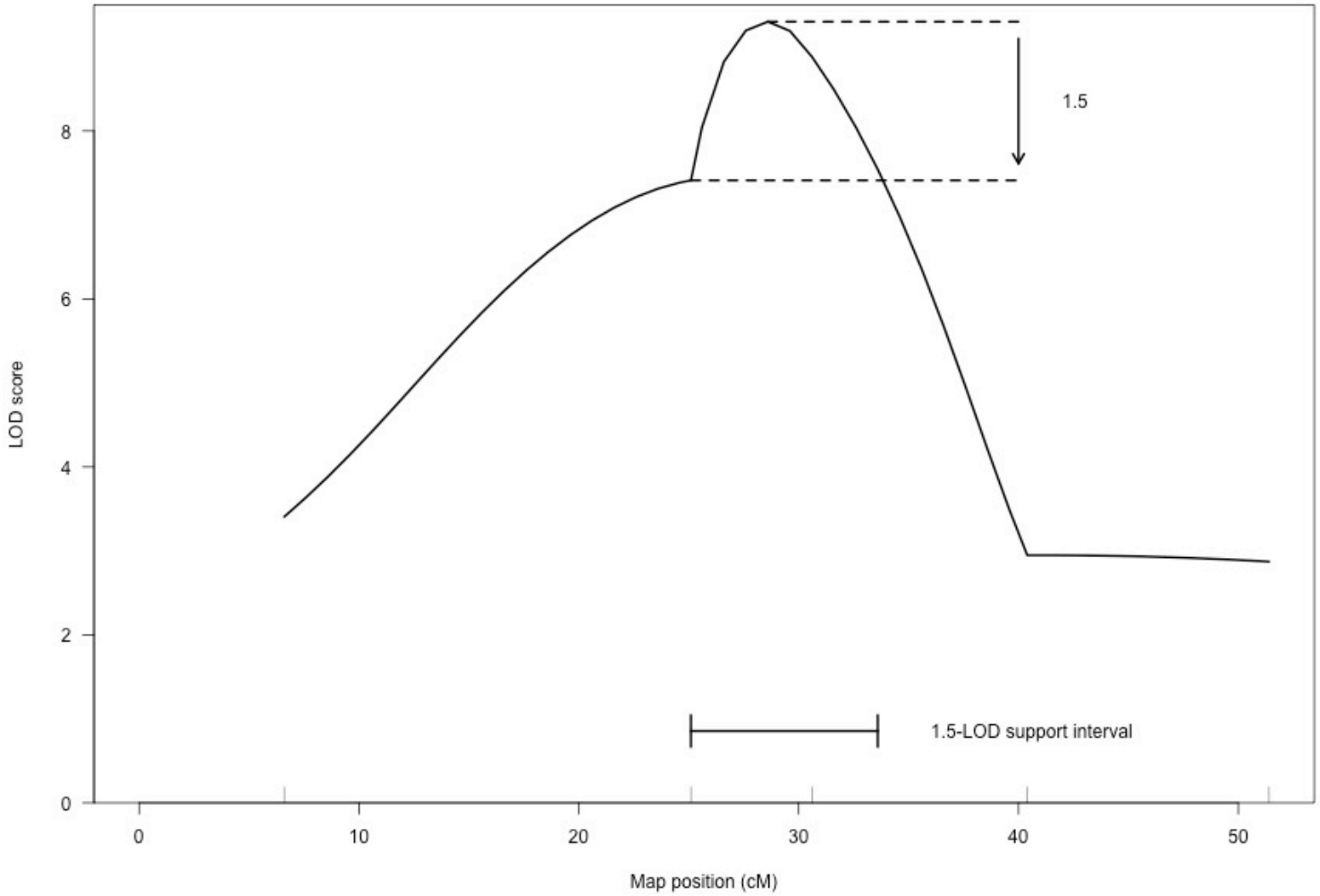




Randomize!

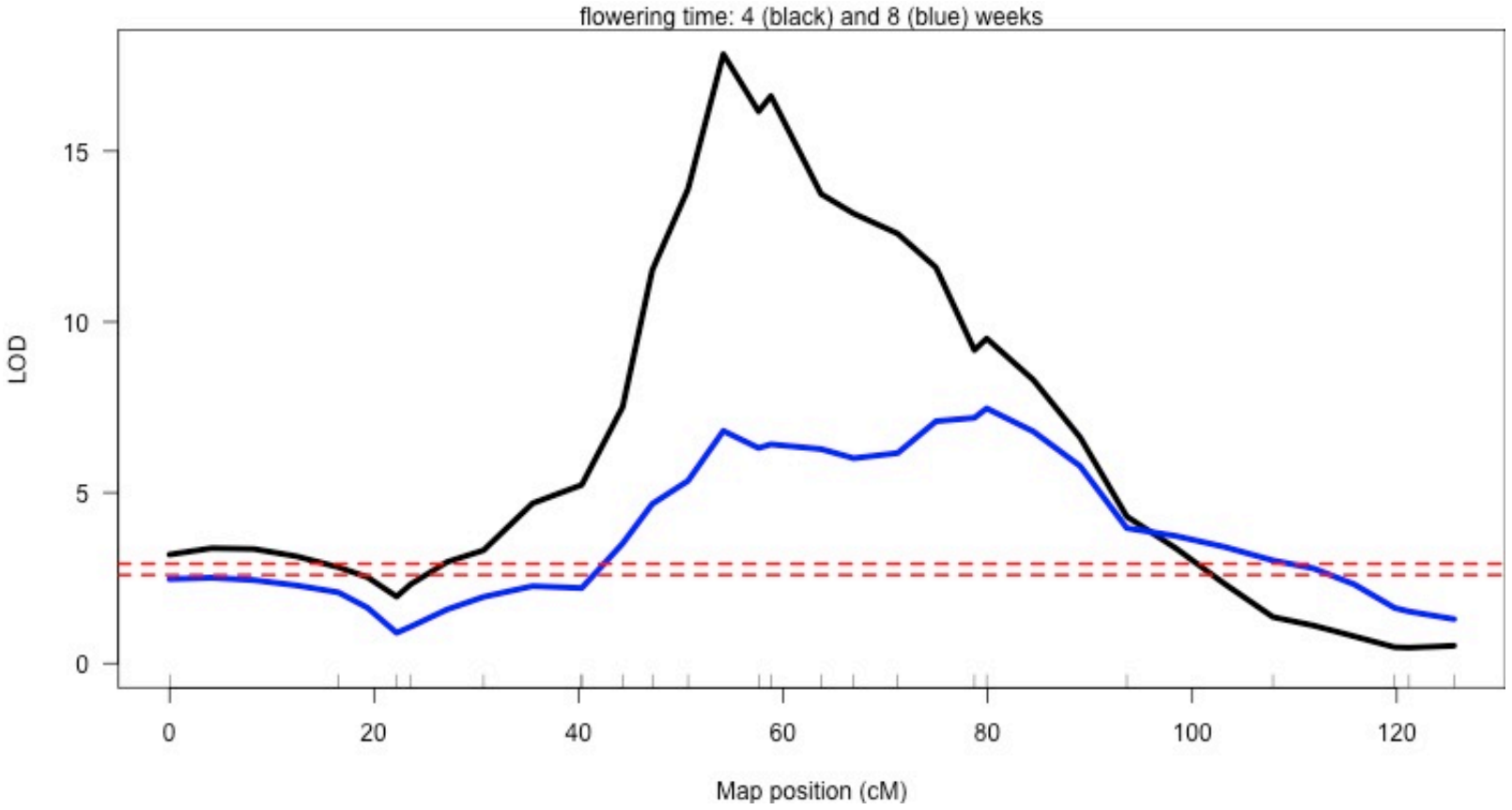
Back

LOD support intervals



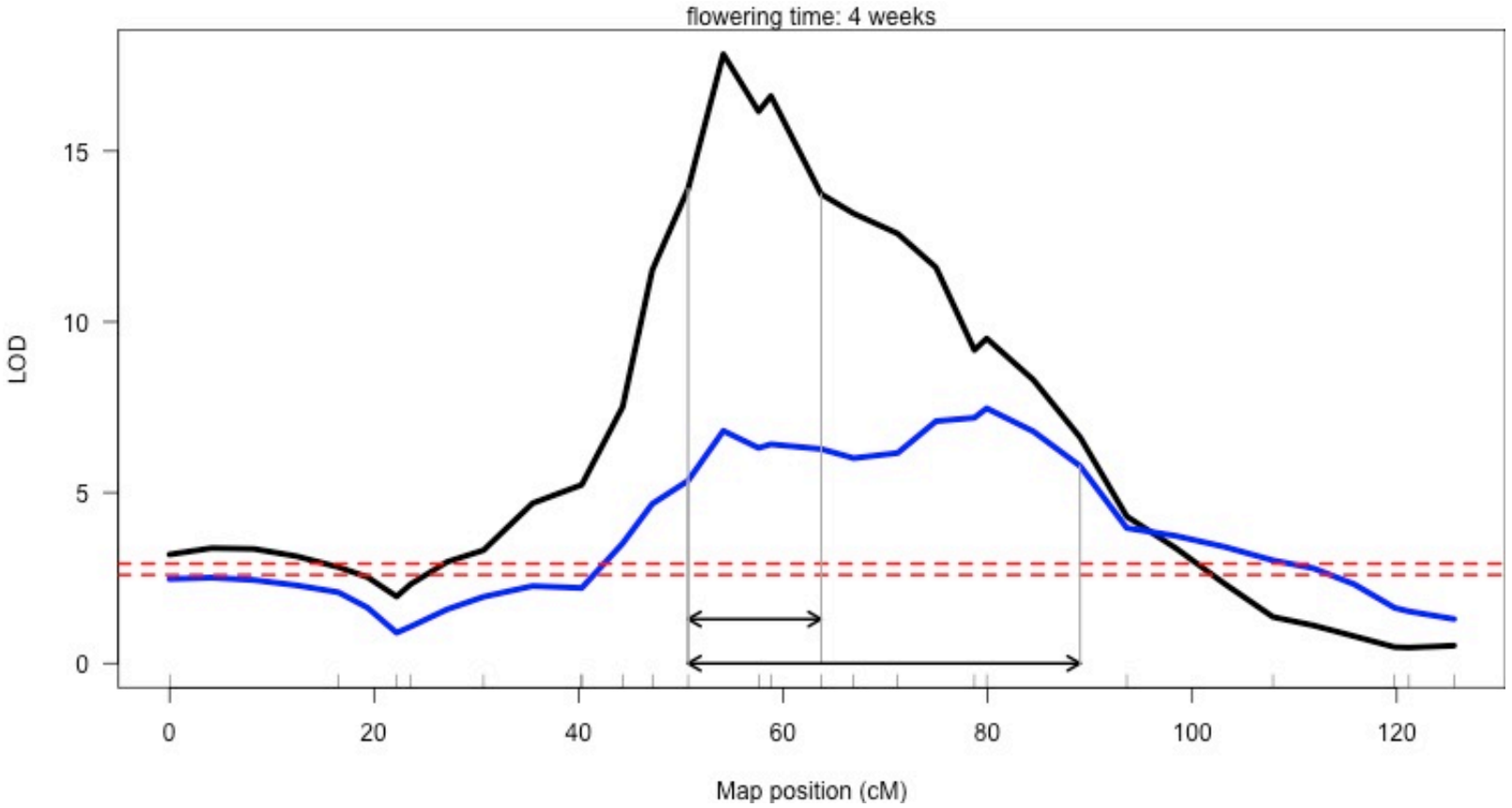
LOD thresholds for flowering time

significant area is quite broad …

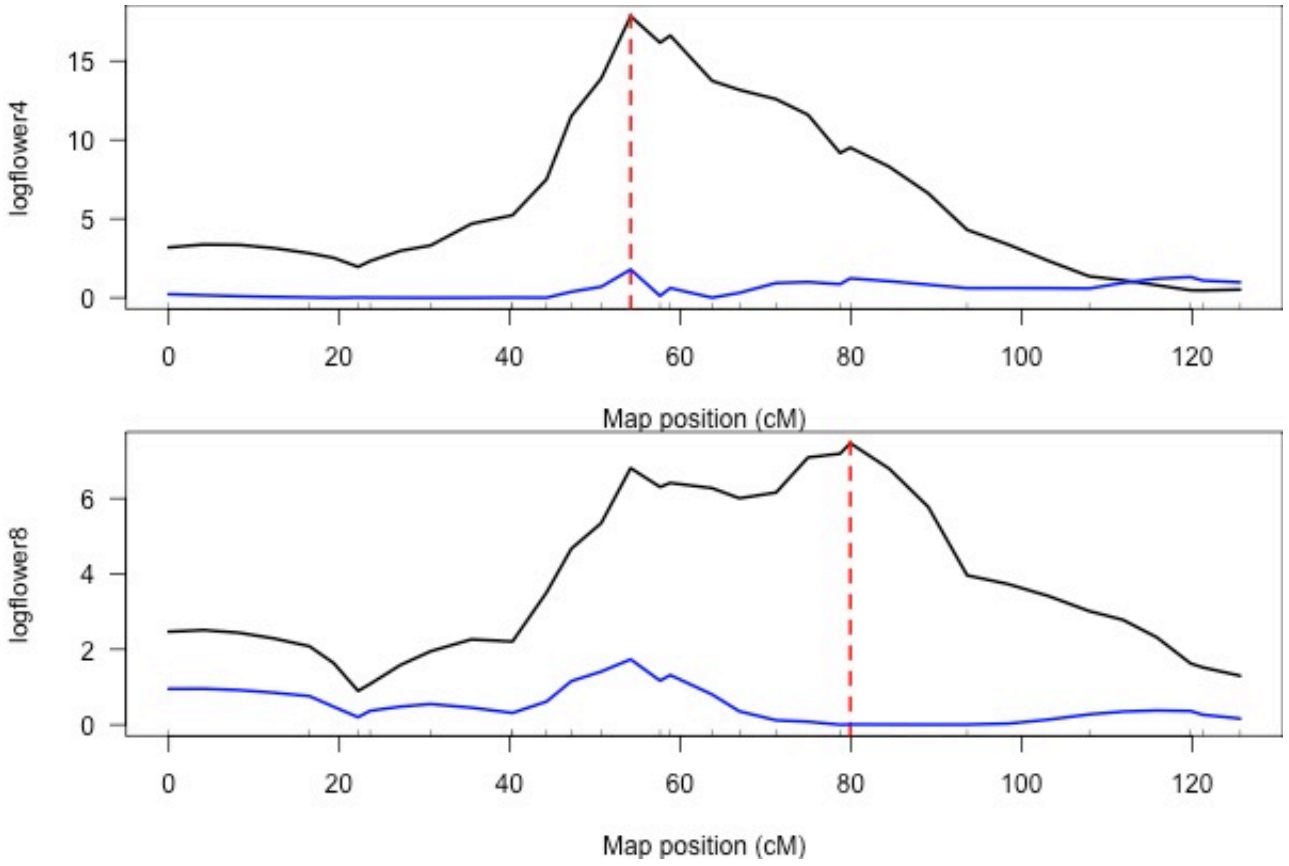


LOD thresholds for flowering time

but 1.5 LOD support interval is narrower



flowering time adjusted for QTL



QTL model search

- Goals
  - identify QTL (and possible interactions among QTL)
  - estimate interval for QTL location
  - estimate QTL effects
- Challenges
  - how many QTL? which ones?
  - more complicated to fit each multiple QTL model
  - need rules to search across many QTL models

pros & cons of multiple QTL models

- benefits
  - reduce residual variation
  - increased power
  - separate linked QTL
  - identify interactions among QTL (epistasis)
- shortcomings
  - only includes significant loci
  - gets complicated very quickly
  - selection bias: overestimate effects of included loci
  - many loci of small effect ignored …

special nature of QTL models

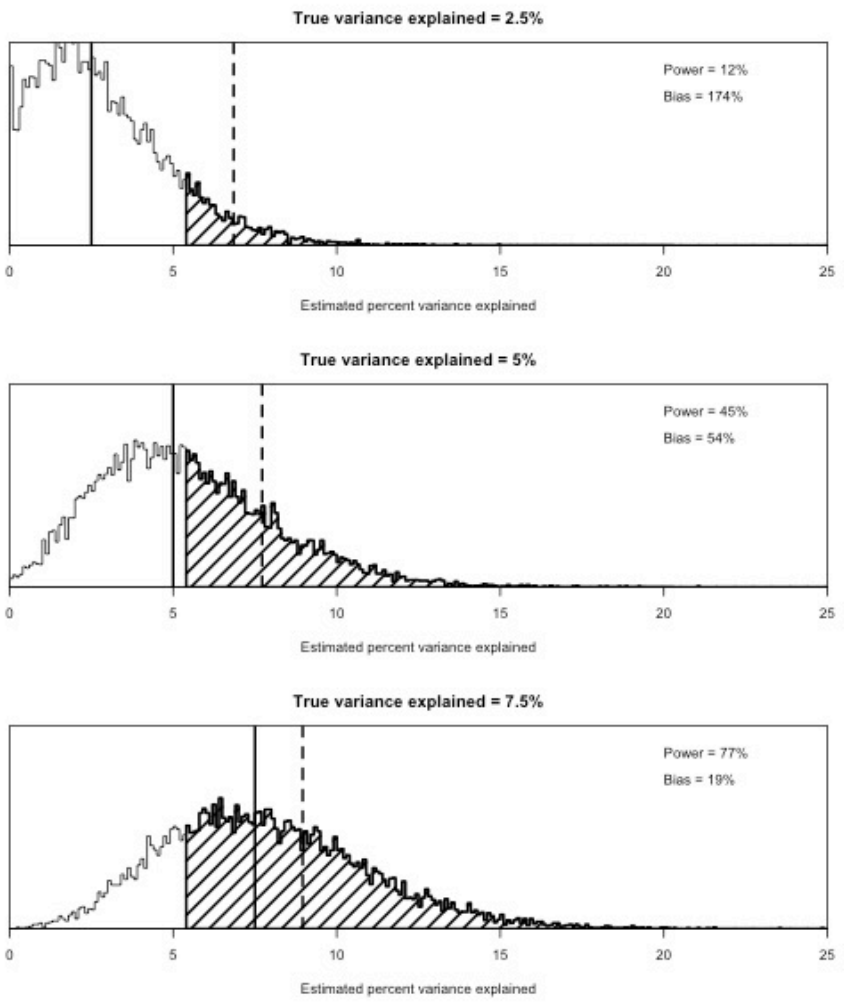
What is special here?

- continuum of ordinal-valued predictors (the genetic loci)
- association among these QTL predictors
- loci on different chromosomes are independent
- along chromosome:
    - simple (and known) correlation structure

See Broman MultiQTL talk for more details

selection bias

- estimated QTL effect QTL varies from true effect
- detect QTL when estimated effect is large
- experiments with detected QTL often have larger estimated than true effect
- selection bias largest in QTLs with small or moderate effects
- true QTL effects smaller than those observed



implications of selection bias

- estimated % variance explained by identified QTLs: too high
- repeating an experiment: different QTL (Beavis effect)
- congenics (or near isogenic lines): off base
- marker-assisted selection: missed effect

See Broman (2003) and Haley, Knott (1992).

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.

Pareto chart: from QTL to GWA