The EM algorithm

QTL mapping with a cure model

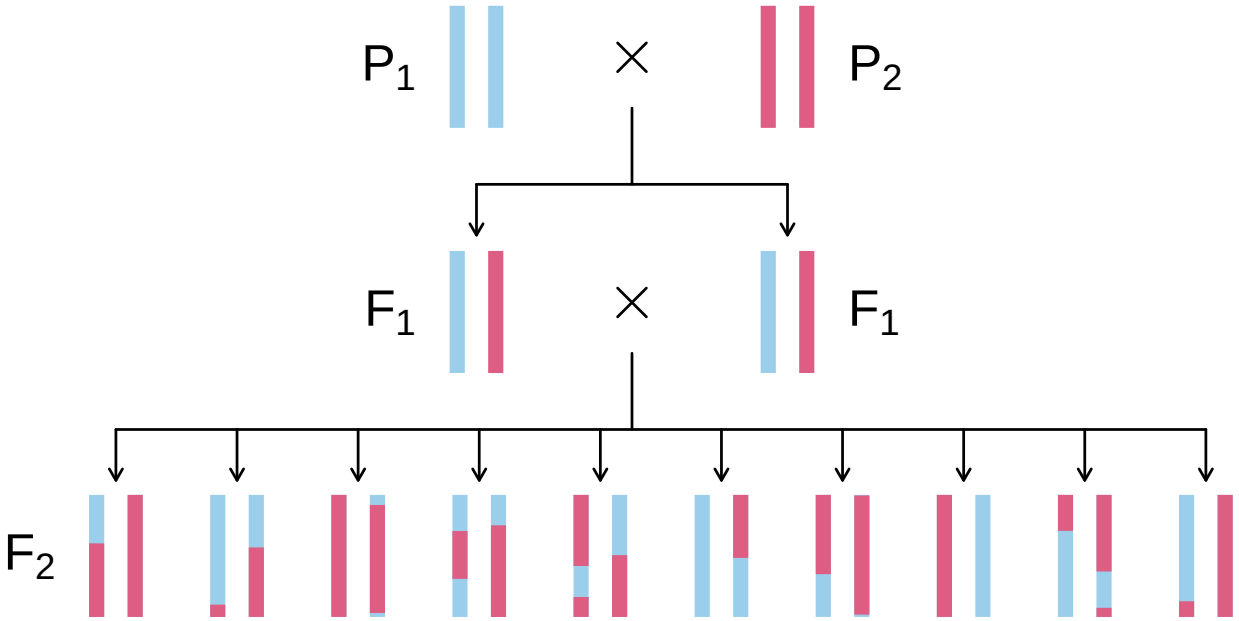
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

In this lecture, I'll illustrate a couple of examples of the EM algorithm, via QTL mapping and with a phenotype that shows a common feature: of having a spike in its distribution. We'll look at a way to deal with that problem.

Intercross

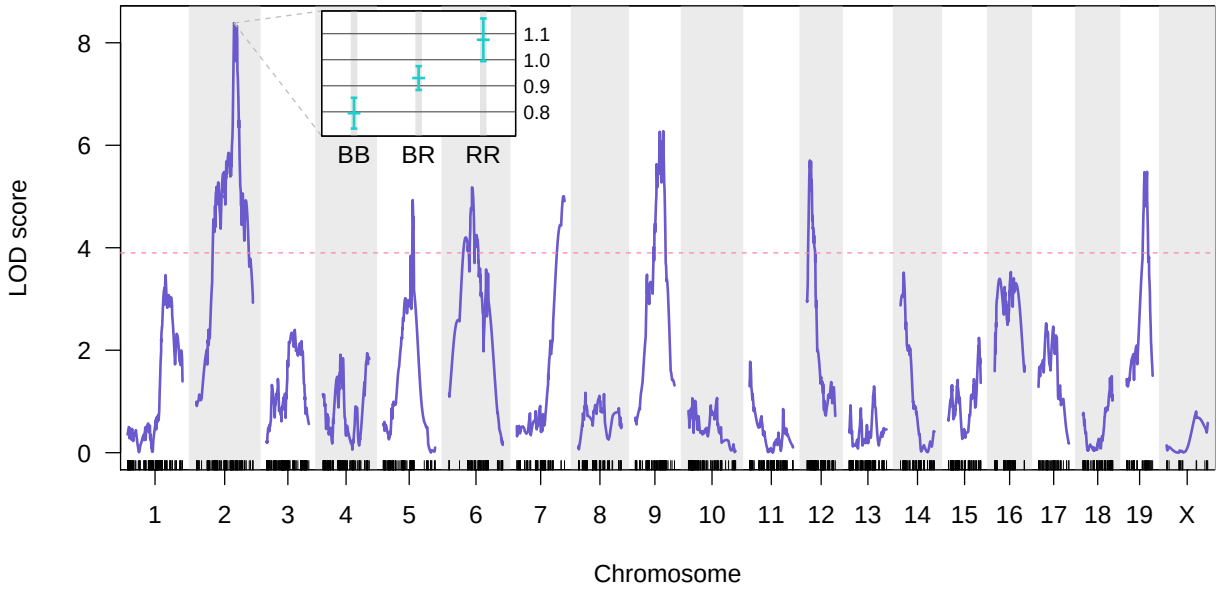


QTL mapping is an effort to identify the genetic loci that contribute to variation in some quantitative trait, like blood pressure. Such loci are called quantitative trait loci (QTL).

We start with two strains that differ in the trait of interest. That they show a consistent difference when raised in the same environment indicates that the difference is genetic. To try to identify genes contributing to the trait difference, we can perform a series of different crosses; the most common is the intercross.

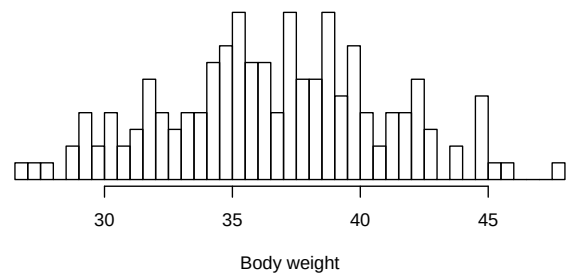
One gathers a number of intercross progeny, measures the trait, and then measures genotype at different positions along the chromosomes. We then look for positions where the genotype is associated with the phenotype.

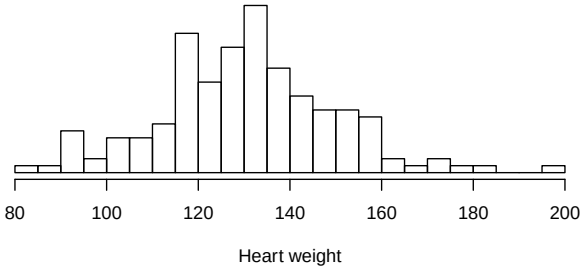
QTL mapping

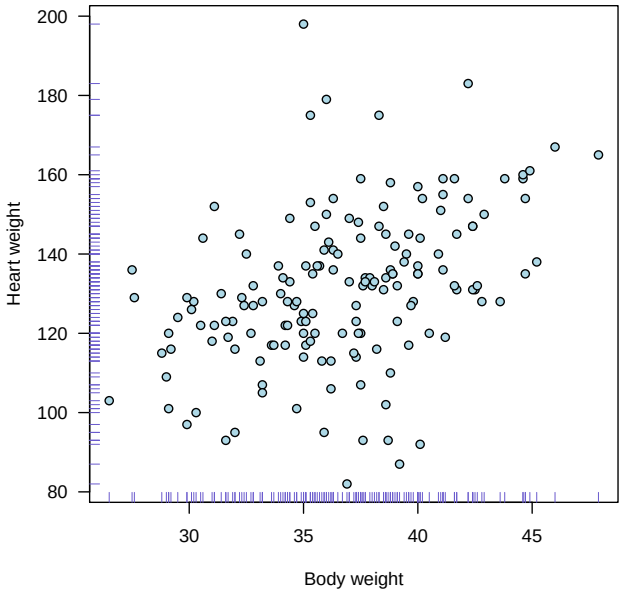


So we will scan across the genome, at each position calculating some test statistic. Regions where the test statistic is large will show an association between genotype and phenotype.

Phenotype data



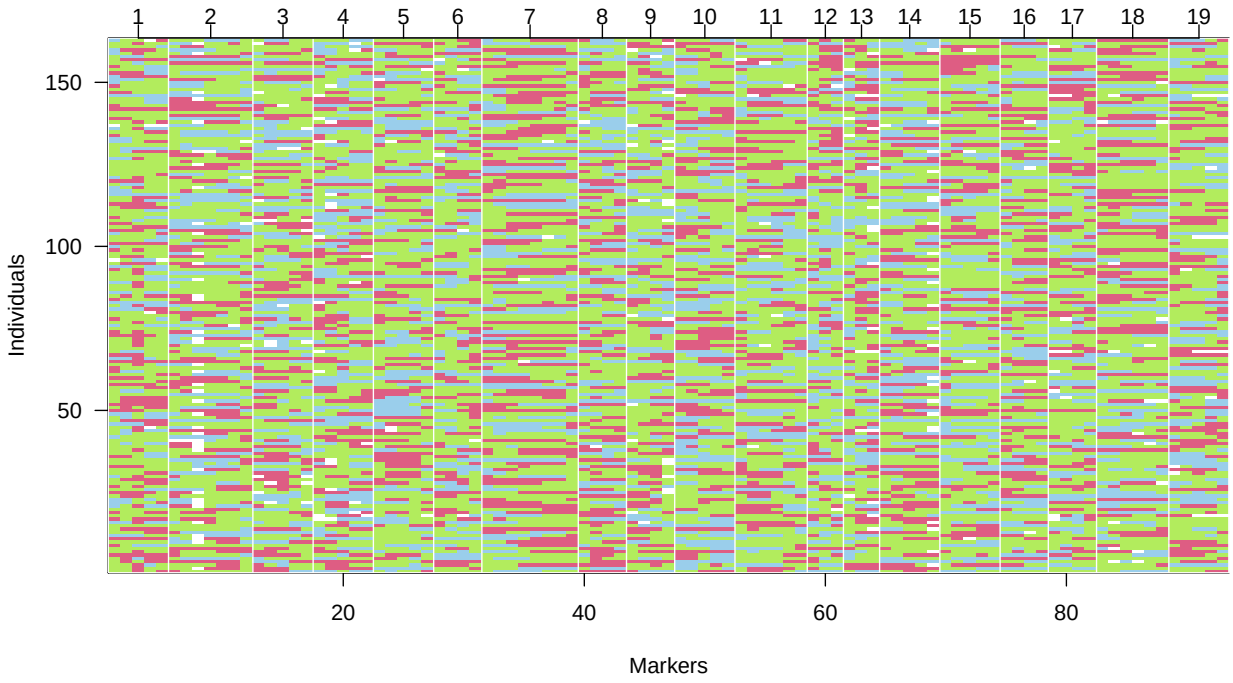




Sugiyama et al. (2002) Physiol Genomics 10:5–12

As an example, here are body weight and heart weight for 150 or so intercross mice.

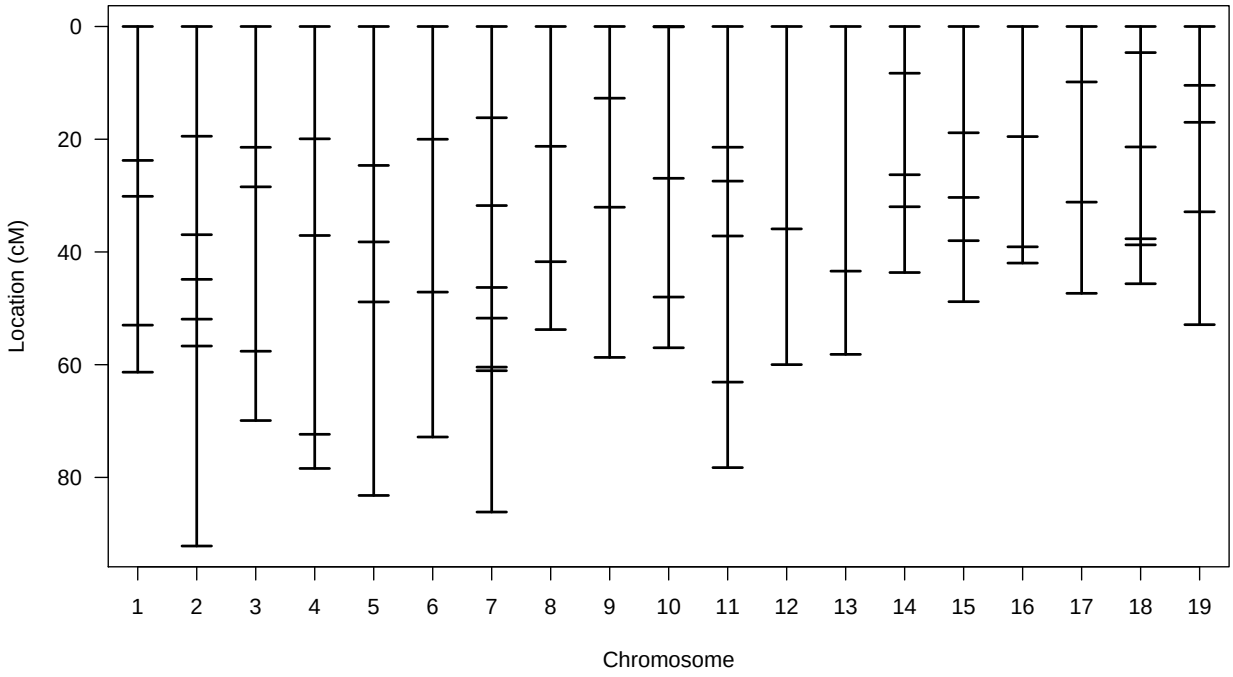
Genotype data



In addition to that phenotype data, we'll have genotype data like this. For each mouse, we'll have data at a variety of positions across the genome, indicating homozygous blue, homozygous pink, or heterozygous (shown in green).

One of the challenges is the missing data: the white pixels sprinkled into the data.

Genetic map

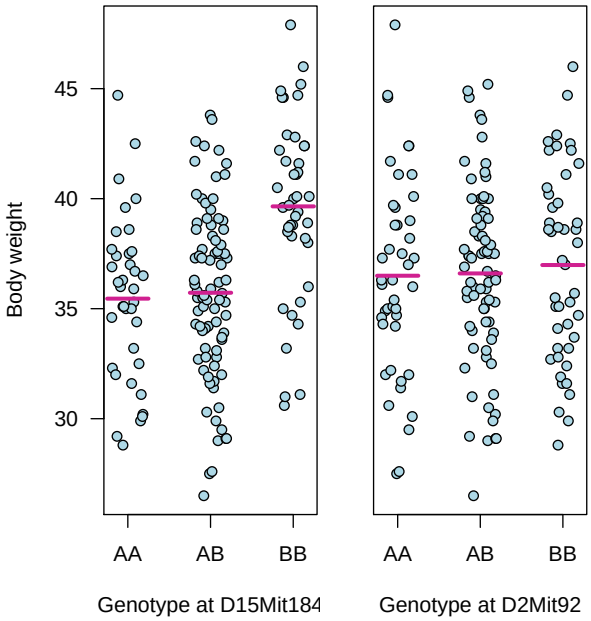


A genetic map will indicate the locations of the genetic markers along the chromosomes.

Our genotype data is rather sparse. On a given chromosome, we get to see genotype only at 3-8 positions, with some pretty large gaps between these positions.

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.



The simplest analysis of this sort of data is to look at each marker, one at a time, split the mice into groups according to their genotype, and do analysis of variance.

We look for markers where the different genotype groups show strong differences in average phenotype, as in the left panel for marker D15Mit184. We are not so interested in markers where there is little difference among the genotypes, as in the right panel for marker D2Mit92.

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.

Advantages of this approach are that it's simple and doesn't require specialized software. You can easily incorporate covariates, but replacing the ANOVA with linear regression and including extra covariates. You can extend the analysis to a more complex model, for example Cox proportional hazards when you have censored survival times. You're just looking for association between the phenotype and a categorical covariate. (Then repeat for each individual marker.) You also don't need a genetic map for the markers; you're just looking at them individually.

Disadvantages include that you need to omit individuals with missing genotype at a marker, and you get imperfect information about QTL location. The approach can suffer in low-density scans where the intervals between markers are large.

The biggest disadvantage is that you're considering just 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(μ_g, σ)
- Given genotypes at linked markers, y $\sim$ mixture of normal dist'ns with mixing proportions Pr(g | marker data)

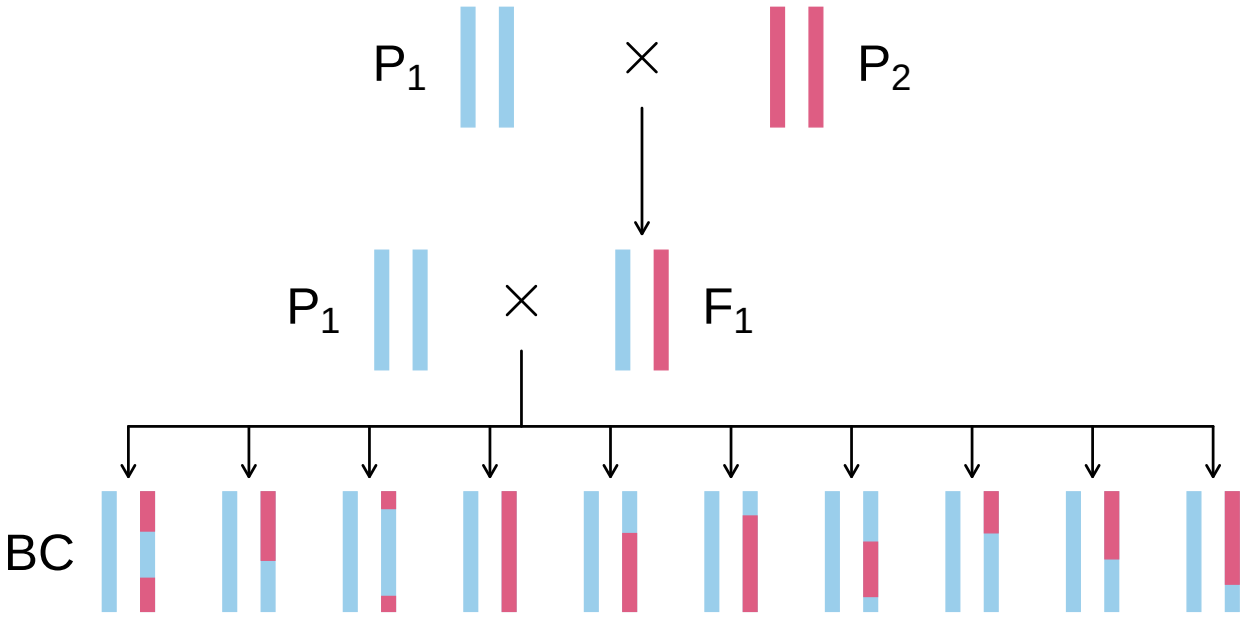
In 1989 Lander and Botstein came up with a method, now called "interval mapping," to solve the first three disadvantages of doing ANOVA at markers.

The idea is to use the same model as in ANOVA: that there is a single QTL that affects the trait and that the residual variation follows a normal distribution. But here, we allow the putative QTL to be at any arbitrary position. We'll consider each position in the genome, one at a time, as the QTL location, and scan along each chromosome.

The challenge is that the genotype at the putative QTL will not be known. However, we can use surrounding markers to calculate the probability of the different possible QTL genotypes, given the available marker data.

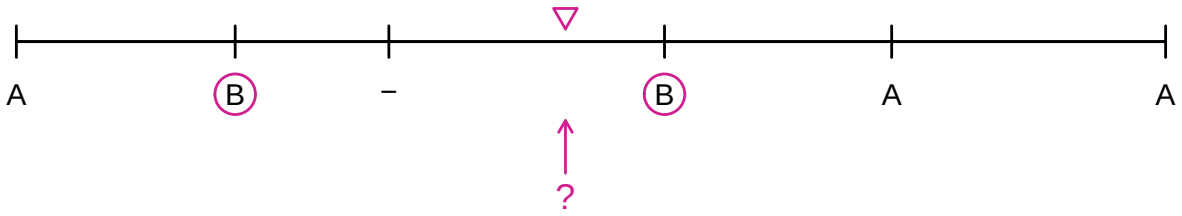
The phenotype, given QTL genotype, follows a normal distribution. Given the marker data (and not knowing QTL genotype), the phenotype follows a mixture of normal distributions with known mixing proportions.

Backcross



It's sometimes useful to think about a backcross, where you cross the two parents and then cross the F_1 back to one of the parents. All the offspring get a blue chromosome plus a recombinant chromosome, so at any one position they have one of two genotypes, homozygous blue or heterozygous. It can be easier to think about this, with just the two possible genotypes. For QTL mapping, we'd do the equivalent of a t-test.

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.)

A crucial ingredient to this interval mapping method is the probability of the different possible QTL genotypes, given the available marker data.

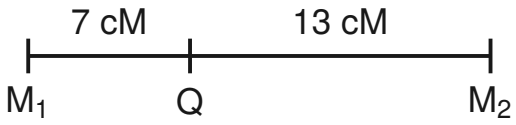
Consider a particular position for a QTL, such as the location of the triangle, and consider the available genotypes along the recombinant chromosome in a backcross.

A common assumption is that there is no crossover interference, meaning that the locations of exchanges are completely random, according to a Poisson process. So whether or not there is an exchange in one interval is indepedent of whether there is an exchange in any other interval.

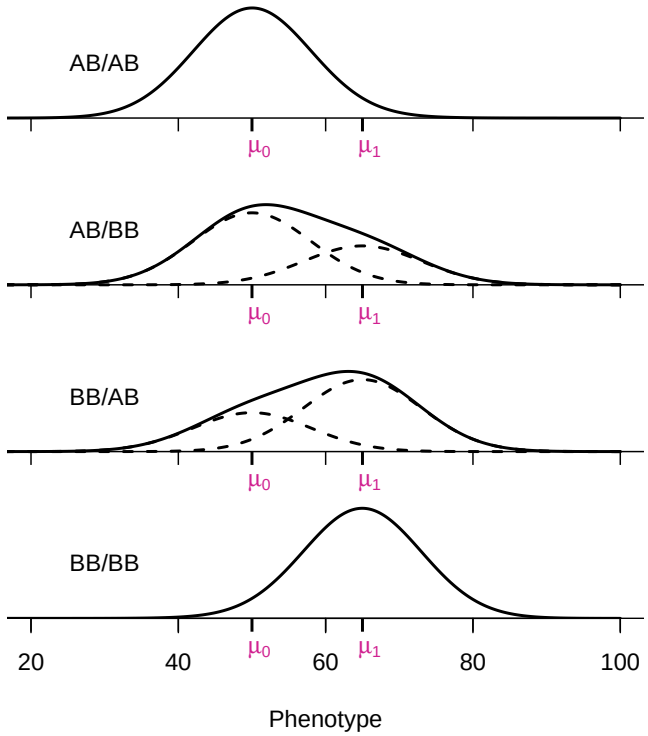
If we make that assumption, plus the assumption that the marker genotypes have no errors, then we just need to look at the two nearest typed markers, and we can do a little back-of-the envelope calculation to get the probability that the individual has an A allele at the QTL, given that they have B alleles at the two flanking positions.

We can relax the "no genotyping errors" assumption with a hidden Markov model, but the no interference assumption is super convenient, though totally wrong.

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.



Another way to think about these genotype probilities, is to think about the normal mixtures. Imagine a pair of markers in a backcross, with complete genotype data at the markers, and a QTL that's in the interval between them, a bit closer to the left marker.

In one sense, there are two kinds of mice: those that are AB at the QTL, and those that are BB at the QTL. But considering the marker data, there are four kinds for mice. The mice that are AB at both markers will mostly also be AB at the QTL; their phenotypes will follow one normal distribution. The mice that are BB at both markers will largely be BB at the QTL, and their phenotypes will follow another normal distribution, shifted over a bit.

The mice that are AB at the left marker and BB at the right marker will include a portion of mice that are AB at the QTL and follow the one normal distribution and another portion that are BB at the QTL and follow the other normal distribution; overall, their phenotypes will follow a blog-like distribution, the mixture of the two normal distributions.

Similarly, mice that are BB at the left marker and AB at the right marker will have phenotypes following a mixture of normal distributions.

Our goal will be to estimate the averages for the two distributions plus their common residual SD.

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.

So in interval mapping, we consider some single position for the one and only QTL, and then for each mouse we calculate the probabilities for each possible QTL genotype given the available marker data.

The phenotypes are assumed to follow a normal distribution, given QTL genotype. Given the marker data, they follow a mixture of normals, with known mixing proportions.

We can write down the log likelihood, which is a sum of log mixture probabilities, where those mixture probabilities are a sum over the individual components.

The parameters to estimate are the phenotype averages for each QTL genotype, plus the residual SD. We seek the MLEs, which are the values for which the log likelihood is maximum.

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:

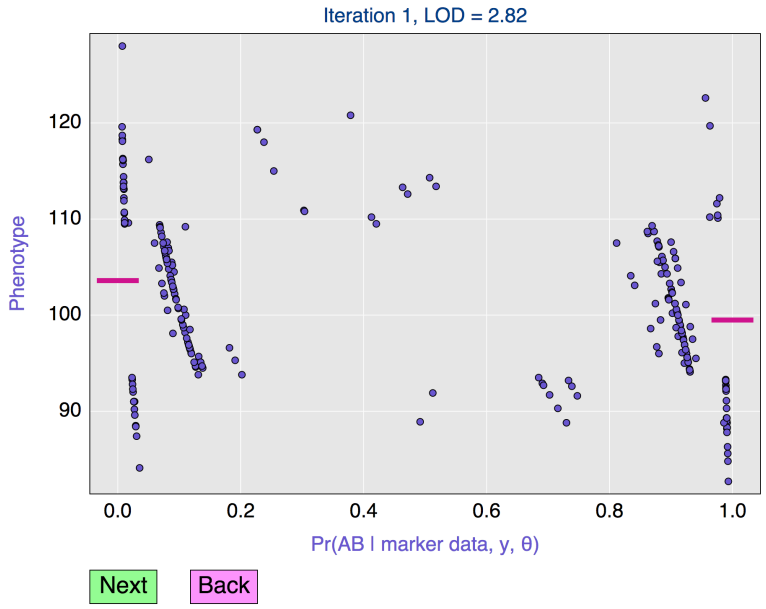
Start with $w_{ij}^{(1)} = p_{ij}$; iterate the E & M steps until convergence.

The EM algorithm is super for this problem, as if we knew the QTL genotype for each mouse, we could calculate the MLEs immediate as the phenotype averages for the mice grouped by QTL genotype, and then the residual SD.

The EM algorithm will alternate between an E step where we calculate probabililties of each possible QTL genotype given not just the marker data but also the phenotype, and using the current estimates of the parameters, and then the M step where we re-estimate the parameters. Here the new estimates of the means will be weighted averages of the phenotypes, with the weights being the conditional probabilities from the E step. The residual SD will be estimated using a weighted sum of residuals.

Rather that start with initial estimates, we actually start by jumping into the M step, using our initial QTL genotype probabilities as the weights.

Interactive illustration



`bit.ly/em_alg`

I've prepared an interactive graph at `bit.ly/em_alg` that illustrates the EM algorithm in this case.

On the y-axis are the phenotypes. And on the x-axis are the initial probabilities for the AB genotype given the marker data. When you click "next" it will step through the iterations of the EM algorithm. Using the available phenotype data, and with the estimated genotype-specific averages indicating that mice on the left (with BB genotype) have higher average phenotype than mice on the right (with AB genotype), the probabilities get titled toward the left. Mice with larger phenotypes get moved a bit to the left, and mice with smaller phenotypers get moved a bit to the right.

After 4 iterations, there's no perceptible further movement.

LOD scores

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

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

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

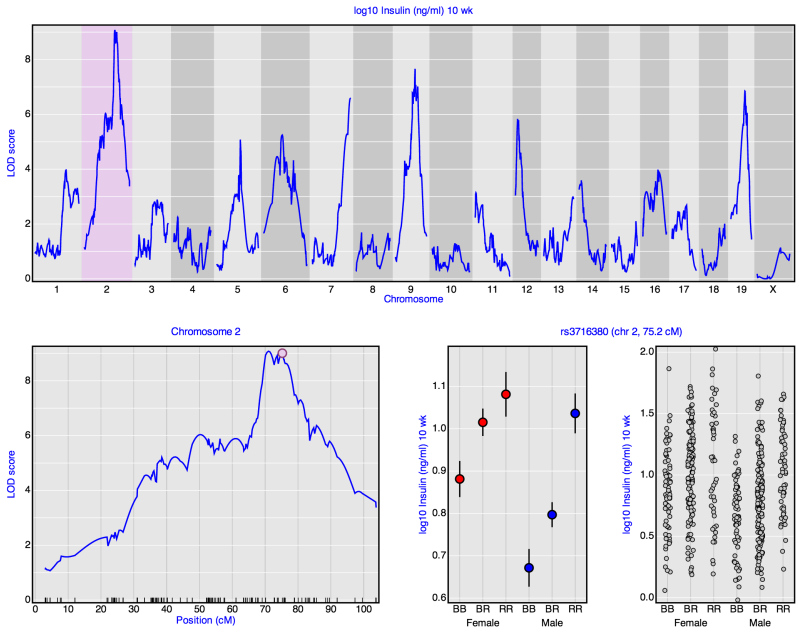
$\hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda$ are the MLEs, assuming a single QTL at position λ.

No QTL model:

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

For historical reasons, we measure evidence for QTL using LOD scores: the $\log_{10}$ likelihood ratio comparing the hypothesis of a single QTL at the given location to the null hypothesis of no QTL.

Interactive plot



bit.ly/D3lod

Plotting these LOD scores across the genome was perhaps the biggest innovation of Lander and Botstein (1989).

I've created an interactive graph to illustrate these LOD curves. In this case, we are also splitting the mice by sex. Click on a chromosome at the top and you'll see a blown-up view on the bottom left. Click on different positions along the chromosome and you can see the phenotype:genotype relationship.

It's important to look at the phenotype:genotype associations and not just the summary LOD curves. While for the locus on chromosome 2, the RR genotype is associated with higher insulin level (which corresponds to what is seen in the B and R founder strains), if you click on chromosome 9 you'll see that there the effect is in the opposite direction. And if you look at chr 7, you'll see that the effect is much larger in males than in females.

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

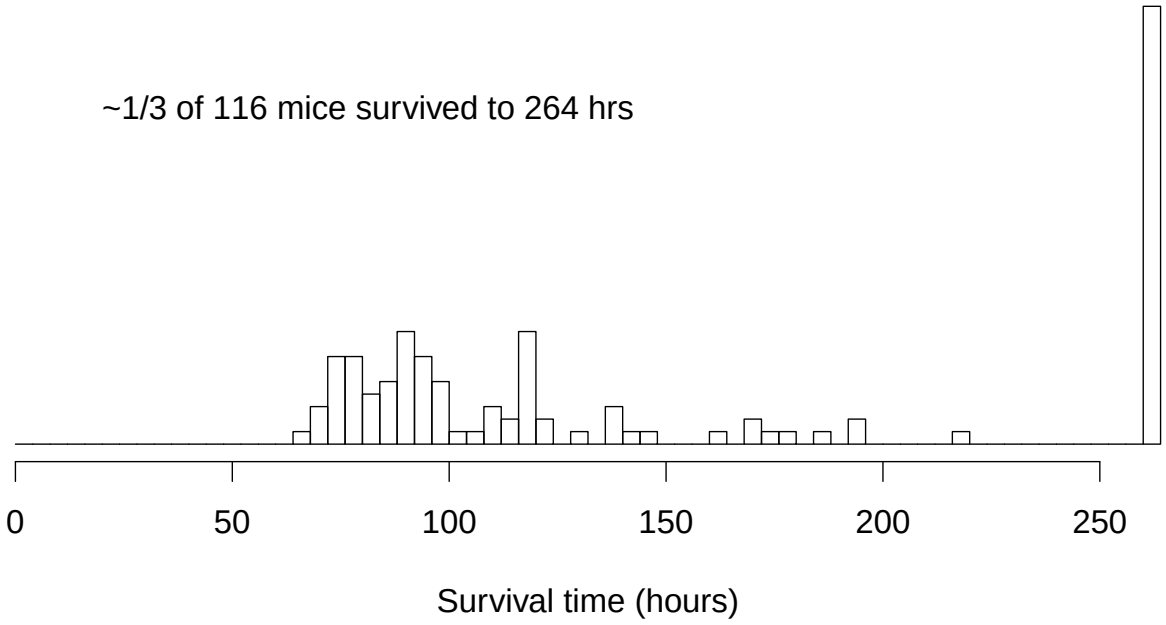
- Increased computation time.
- Requires specialized software.
- Difficult to generalize.
- Only considers one QTL at a time.

This interval mapping approach solves many of the weakness of just doing ANOVA at the markers: it takes account of missing data, it allows you to inspect positions between markers, at it gives improved estimates of QTL effects because you're looking directly at the QTL position. The biggest innovation, though, is graphing the test statistic across the genome.

Disadvantages include that it requires increased computation time, it requires special software, and it can be difficult to generalize. These aren't so important anymore.

The main disadvantage is that you're still just thinking about a single QTL. If you have reasonably dense markers and reasonably complete genotype data, this is really no different than just doing ANOVA with each individual marker.

Survival after Listeria infection



Okay, now to the main event for this lecture.

A collaborator asked me about QTL mapping with survival time after infection with Listeria, where about 1/3 of the mice survived to 10 days (264 hours) and at that point had basically cleared the infection.

Here, the phenotype distribution clearly doesn't fit the normality assumption that underlies ANOVA. What should we do?

Normal assumption in ANOVA

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

ANOVA is remarkably robust to non-normality. But alternatives to deal with non-normality include transforming the outcome (like taking square-roots or logs), using rank-based methods, or switching to specially-tailored models like generalized linear models (GLM).

Censoring?

Another thought here is whether to treat those surviving mice as censored. But they are not really censored in the usual sense, but rather had simply been cured of the infection.

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

This sort of phenotype distribution, particular the case with a spike at 0, is really quite common.

Consider the mass of gallstones, where some subjects have no gallstones. Or gene expression, where a gene might be turned off. Or microbiome data, where a particular microbe might be completely absent in some subjects.

Outside of biology, I like the example of "square feet of the garage" for a house, where some houses have no 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

My solution to this problem was to split the trait into two parts: a binary trait (survived or not) and then the quantitative survival time, if it didn't survive.

Given QTL genotype g, we have some probability of surviving, and then if it didn't survive, its survival time follows a normal distribution with some mean depending on the survival time.

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

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

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

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

We can expand our EM algorithm for QTL analysis to include this two-part, "cure" model. In the E step, we're still trying to get the probability a mouse has QTL genotype *j* given its marker data, its phenotype, and given the current estimates of the parameters. We treat the surviving and non-survivng mice a bit differently.

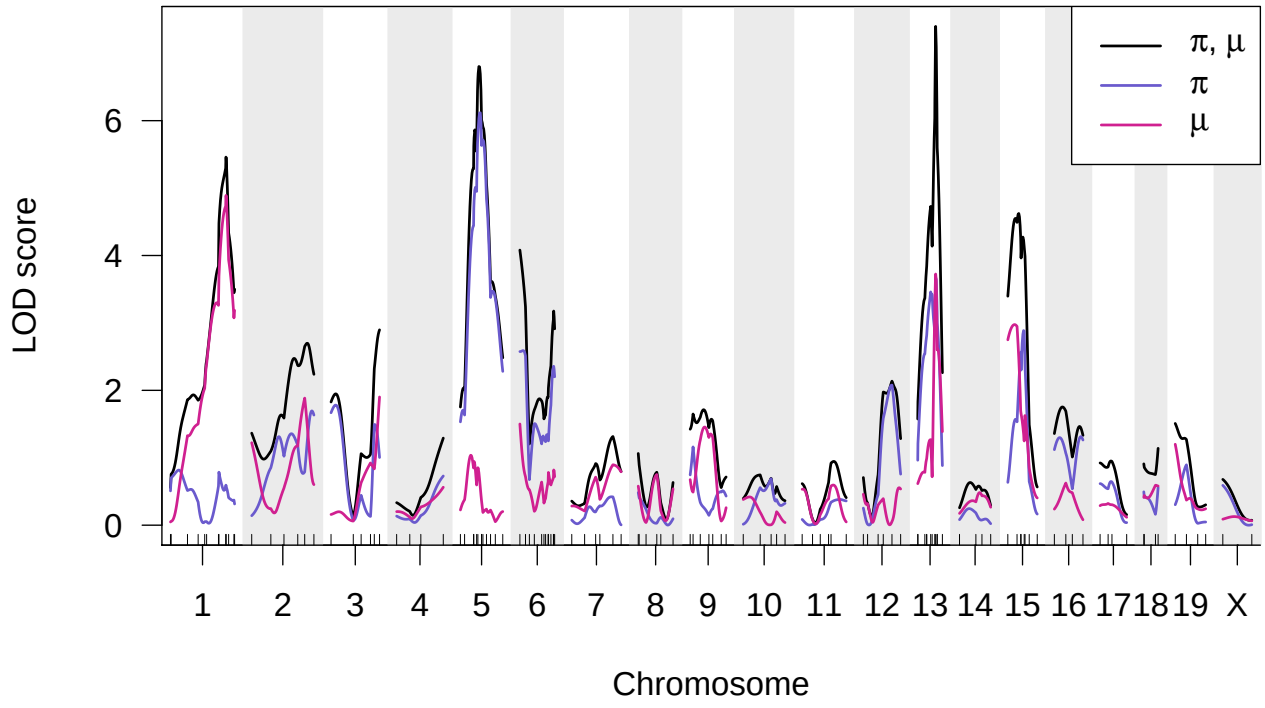
Then in the M step, we get updated probabilities of surviving as weighted proportions of surviving individuals. The updated estimates of the average survival times in each genotype group are weighted averages of the survival times.

Tests

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

And we can do three separate statistical tests. We can test whether the QTL has any effect at all, or we can look just at the probabilities or just at the survival time means.

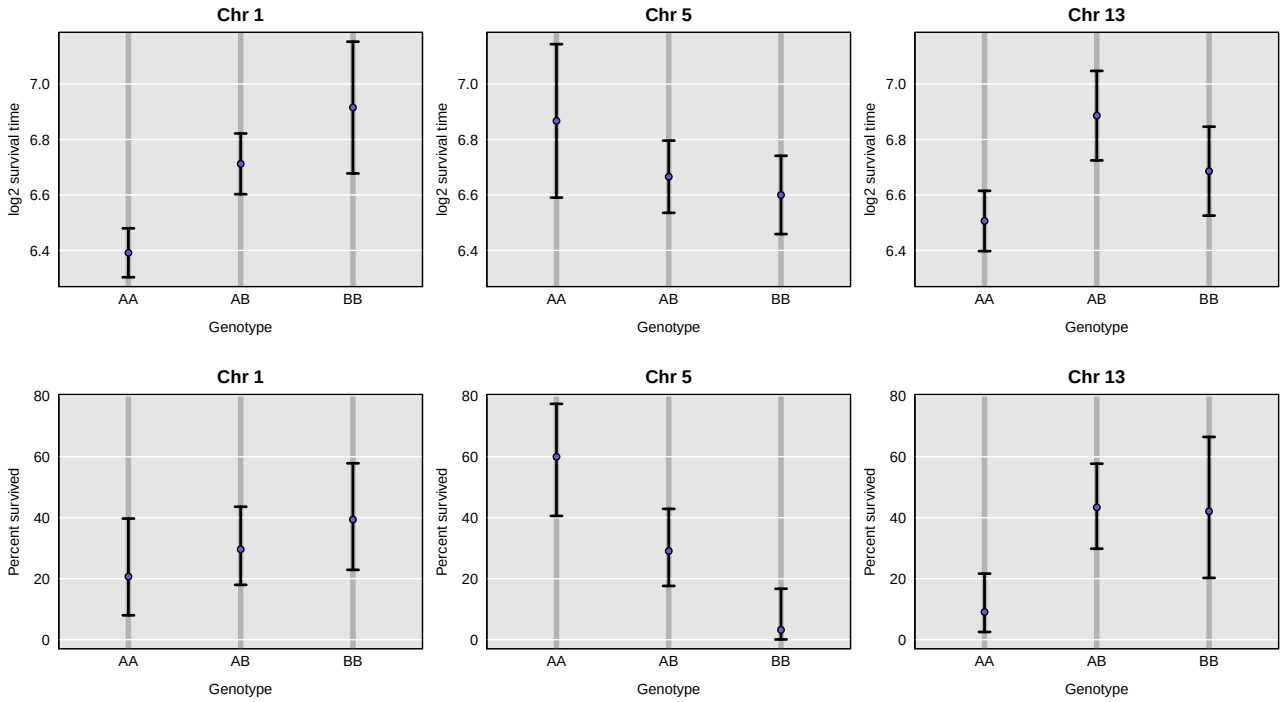
LOD curves



And so here are the results for these data. The black curves are for the overall tests, combining effects on probability of survival and on survival time. The blue curves look just at probability of survival, and the red curves just look at the survival times.

It looks like the QTL on chr 1 largely affects survival time, while the QTL on chr 5 largely affects the probability of survival. The loci on 13 and 15 seem to affect both aspects.

QTL effects



This shows up when you look at the estimated QTL effects on log survival time (top panels) and probability of survival (bottom panels) for the QTL on chr 1, 5, and 13.

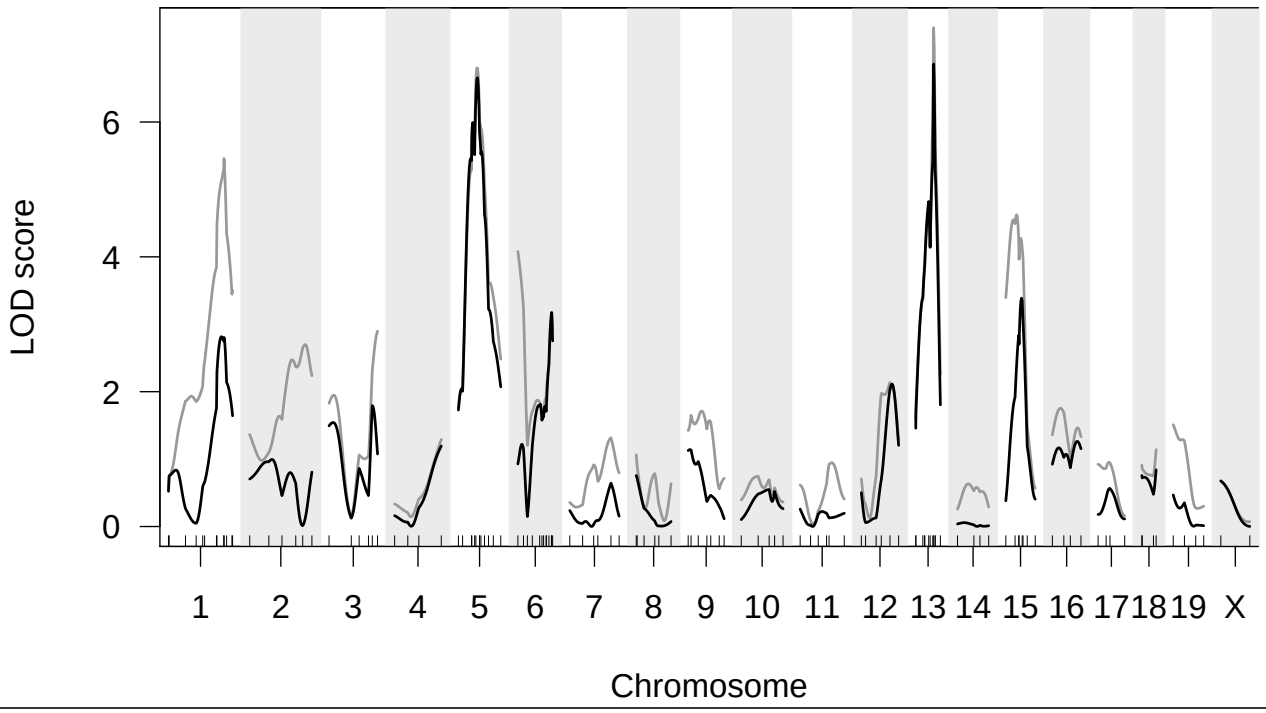
Lessons

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

The lesson here is, don't just cram your data into the standard approach.

A second lesson is, well actually you might be able to just cram your data into the standard approach.

Standard approach



Here are the results with the standard approach. In gray is the new fancy two-part method. The loci on 5 and 13 are still clearly seen, you just lose out on the chr 1 locus.

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

Here are some papers related to the work I've discussed today.

Lander and Botstein (1989) introduced this idea of interval mapping of QTL, and the use of the EM algorithm to deal with missing genotype data in this context.

The second is a gentler review of QTL mapping.

The last two papers concern this "spike in the phenotype distribution" situation.