

The EM algorithm

QTL mapping with a cure model

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

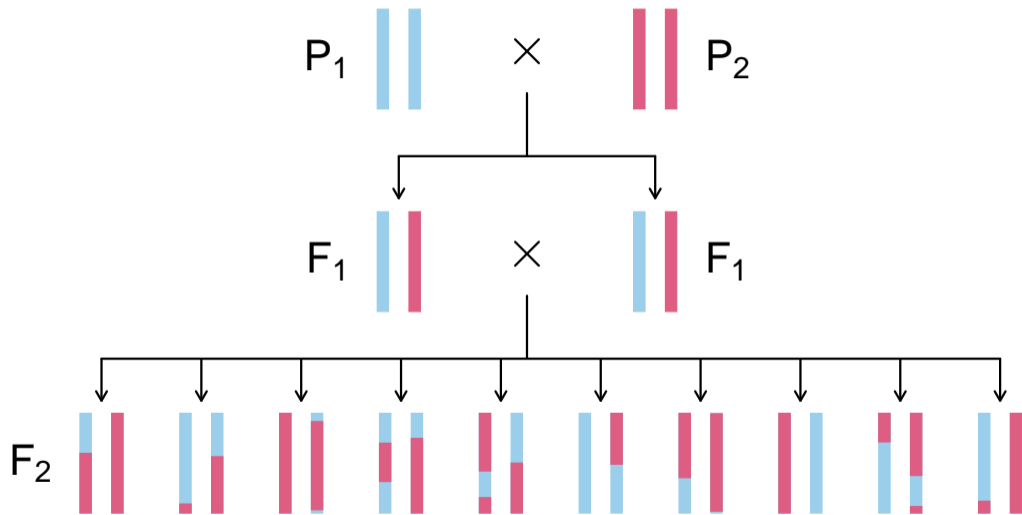
kbroman.org

github.com/kbroman

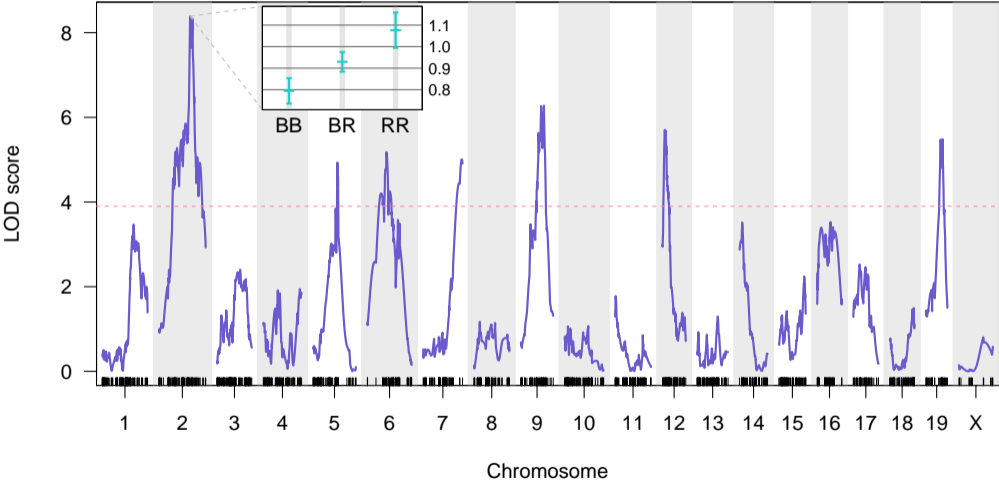
@kwbroman

Course web: kbroman.org/AdvData

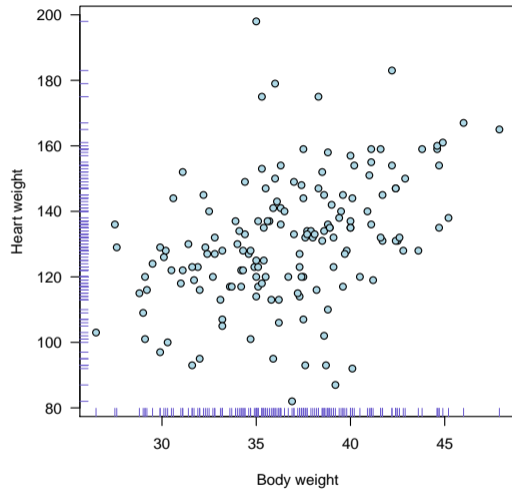
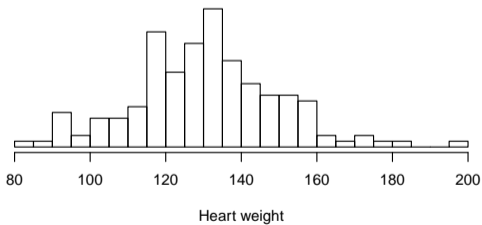
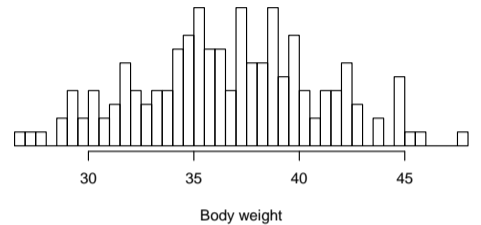
Intercross



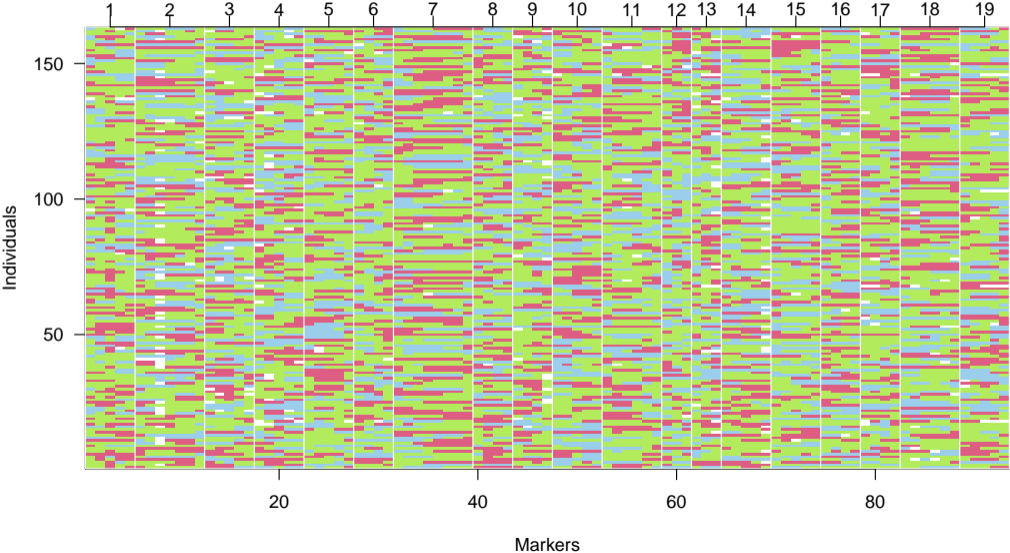
QTL mapping



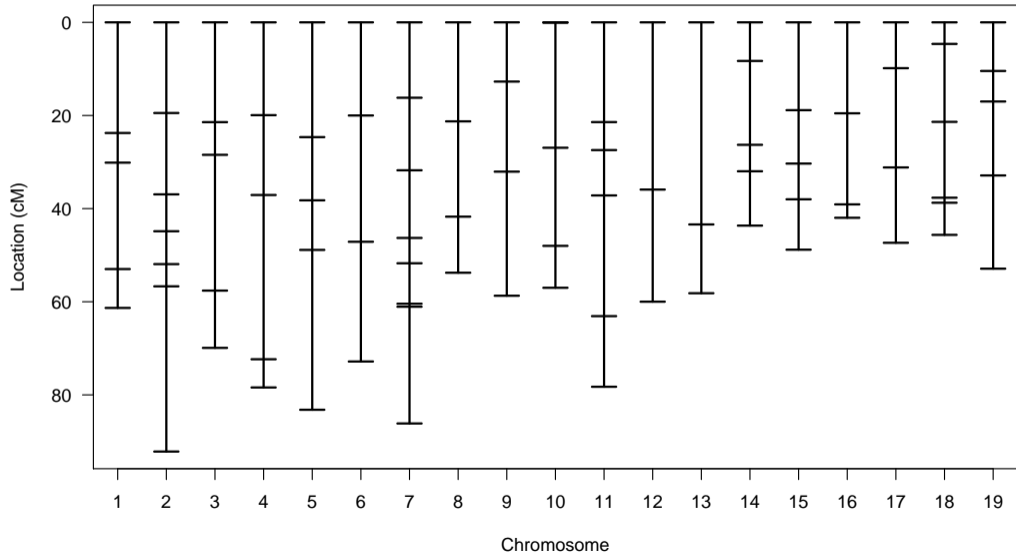
Phenotype data



Genotype data

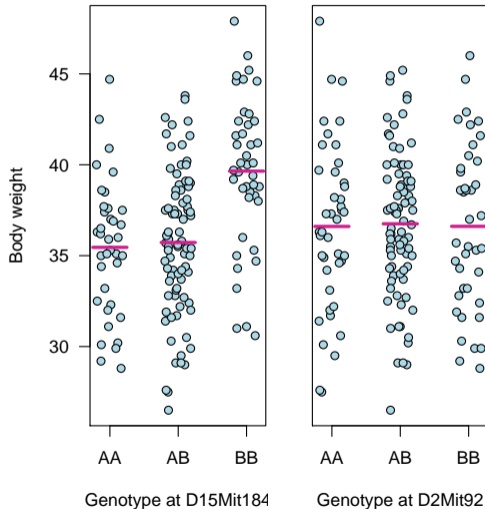


Genetic map



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

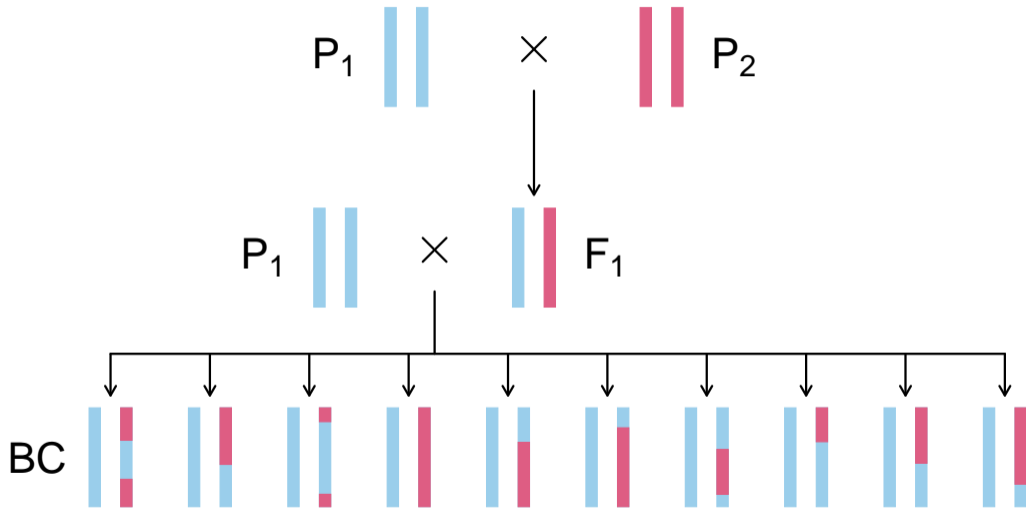
Lander & Botstein (1989)

- ▶ Assume a **single** QTL model.
- ▶ Each position in the genome, one at a time, is posited as the putative QTL.
- ▶ Let $g = 0/1/2$ if the (unobserved) QTL genotype is AA/AB/BB.

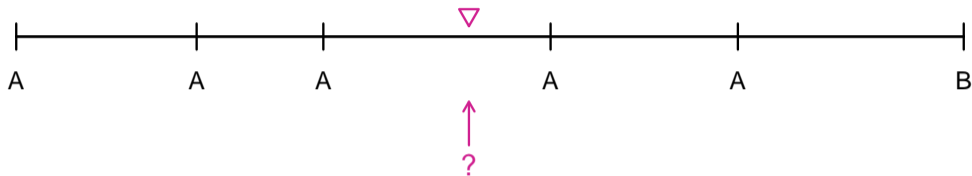
Assume $y|g \sim N(\mu_g, \sigma)$

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

Backcross



Genotype probabilities



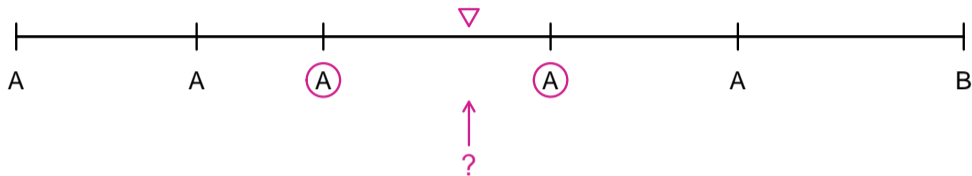
Calculate $\Pr(g \mid \text{marker data})$, assuming

- ▶ 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



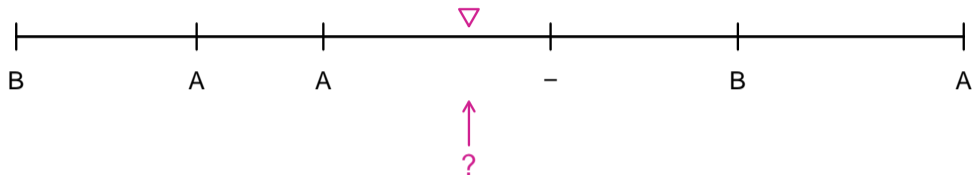
Calculate $\Pr(g \mid \text{marker data})$, assuming

- ▶ 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



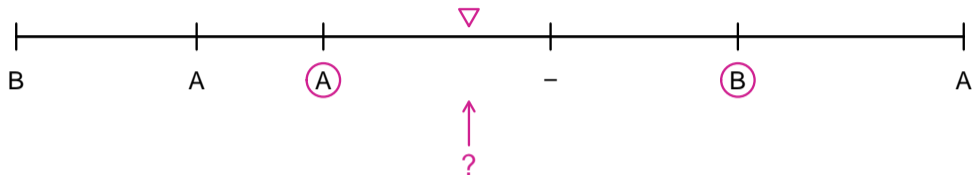
Calculate $\Pr(g \mid \text{marker data})$, assuming

- ▶ 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



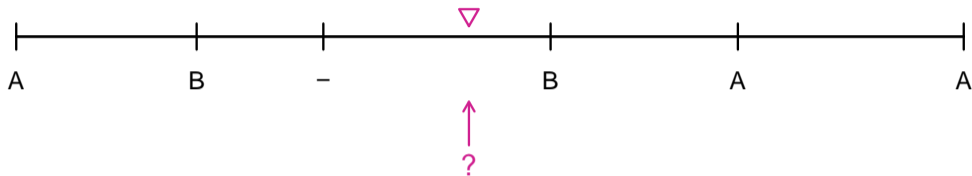
Calculate $\Pr(g \mid \text{marker data})$, assuming

- ▶ 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



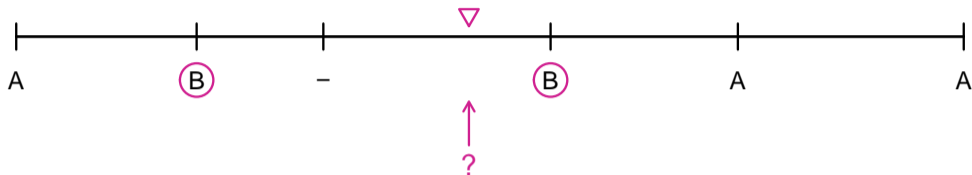
Calculate $\Pr(g \mid \text{marker data})$, assuming

- ▶ 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



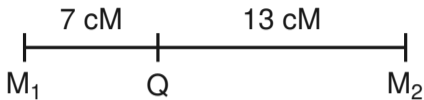
Calculate $\Pr(g \mid \text{marker data})$, assuming

- ▶ No crossover interference
- ▶ No genotyping errors

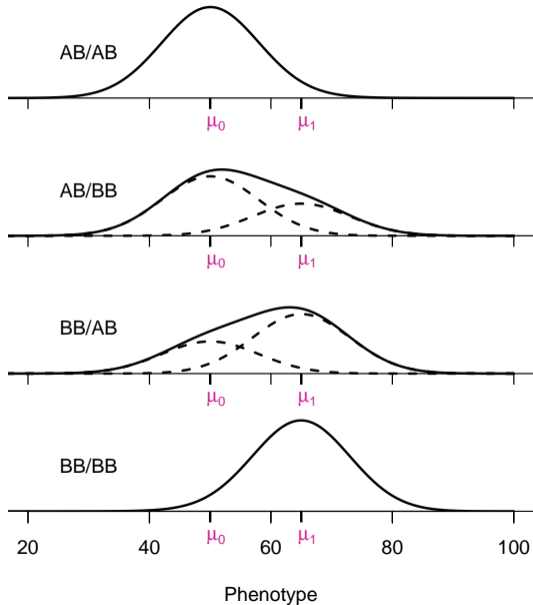
Or use the **hidden Markov model (HMM)** technology

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

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

$\Pr(y_i | \text{marker data}, \mu_0, \mu_1, \sigma) = \sum_j p_{ij} f(y_i; \mu_j, \sigma)$

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

Log likelihood: $l(\mu_0, \mu_1, \sigma) = \sum_i \log \Pr(y_i | \text{marker data}, \mu_0, \mu_1, \sigma)$

Maximum likelihood estimates (**MLEs**) of μ_0, μ_1, σ :
values for which $l(\mu_0, \mu_1, \sigma)$ is maximized.

EM algorithm

E step:

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

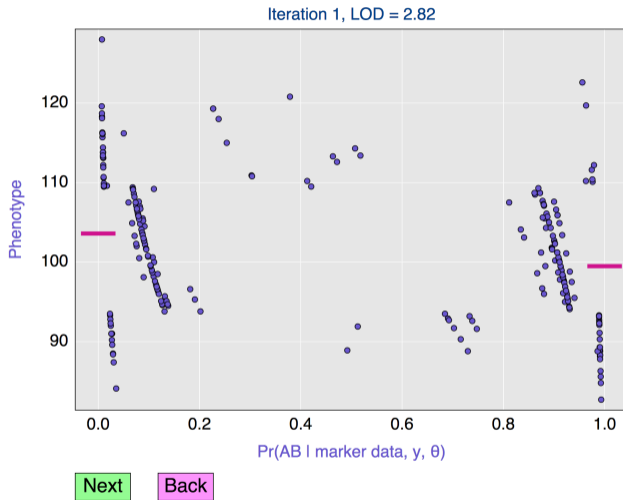
M step:

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

The algorithm:

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

Interactive illustration



bit.ly/em_alg

LOD scores

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

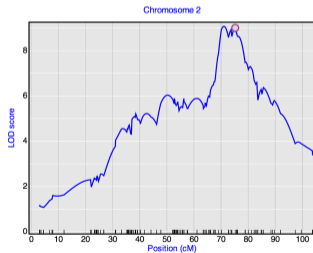
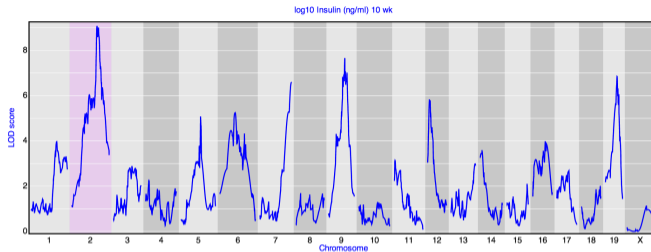
$$\begin{aligned}\text{LOD}(\lambda) &= \log_{10} \text{likelihood ratio comparing the hypothesis of a} \\ &\quad \text{QTL at position } \lambda \text{ versus that of no QTL} \\ &= \log_{10} \left\{ \frac{\text{Pr}(y|\text{QTL at } \lambda, \hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda)}{\text{Pr}(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}\end{aligned}$$

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

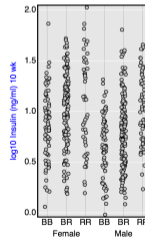
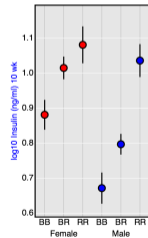
No QTL model:

The phenotypes are independent and identically distributed (iid) $N(\mu, \sigma^2)$.

Interactive plot



rs3716380 (chr 2, 75.2 cM)



bit.ly/D31lod

Interval mapping

Advantages

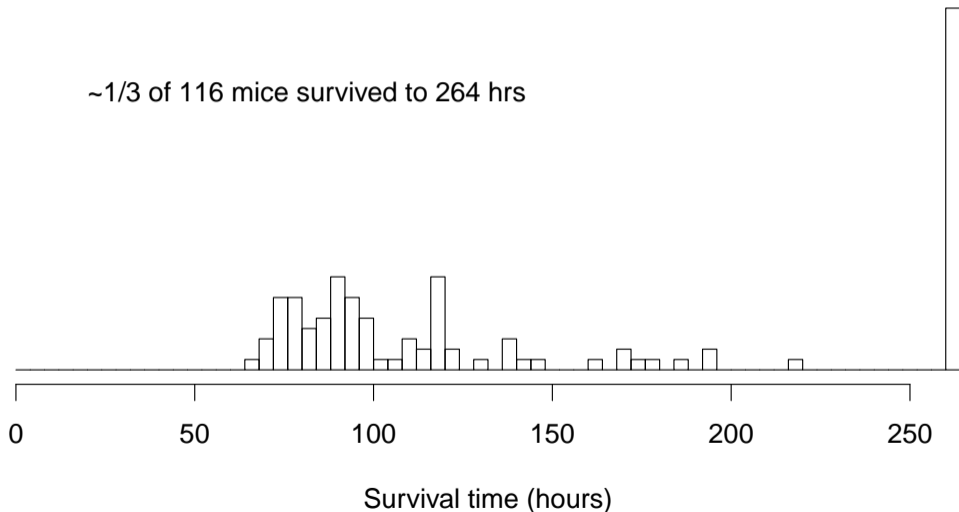
- ▶ Takes proper account of missing data.
- ▶ Allows examination of positions between markers.
- ▶ Gives improved estimates of QTL effects.
- ▶ Provides pretty graphs.

Disadvantages

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

Survival after Listeria infection

~1/3 of 116 mice survived to 264 hrs



Normal assumption in ANOVA

- ▶ ANOVA is remarkably robust
- ▶ Transformation
- ▶ Rank-based methods
- ▶ Specially-tailored models (e.g. GLM)

Censoring?

Measurements with a spike at 0

- ▶ Mass of gallstones
- ▶ Gene expression, when a gene might be turned off
- ▶ Microbiome data, when a microbe might be absent
- ▶ Area of garage

Two-part (“cure”) model

- ▶ Let $z_i = 1$ if mouse i survived the infection

y_i = survival time

- ▶ Assume $\Pr(z_i|g) = \pi_g$

$y_i|z_i = 0, g \sim \text{Normal}(\mu_g, \sigma)$

$\{(y_i, z_i, g)\}$ mutually independent

EM algorithm

E step

$$w_{ij}^{(s+1)} = \Pr(g_i = j | y_i, z_i, \mathbf{m}_i, \hat{\boldsymbol{\theta}}^{(s)})$$
$$= \begin{cases} \frac{p_{ij}(1 - \hat{\pi}_j^{(s)})}{\sum_k p_{ik}(1 - \hat{\pi}_k^{(s)})} & \text{if } z_i = 0 \\ \frac{p_{ij}\hat{\pi}_j^{(s)} f(y_i; \hat{\boldsymbol{\mu}}_j^{(s)}, \hat{\boldsymbol{\sigma}}^{(s)})}{\sum_k p_{ik}\hat{\pi}_k^{(s)} f(y_i; \hat{\boldsymbol{\mu}}_k^{(s)}, \hat{\boldsymbol{\sigma}}^{(s)})} & \text{if } z_i = 1. \end{cases}$$

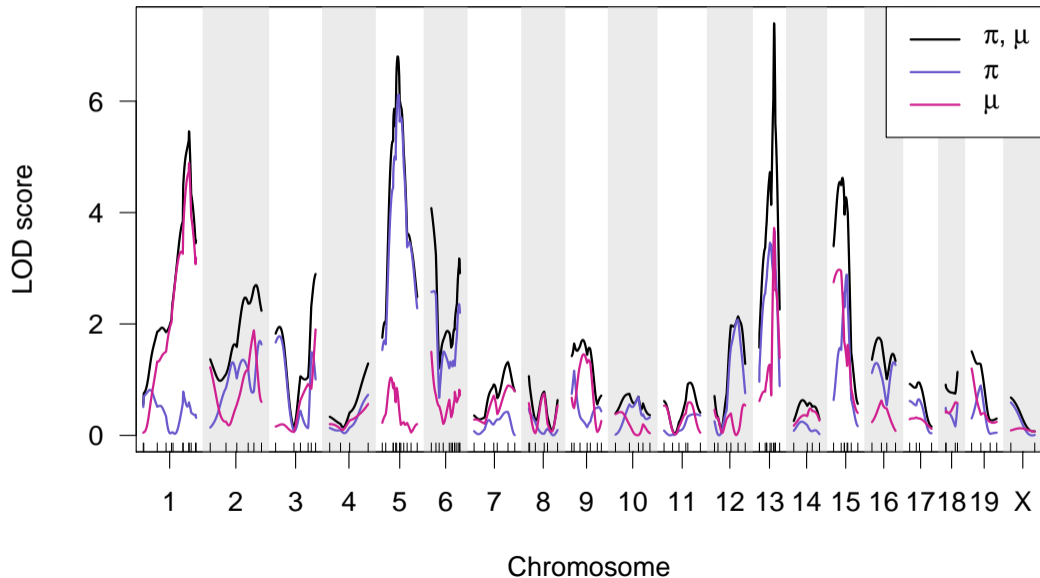
M step

$$\hat{\pi}_j^{(s+1)} = \frac{\sum_i w_{ij}^{(s+1)} z_i}{\sum_i w_{ij}^{(s+1)}}$$
$$\hat{\boldsymbol{\mu}}_j^{(s+1)} = \frac{\sum_i y_i w_{ij}^{(s+1)} z_i}{\sum_i w_{ij}^{(s+1)} z_i}$$
$$\hat{\boldsymbol{\sigma}}^{(s+1)} = \sqrt{\frac{\sum_i \sum_j (y_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})^2 w_{ij}^{(s+1)} z_i}{\sum_i z_i}}.$$

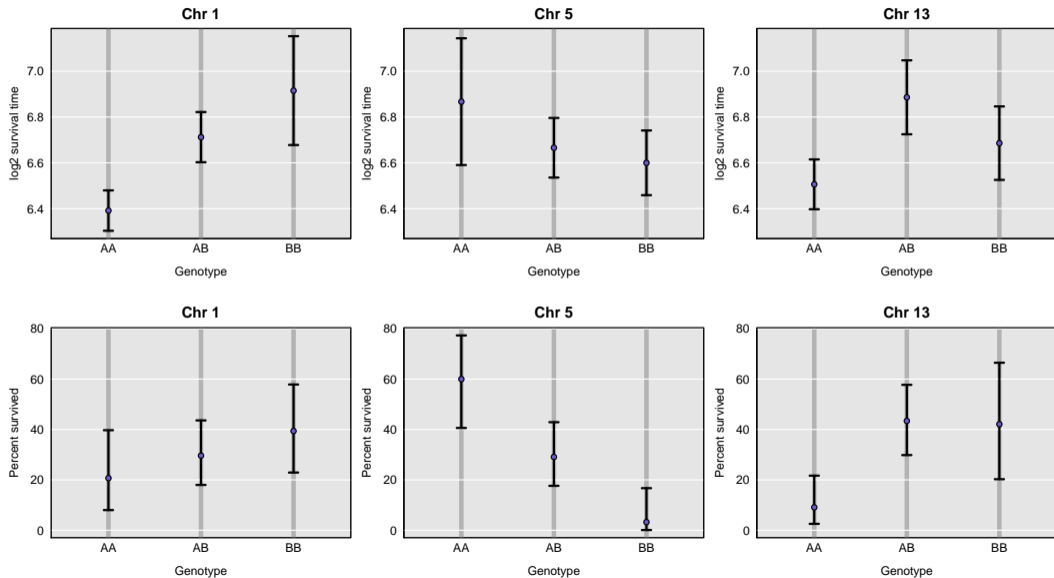
Tests

- ▶ $\pi_{AA} = \pi_{AB} = \pi_{BB}$
- ▶ $\mu_{AA} = \mu_{AB} = \mu_{BB}$
- ▶ $\pi_{AA} = \pi_{AB} = \pi_{BB}$ and $\mu_{AA} = \mu_{AB} = \mu_{BB}$

LOD curves



QTL effects



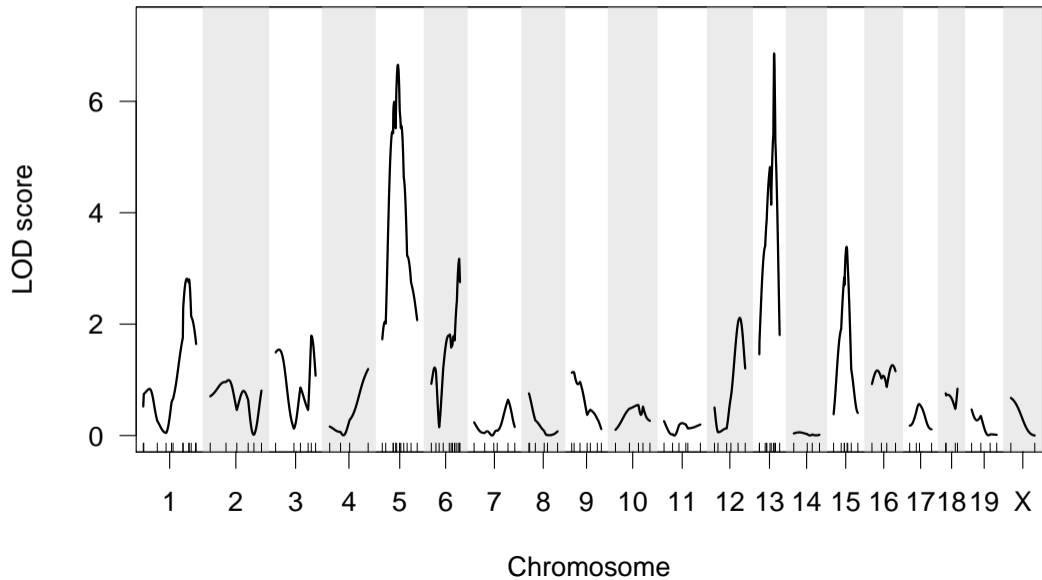
Lesson

- ▶ Don't just cram your data into the standard approach.

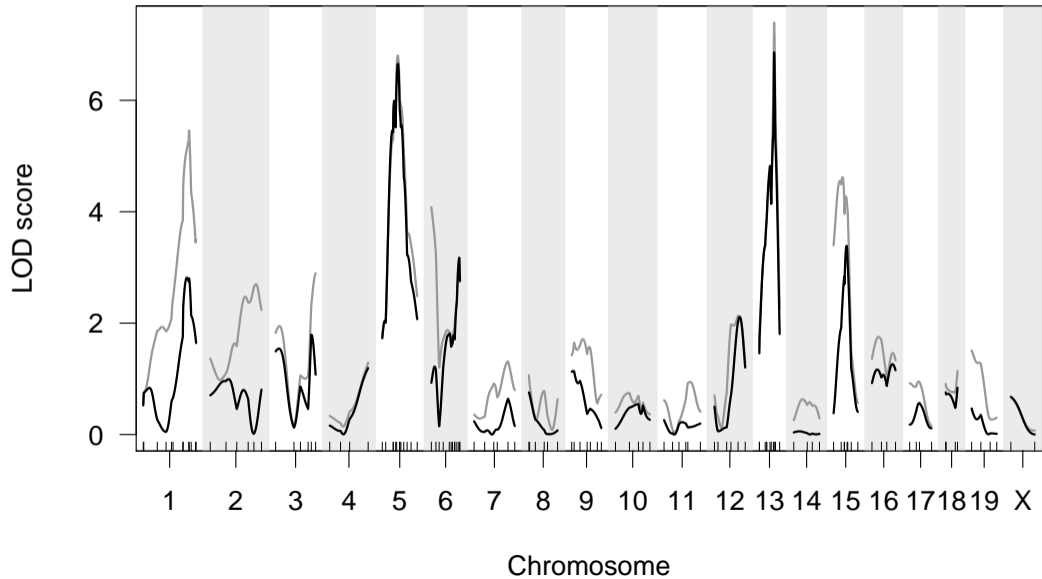
Lessons

- ▶ Don't just cram your data into the standard approach.
- ▶ Cramming your data into the standard approach might work fine.

Standard approach



Standard approach



References

- ▶ Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185-199
[PMCID: PMC1203601](#)
- ▶ Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal* 30(7):44-52
[PMID: 11469113](#)
- ▶ Boyartchuk VL, et al. (2001) Multigenic control of *Listeria monocytogenes* susceptibility in mice. *Nat Genet* 27:259-260
[doi:10.1038/85812](#)
- ▶ Broman KW (2003) Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics* 163:1169-1175
[PMCID: PMC1462498](#)