The EM algorithm

QTL mapping with a cure model

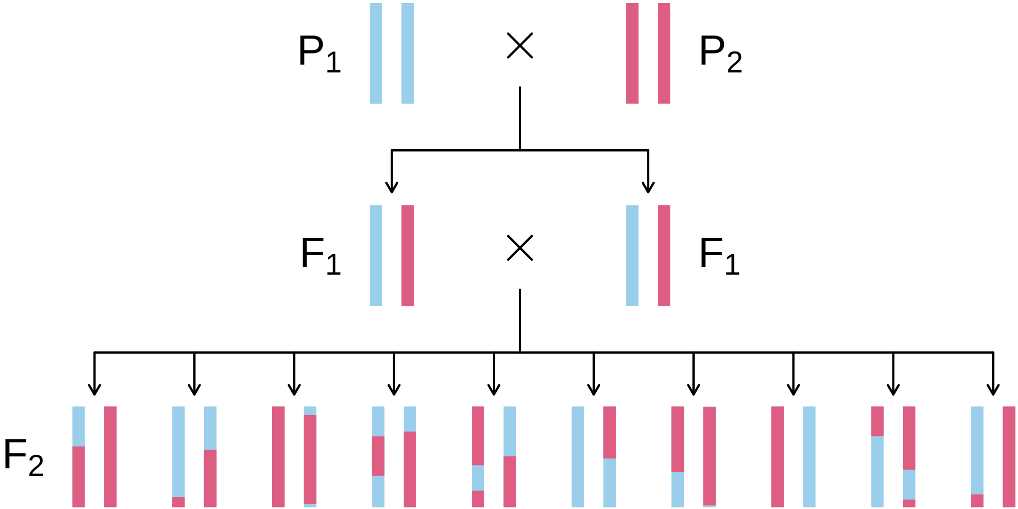
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

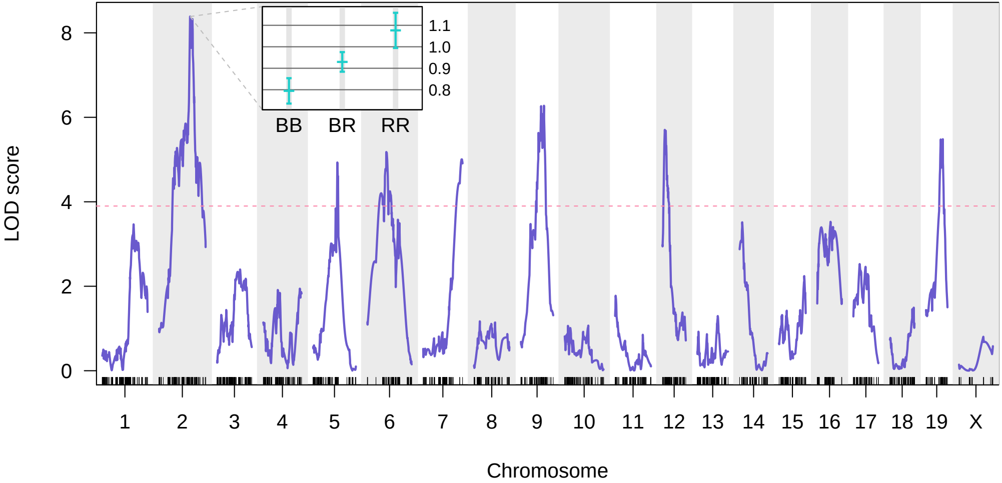
kbroman.org
github.com/kbroman
@kwbroman

Course web: kbroman.org/AdvData

Intercross

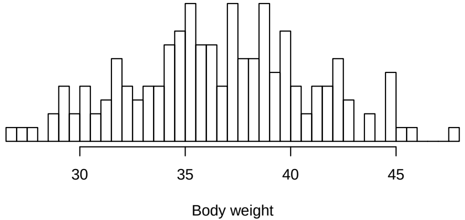


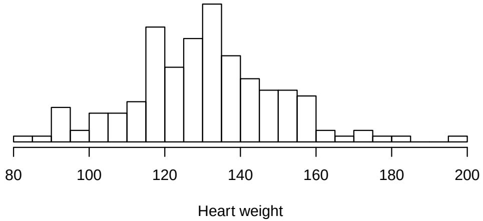
QTL mapping

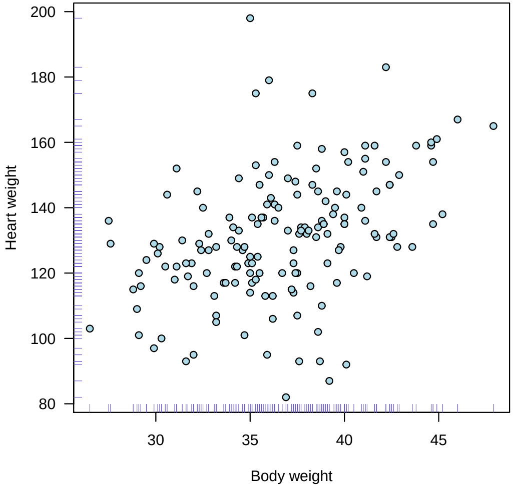


LOD score
8
6
4
2
0
1.1
1.0
0.9
0.8
BB
BR
RR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

Phenotype data

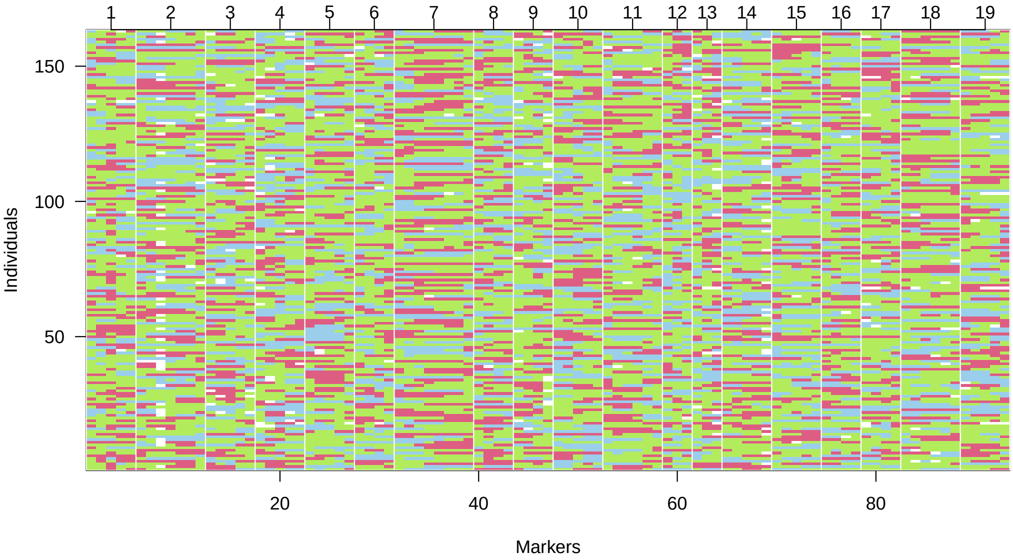




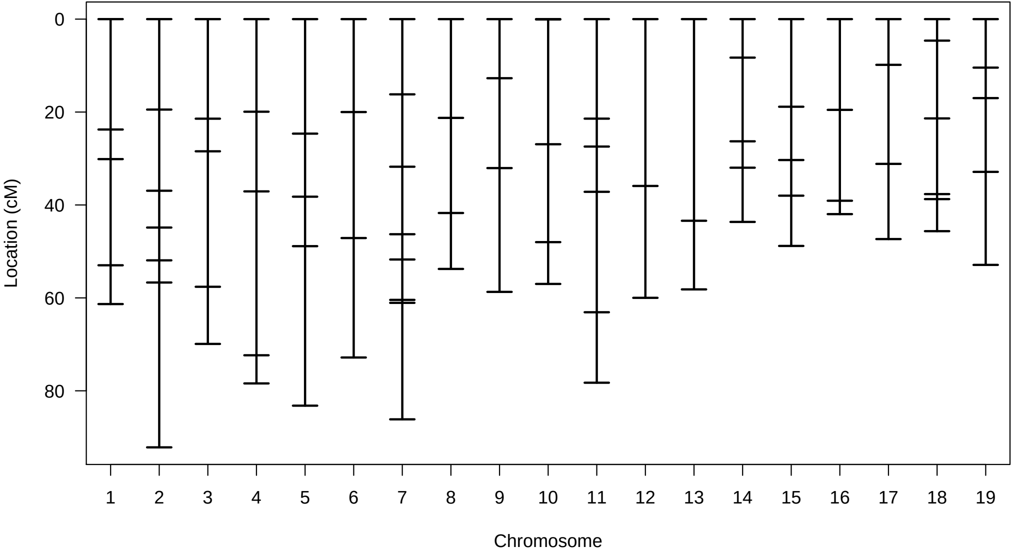


Sugiyama et al. (2002) Physiol Genomics 10:5–12

Genotype data

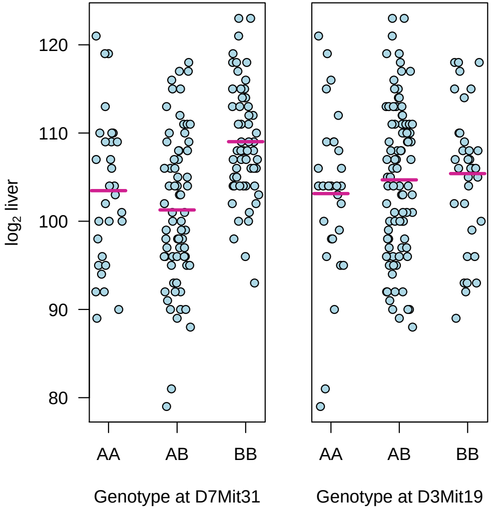


Genetic map


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

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.

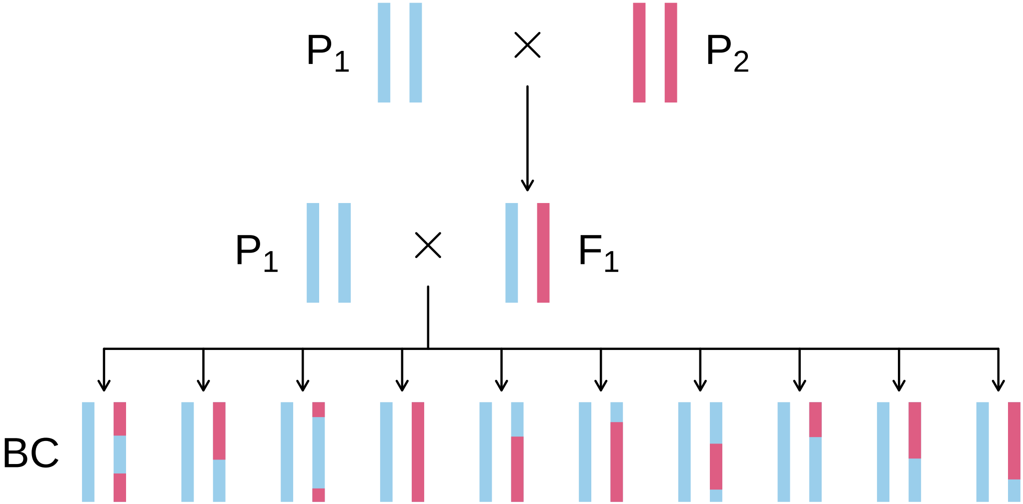
Interval mapping

Lander & Botstein (1989)

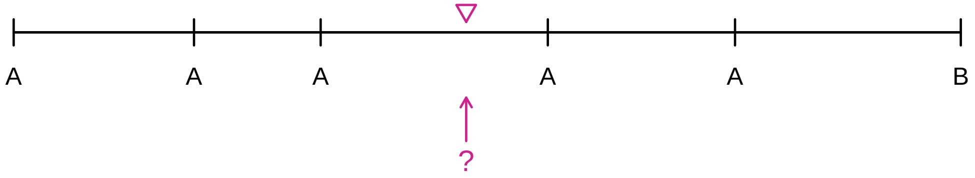
- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let $g = 0/1/2$ if the (unobserved) QTL genotype is AA/AB/BB.

  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



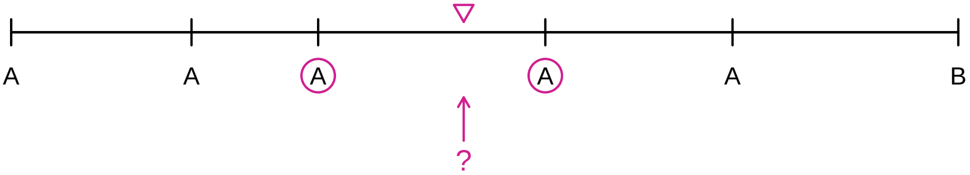
Calculate Pr(g | marker data), assuming

- 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



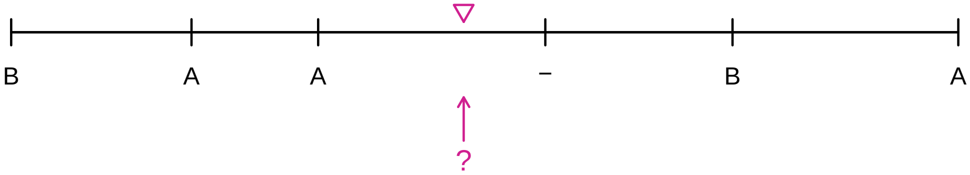
Calculate Pr(g | marker data), assuming

- 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



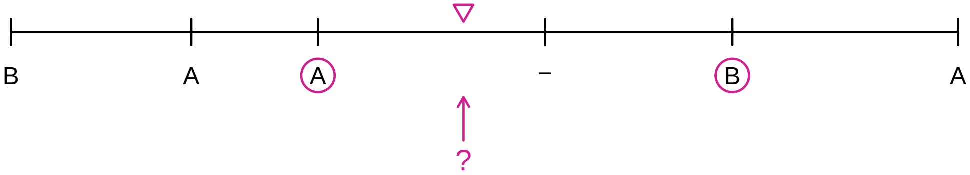
Calculate Pr(g | marker data), assuming

- 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



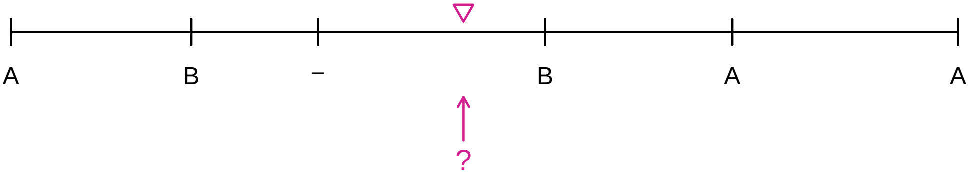
Calculate Pr(g | marker data), assuming

- 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



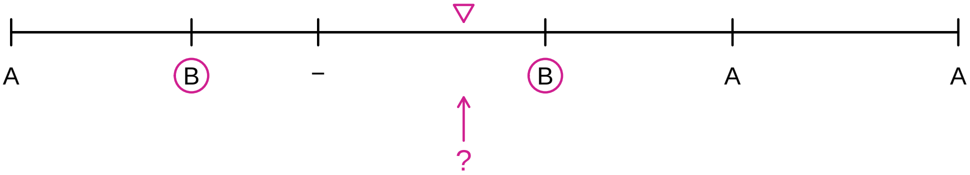
Calculate $\Pr(g \mid \text{marker data})$, assuming

- 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



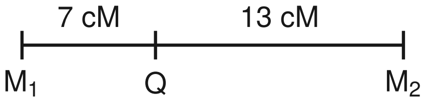
Calculate Pr(g | marker data), assuming

- No crossover interference
- No genotyping errors

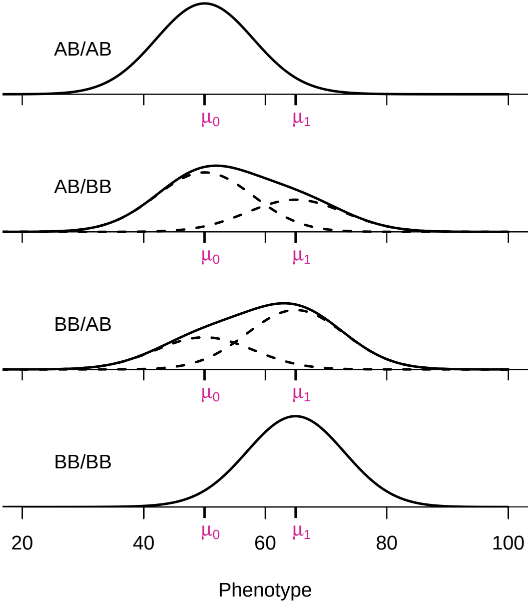
Or use the hidden Markov model (HMM) technology

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

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

$\mathsf{Pr}(\mathsf{y_i} | \text{marker data}, \mu_0, \mu_1, \sigma) = \sum_{\mathsf{j}} \mathsf{p_{ij}} \, \mathsf{f}(\mathsf{y_i}; \mu_{\mathsf{j}}, \sigma)$

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

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

Maximum likelihood estimates (MLEs) of μ_0, μ_1, σ:

values for which $\mathsf{l}(\mu_0, \mu_1, \sigma)$ is maximized.

EM algorithm

E step:

Let $$w_{ij}^{(k)} = \Pr(g_i = j | y_i, \text{marker data}, \hat{\mu}_0^{(k-1)}, \hat{\mu}_1^{(k-1)}, \hat{\sigma}^{(k-1)})$$

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

M step:

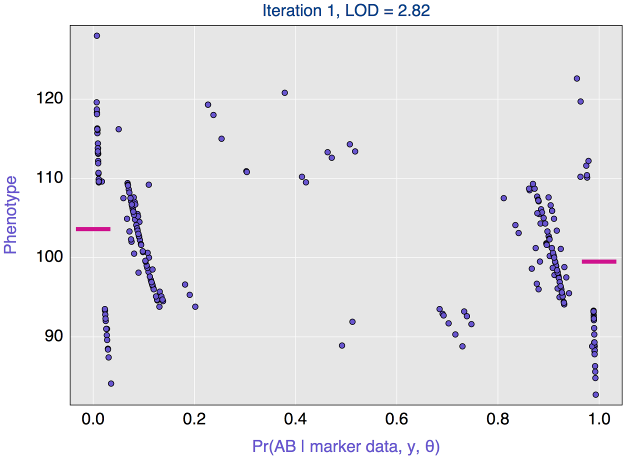
Let $$\hat{\mu}_j^{(k)} = \sum_i y_i w_{ij}^{(k)} / \sum_i w_{ij}^{(k)}$$

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

The algorithm:

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

Interactive illustration



Next

Back

`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.

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

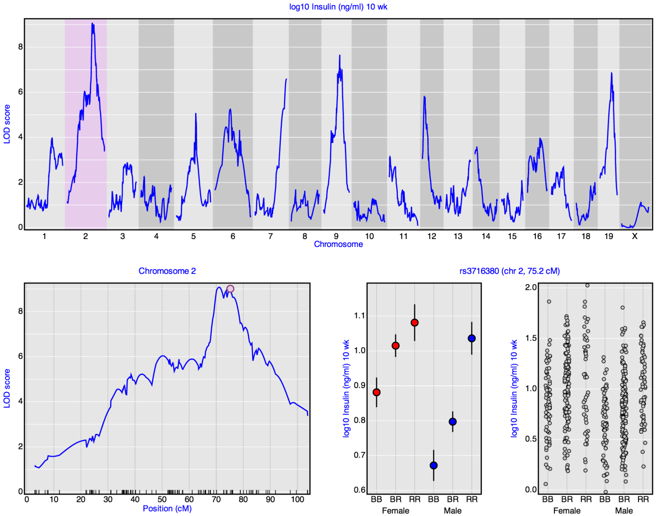
$$= \log_{10} \left\{ \frac{\Pr(y|\text{QTL at } \lambda, \hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda)}{\Pr(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

$\hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda$ 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).

Interactive plot



bit.ly/D3lod

Interval mapping

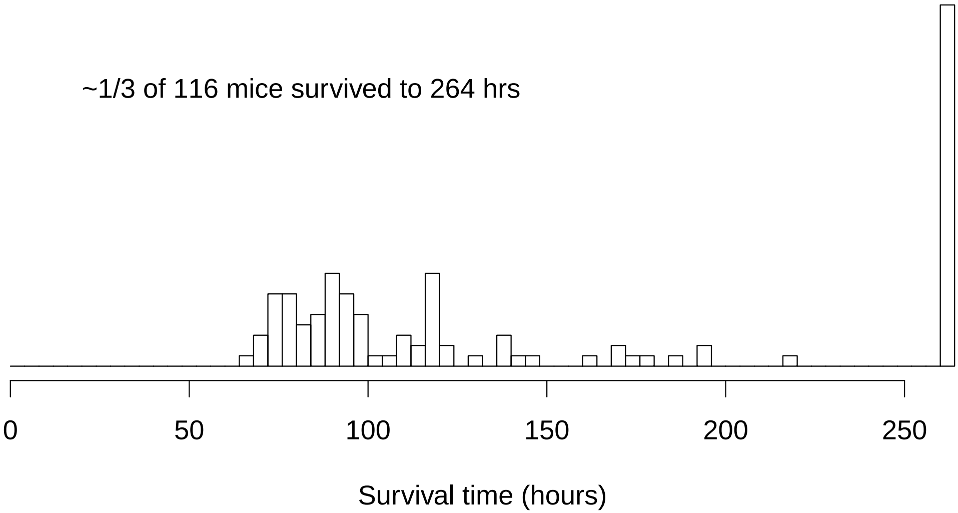
Advantages

- 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

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

M step

$$\hat{\pi}_j^{(s+1)} = \frac{\sum_i w_{ij}^{(s+1)} z_i}{\sum_i w_{ij}^{(s+1)}}$$

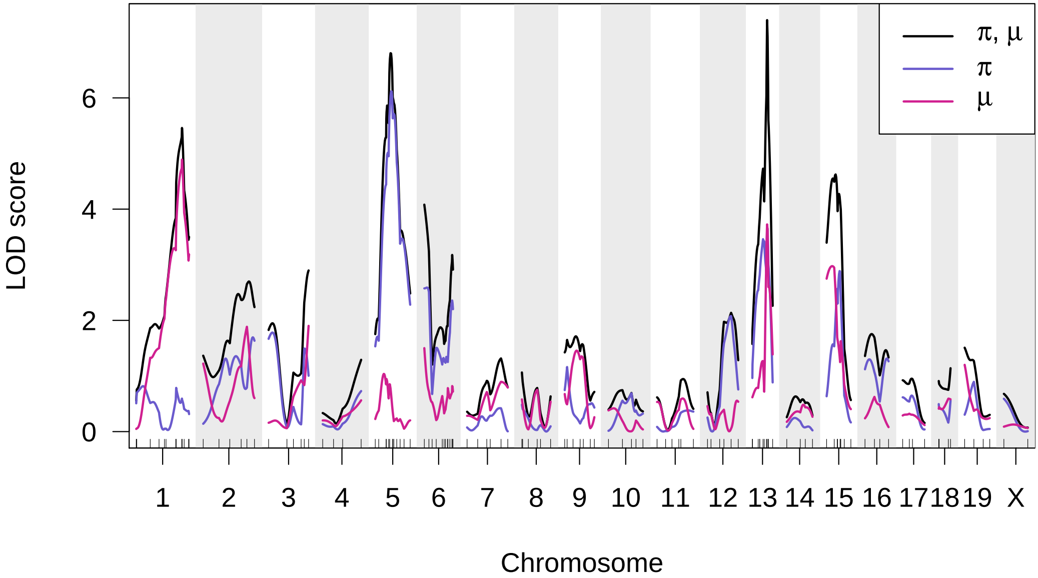
$$\hat{\mu}_j^{(s+1)} = \frac{\sum_i y_i w_{ij}^{(s+1)} z_i}{\sum_i w_{ij}^{(s+1)} z_i}$$

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

Tests

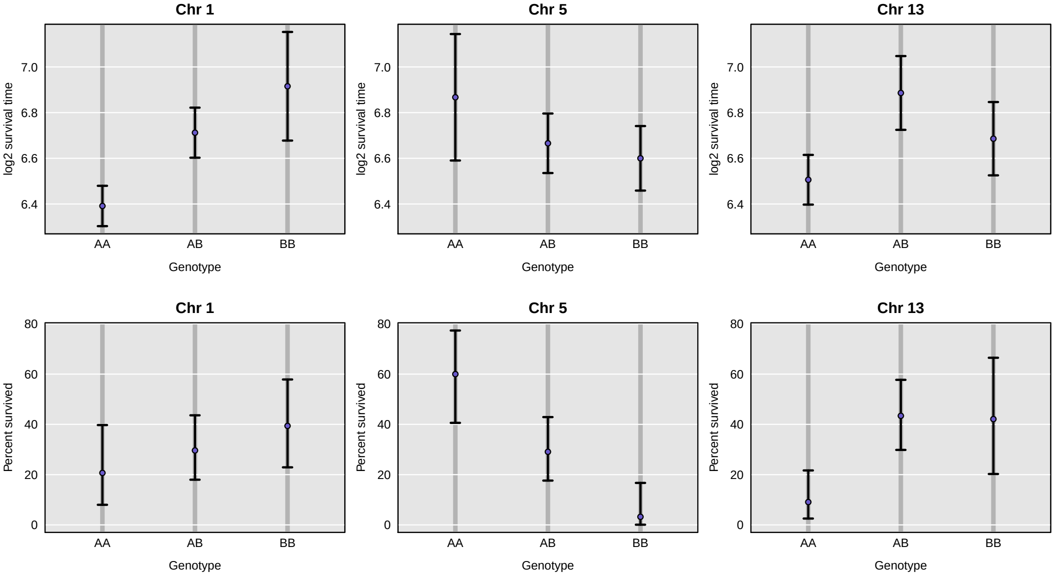
- $\pi_{AA} = \pi_{AB} = \pi_{BB}$
- $\mu_{AA} = \mu_{AB} = \mu_{BB}$
- $\pi_{AA} = \pi_{AB} = \pi_{BB}$ and $\mu_{AA} = \mu_{AB} = \mu_{BB}$

LOD curves



LOD score
6
4
2
0
π, μ
π
μ
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

QTL effects


Chr 1
Chr 5
Chr 13
log2 survival time
6.4
6.6
6.8
7.0
AA
AB
BB
Genotype
Chr 1
Chr 5
Chr 13
Percent survived
0
20
40
60
80
AA
AB
BB
Genotype

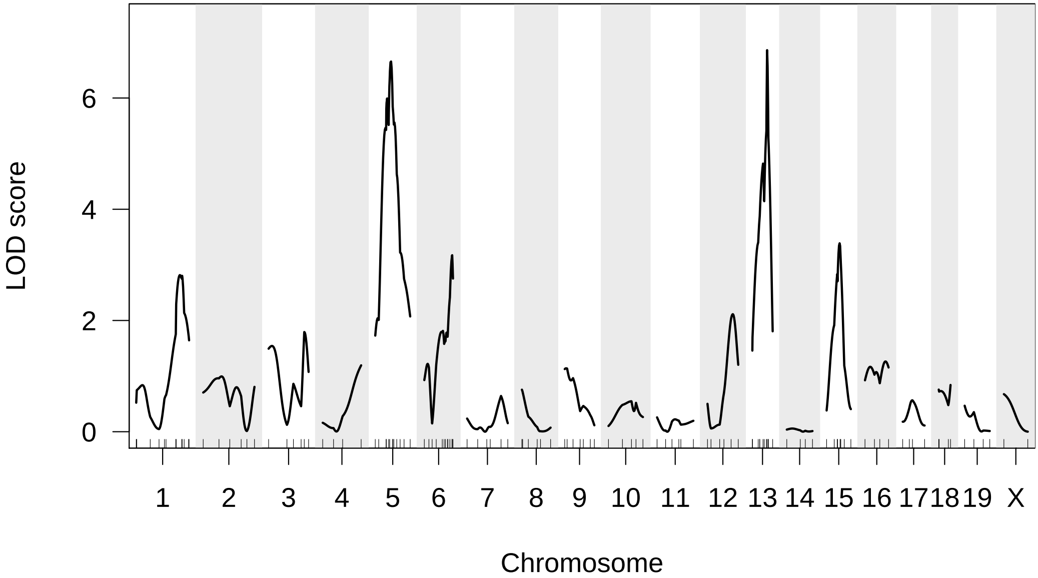
Lesson

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

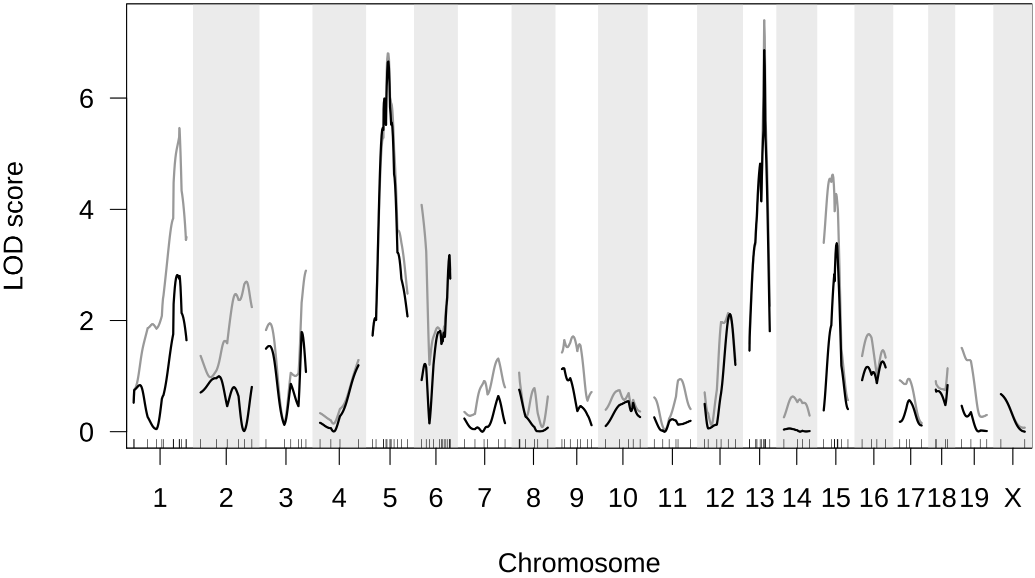
Lessons

- 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