Building regression models

Mapping multiple QTL

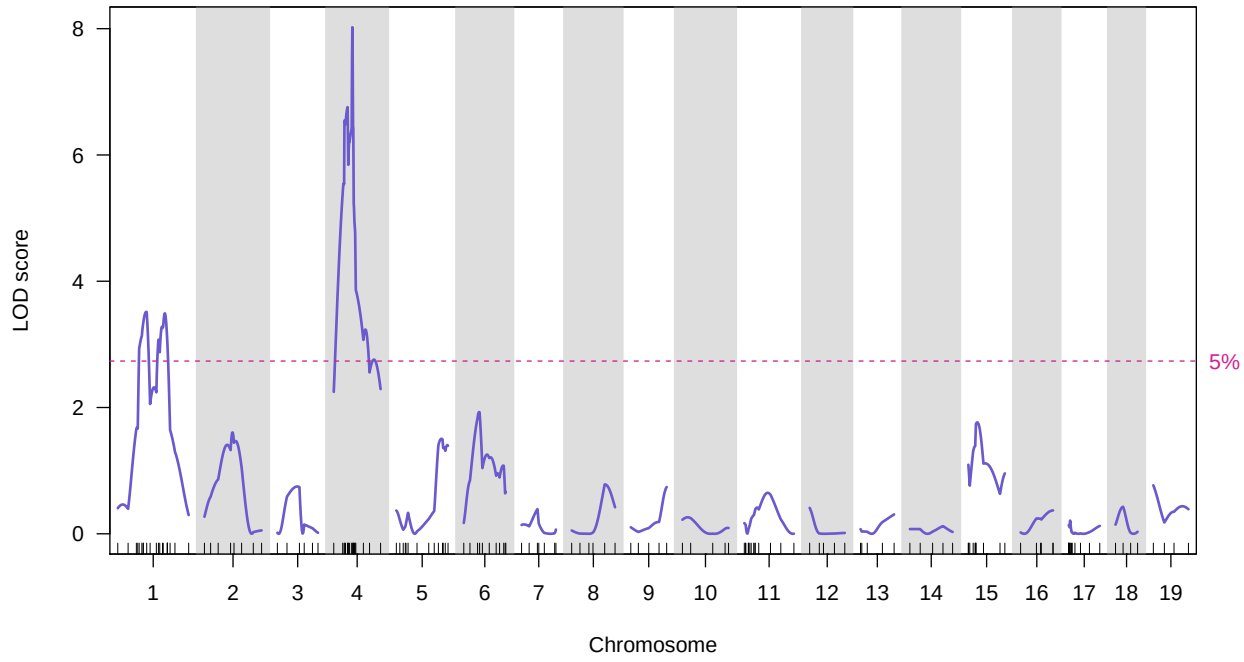
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

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

In this lecture, we'll look at building regression models, and particularly at the case of trying to map multiple QTL. Our interest here is not in prediction but rather in identifying the important variables.

LOD curves



I've talked a lot about QTL mapping: seeking to identify the set of genetic loci in the genome that affect some quantitative trait, by doing an experimental cross and then performing ANOVA at each locus, one at a time.

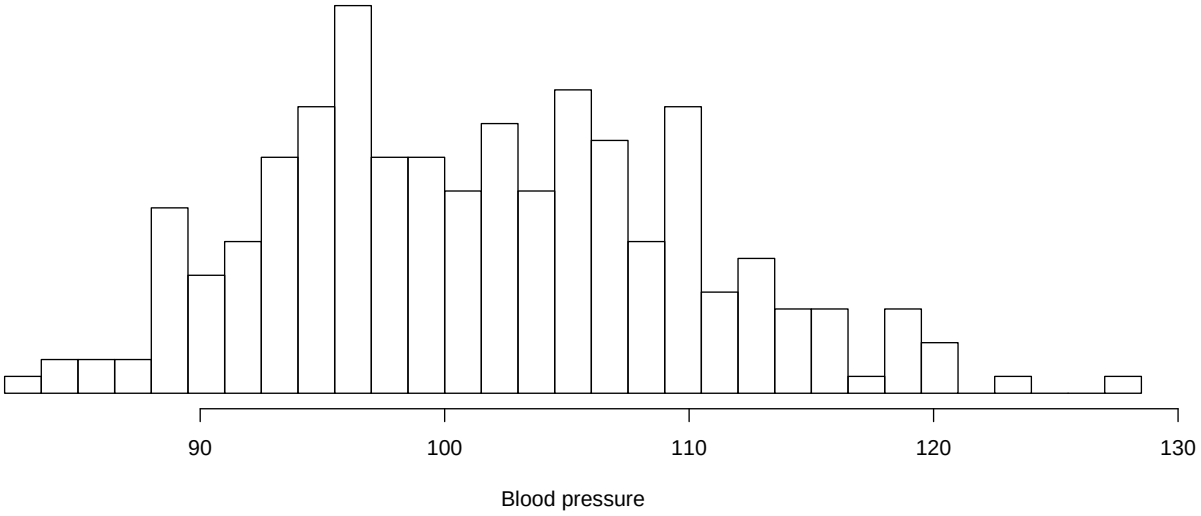
This basic analysis considers each predictor in isolation, though we expect there to be multiple loci that are influencing the trait.

Today, I want to talk about efforts to go after multiple loci at once.

Example

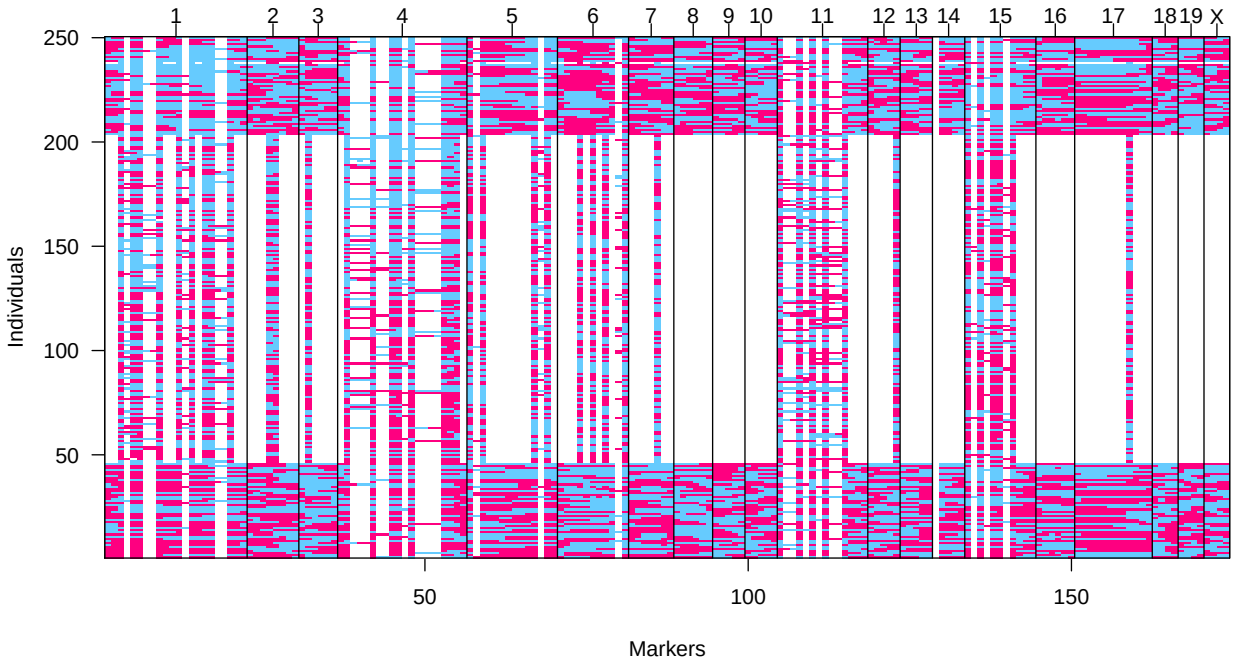
Sugiyama et al. Genomics 71:70-77, 2001

- 250 male mice from the backcross (A $\times$ B) $\times$ B
- Blood pressure after two weeks drinking water with 1% NaCl



As an example, I'll focus on this backcross of 250 mice, with the outcome being blood pressure. At each position, mice have one of two genotypes.

Genotype data



I like this example because the genotyping strategy makes the analysis quite difficult. The mice in this figure are sorted by blood pressure. There was a selective genotyping approach, in which only the 46 mice with highest blood pressure and the 46 mice with lowest blood pressure were genotyped. But then they did an initial analysis and genotyped all the mice at regions of potential interest. And then they further genotyped apparent recombinant individuals in selected regions, to try to further hone in on QTL locations.

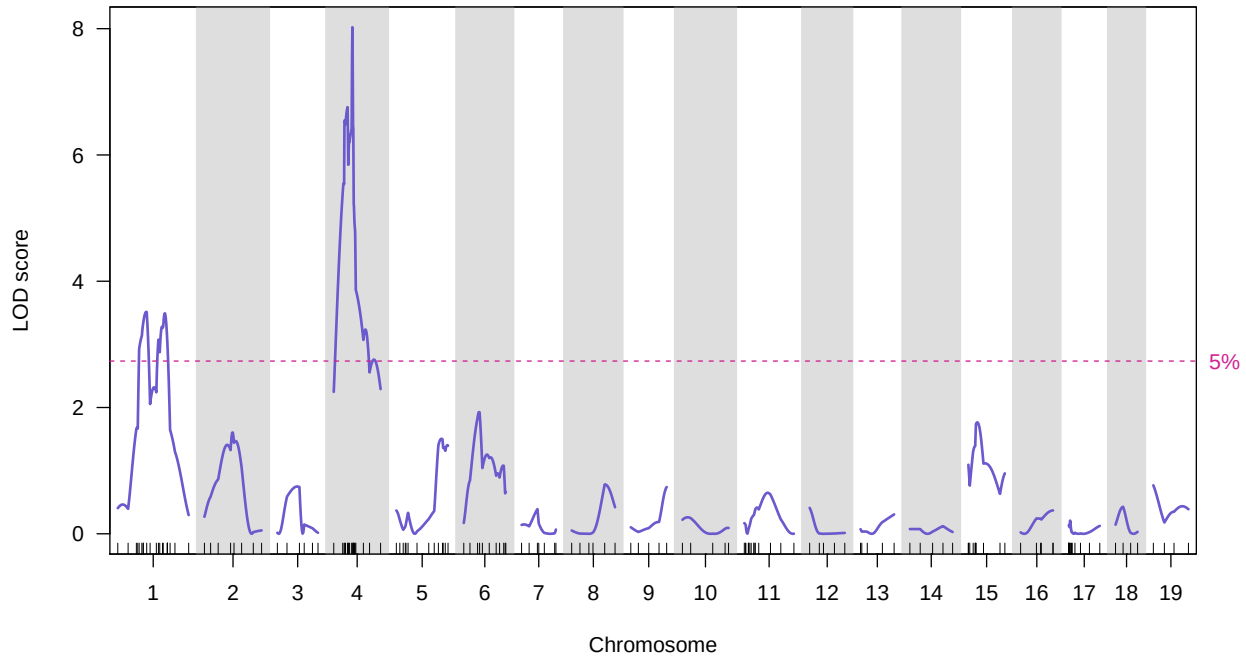
Goals

- Identify quantitative trait loci (QTL)

  (and interactions among QTL)
- Interval estimates of QTL location
- Estimated QTL effects

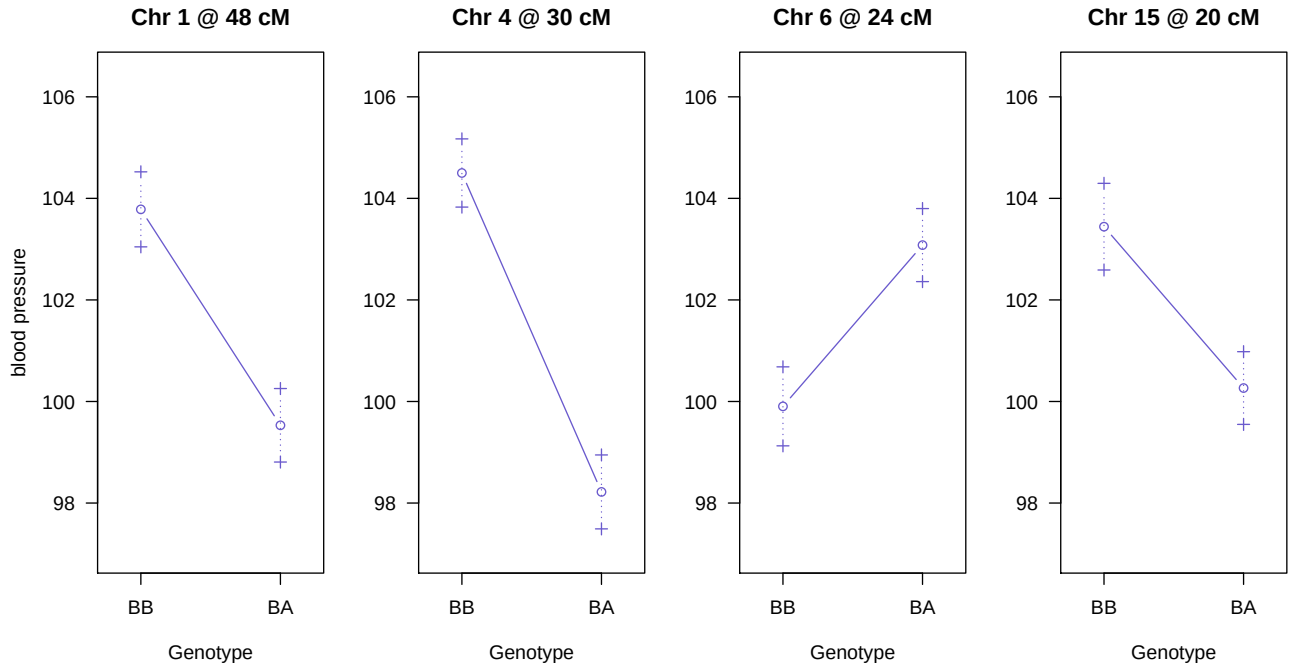
Our goals are first to identify QTL and potentially interactions among QTL, and secondly to get interval estimates of their location and to further estimate their effects. But we're going to focus solely on the first of these.

LOD curves



The genome scan results for this example shows a strong QTL on chromosome 4, and a double-humped QTL on chromosome 1. Next in line are QTL on chr 6 and 15, but both are quite far below the genome-wide threshold for statistical signifance.

Estimated effects



Here are the estimated effects of these loci. For the QTL on chromosomes 1, 4, and 15, the A allele leads to lower blood pressure. This matches the parental strains' traits: the A strain has lower blood pressure than the B strain.

On chr 6, though, the effect is in the opposite direction: the A allele causes an increase in blood pressure. This is called a transgressive QTL. The A strain has lower blood pressure, but it appears to include alleles that cause increased blood pressure, relative to the B allele.

Modeling multiple QTL

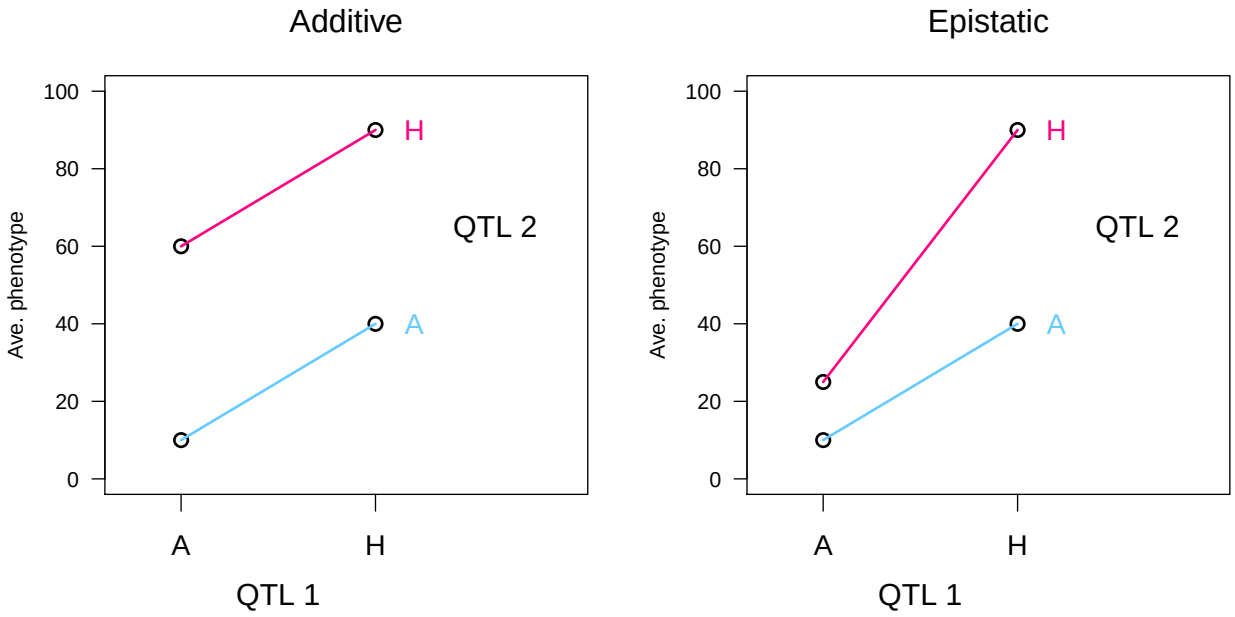
- Reduce residual variation $\longrightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)

Why go after multiple QTL? First, by taking account of large-effect QTL, we can potentially increase our power to detect further, more modest effect loci.

Second, we can better determine whether there are linked QTL, such as the loci on chromosome 1. Does that double-hump mean there are two QTL? The best way to determine that is to explicitly model two QTL and compare the best two-locus model to the best single-locus model.

Third, there is the possibility of interactions between QTL. That the effect of one locus depends on the genotype at a second locus. We can't learn about such interactions without explicitly considering multiple QTL simultaneously.

Epistasis in BC



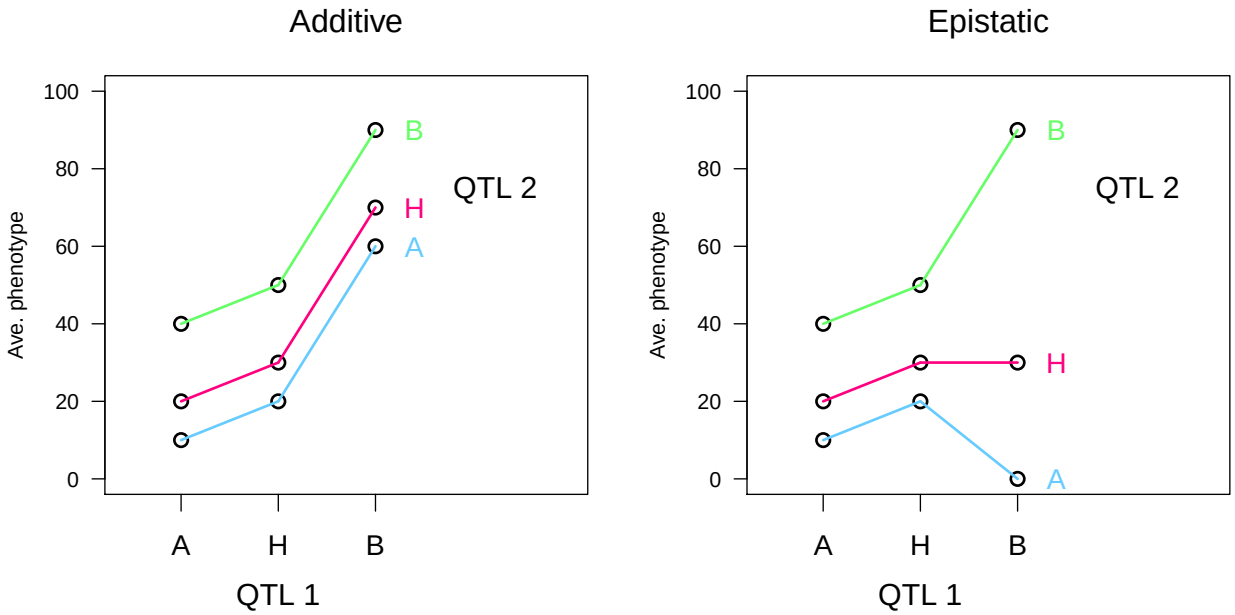
Genetics call interactions between two genetic loci "epistasis."

Here, the dots are at the average phenotype for each of the 4 possible two-locus genotypes in a backcross. The difference between the dots on the blue curve is the effect of QTL 1 when QTL 2 is homozygous AA; the difference between the dots on the pink curve is the effect of QTL 1 when QTL 2 is heterozygous AB.

In the left panel, the effect of QTL 1 is the same, irrespective of the genotype at QTL 2, and so the QTL are said to be "additive." (Some people might say that they have independent effects, but I really don't like that terminology; much better to say additive.) Note that the effect of QTL 2 is also the same, no matter the genotype at QTL 1.

In the right panel, the effect of QTL of QTL 1 is larger when QTL 2 is heterozygous than what QTL 2 is homozygous. Similarly for QTL 2: its effect depends on the genotype at QTL 1. So the QTL are said to be epistatic: they "interact."

Epistasis in F_2



Here we illustrate epistasis in an intercross, where there are three possible genotypes at each QTL.

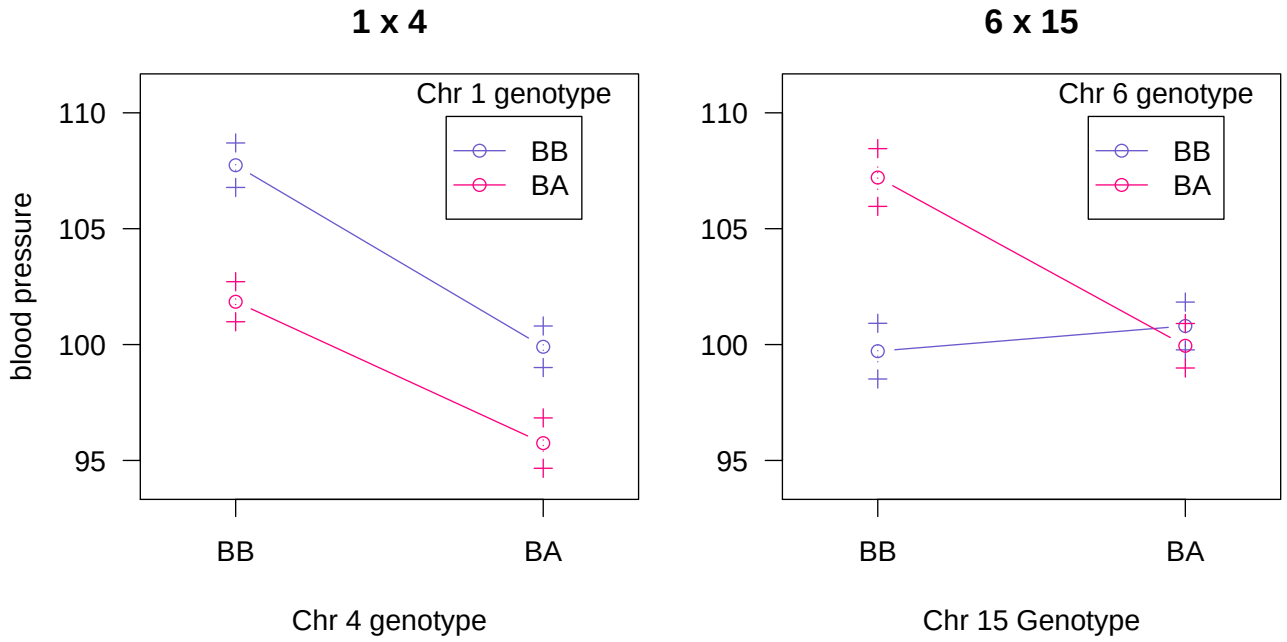
In the left panel, the pattern of effect of QTL 1 is the same, no matter the genotype at QTL 2. So the QTL are additive.

In the right panel, the pattern of effect of QTL 1 is different for different genotypes at QTL 2, and so the QTL are epistatic; they interact.

Note the potential dependence on the scale of the phenotype. If you transformed the phenotype by taking logs or square-roots, the figure on the left would no longer have nice parallel curves.

The figure on the right, though, cannot be repaired by transformation.

Estimated effects



Back to our example data, the QTL on chr 1 and 4 look almost perfectly additive. The QTL on chr 6 and 15, however, show a strong interaction. Basically you see that if you're het at chr 6 and homozygous at chr 15 you have high blood pressure; otherwise you have low blood pressure.

Note that the effect here is as large as the effect of the chr 4 locus, but it's hard to see if you don't take the possibility of the interaction into account. It gets dampened by 1/2 if you consider the chr 6 and 15 loci separately.

Model selection (or variable selection)

- Subset selection
- L_1-penalized regression (the LASSO)
- regression forests
- Bayes
- ...

Ultimately, the question here is one of variable selection: which genetic loci seem to affect the phenotype, and how do they interact?

There are lots of methods for model or variable selection, but I still prefer subset selection.

And much of this work is focusing on minimizing prediction error. The idea is that all predictors have an effect, but there are so many of them and some have such small effects that it may be better to omit them, introducing a bit of bias but reducing the prediction variance.

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

Focusing on variable subset selection, I like to think of the problem as being split into four parts.

First, the selection of a class of models (strictly additive models, or allow pairwise interactions or higher-order terms, or consider a rather different set of models such as regression trees?)

Second, how to fit the models? If there is no missing data, this would just be linear regression. But with missing data in the predictors, there are different ways of dealing with that.

Third, how to compare models? This ends up being the most important thing. Larger models will provide a better fit, but has the fit improved enough for you to incorporate the additional terms?

Finally, there are far more models than can be investigated exhaustively, and so you need some way of searching the space of models. In forward selection, you start with a search over all single-term models and pick the best one. Then you do additional single-term searches of terms to add. This creates a nested sequence of models of increasing size. Backward elimination is the reverse: start with a large model and then omit one term at a time. Stepwise selection goes forward and back; there are also 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.

The key thing is: what is our goal? In many cases one is interested in developing a model that gives good predictions. For QTL mapping, I think we're not interested in prediction but rather in identifying the important terms. I view this in a hypothesis testing like way: thinking about false positives and false negatives.

I view the goal to be to find as many QTL as possible, which controlling the rate of inclusion of false loci.

What is special here?

- Goal: identify the major players
  - not prediction
- 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

There is a vast literature on variable selection in regression, but most of it is not relevant to us, as it tends to focus on minimizing prediction error.

That's not the only difference here, though. We also have this strange continuum of ordinal-valued covariates. And our covariates have a quite simple correlation structure. Between chromosomes, they are completely independent. And along chromosomes, they have a very simple correlation structure.

Compare this to the correlations among covariates in an epidemiologic study, such as of health-related covariates for predicting diabetes.

The simple associations among our covariates in QTL mapping mean that some otherwise quite badly behaved methods actually work well. For example, forward selection has a very bad reputation in statistics. But in QTL analysis, it seems to perform quite well.

Exploratory methods

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

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

- Two-dimensional, two-QTL scan to investigate linked loci or interactions.

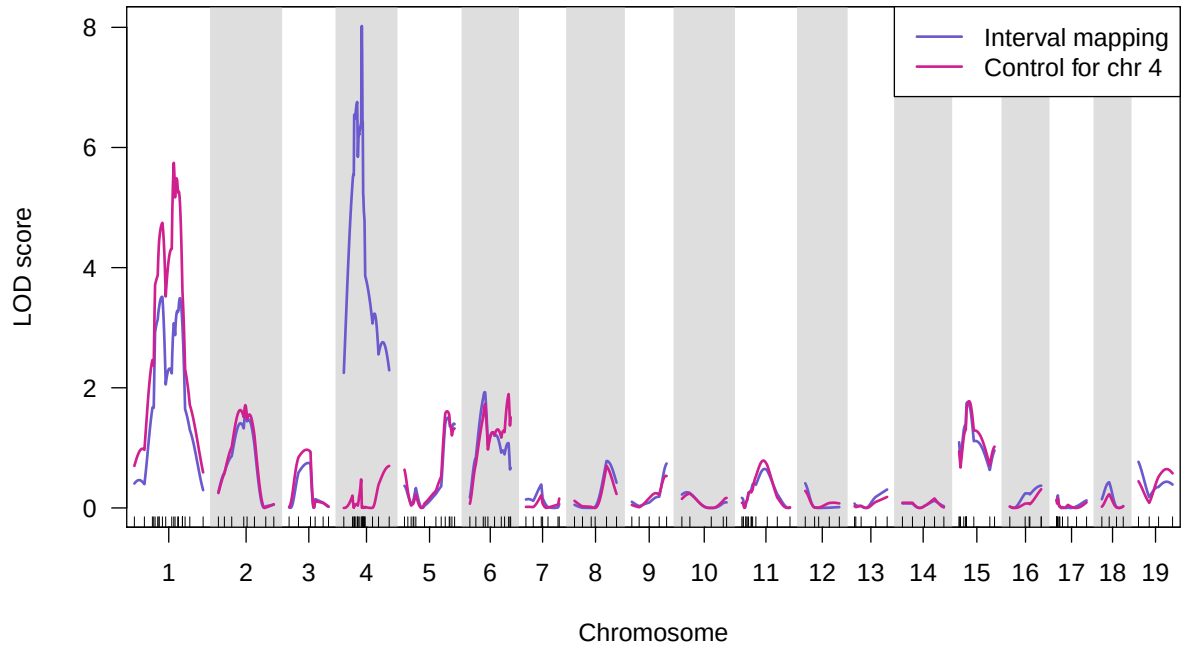
- Piece together the putative QTL from the 1d and 2d scans
  - Omit loci that no longer look interesting (drop-one-at-a-time analysis)
  - Study potential interactions among the identified loci
  - Scan for additional loci (perhaps allowing interactions), conditional on these

There are a variety of exploratory methods that one can consider. For example, you could use a version of forward selection: condition on some large effect QTL and scan for further loci.

You could perform two-dimensional, two-QTL scan; sort of the next step after the single-dimensional genome scan we've often discussed.

Further, you can piece together a multiple-QTL model and then look at scanning for additional loci or adding interactions, or dropping one term at a time, to see which ones are important.

Controlling for chr 4



For our example data, here are the results we get when we condition on the chr 4 locus. The blue curves are the original results, and the pink curves are the results conditional on the chr 4 locus.

The evidence for the chr 1 locus increases, and the size of the two humps are no longer equal; evidence points more strongly towards the second hump.

Not much else has changed in the rest of the genome, though. The biggest change is that there are now two peaks on chr 6.

Drop-one-QTL table

	df	LOD	%var
1@68.3	1	6.30	11.0
4@30.0	1	12.21	20.1
6@61.0	2	7.93	13.6
15@17.5	2	7.14	12.3
6@61.0 : 15@17.5	1	5.68	9.9

This is what I mean by a drop-one-QTL analysis.

We take a model with chromosomes 1, 4, 6 and 15, and with the 6 and 15 locus interacting. We then drop one term at a time and see how much the log likelihood has changed. Note that when we drop a QTL, we also drop any interaction that it's involved in.

Automation

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

But I'm particularly interested in developing fully automated methods: so that you have something that non-specialists can use and get reasonable results, so that you can understand its performance (for example, through simulation studies), and because you might want to apply it to cases where you have tens of thousands of phenotypes.

Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$y = \mu + \sum \beta_j q_j + \epsilon \qquad \text{which } \beta_j \neq 0?$$

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

0 vs 1 QTL:

$$\text{pLOD}(\emptyset) = 0$$
$$\text{pLOD}(\{\lambda\}) = \text{LOD}(\lambda) - T$$

Consider the case of strictly additive QTL. And suppose I have dense markers and complete genotype data, so I can ignore the whole issue of missing data and just focus on variable selection in linear regression.

The question is: which QTL locations have $\beta \neq 0$?

I've focused on a penalized LOD score approach. Consider a model (a set of QTL locations) denoted γ, and let $\text{LOD}(\gamma)$ denote the $\log_{10}$ likelihood ratio of that model vs the null model. The penalized LOD score is the LOD score minus some penalty on the size of the model. As I add more terms to the model, the LOD score increases but so does the penalty.

The question is: what penalty to use? Well, what I want is to ensure that the chance of including a false positive is controlled at some rate. Imagine the null hypothesis is true, that there are no QTL, and I do a search among single-QTL models (with a QTL at some position λ), if I choose the penalty T to be the significance threshold from the genome-scan permutation test, that will do just what I want. I then cross my fingers and hope that it works for larger models and more extensive searches.

Experience

- Controls rate of inclusion of extraneous terms
- Forward selection over-selects
- Forward selection followed by backward elimination works as well as MCMC
- Need to define performance criteria
- Need large-scale simulations

Broman & Speed, JRSS B 64:641-656, 2002
10.1111/1467-9868.00354

Our experience in applying this approach has been remarkably good. The penalized LOD score criterion does a good job of controlling the false positive rate, and while forward selection tends to over-select (including some extra terms), if it's followed by a backward elimination step, it works as well as the best MCMC algorithm we can contrive.

It's also been clear that defining performance criteria and performing large-scale simulations are important for establishing the relative value of different methods.

Epistasis

$$y = \mu + \sum \beta_j q_j + \sum \gamma_{jk} q_j q_k + \epsilon$$

$$\text{pLOD}(\gamma) = \text{LOD}(\gamma) - T_m |\gamma|_m - T_i |\gamma|_i$$

T_m = as chosen previously

T_i = ?

If we want to expand the approach to include interactions, we need to also establish a penalty on interaction terms.

I focus on models where if you include an interaction term, you always include both main effects.

Idea 1

Imagine there are two additive QTL and consider a 2d, 2-QTL scan.

T_i = 95th percentile of the distribution of

$$\max \mathrm{LOD}_f(s,t) - \max \mathrm{LOD}_a(s,t)$$

For the mouse genome:

T_m = 2.69 (BC) or 3.52 (F_2)

T_i^H = 2.62 (BC) or 4.28 (F_2)

My first idea for deriving a penalty on interactions was to use the same approach as for main effects. So we consider a two-dimensional, two QTL scan, and we derive a threshold on interaction terms, as the 95th percentile of the distribution of the $\log_{10}$ likelihood ratio comparing the best two-locus interactive model to the best two-locus additive model.

The penalty is pretty harsh, but it has the property that it will control the rate of inclusion of extraneous interactions.

Idea 2

Imagine there is one QTL and consider a 2d, 2-QTL scan.

$T_m + T_i$ = 95th percentile of the distribution of

$$\max \mathrm{LOD}_f(s,t) - \max \mathrm{LOD}_1(s)$$

For the mouse genome:

T_m = 2.69 (BC) or 3.52 (F_2)

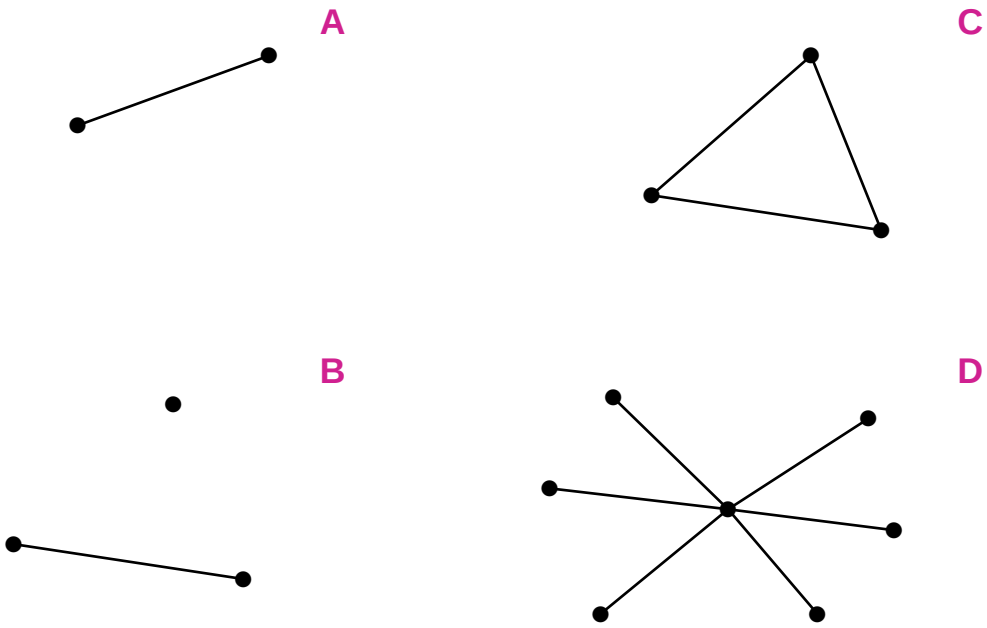
T_i^H = 2.62 (BC) or 4.28 (F_2)

T_i^L = 1.19 (BC) or 2.69 (F_2)

But if that idea is good, I have an even better one: compare the best two-locus interactive model to the best single-locus model, and take that to be the sum of the new interactive penalty and the previous main-effect penalty.

This leads to a much lighter penalty on interactions. It will include false interaction terms at a high rate, but it will allow us greater power to detect QTL. If we care mostly about not including extraneous QTL, and we don't mind falsely identifying a few interactions, this approach will work well.

Models as graphs



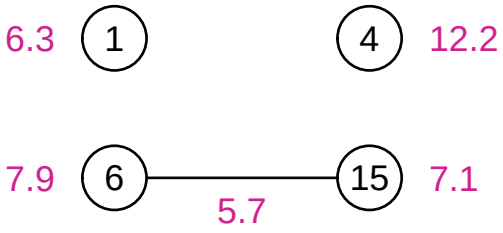
In considering multiple-QTL models with the requirement that when you include an interaction you always include both main effects, it's useful to display QTL models as graphs, where the nodes are the QTL and edges connecting them indicate that they interact.

We found that the approach of the light interaction penalty tends to lead to models like that in panel D, with a single false QTL interacting with a bunch of other QTL.

Our solution to this problem was a bit ad hoc: considering a model as a graph, we allow up to one light interaction penalty for each connected component, and give all the rest of the interactions a heavy penalty.

So imagine the four panels here were one big QTL model with 15 QTL and 11 interactions. We'd give 15 main effect penalties, 4 light interaction penalties, and 7 heavy interaction penalties. (I ended up mostly skipping over this detail in class, because it's a bit of a pain and I was running out of time.)

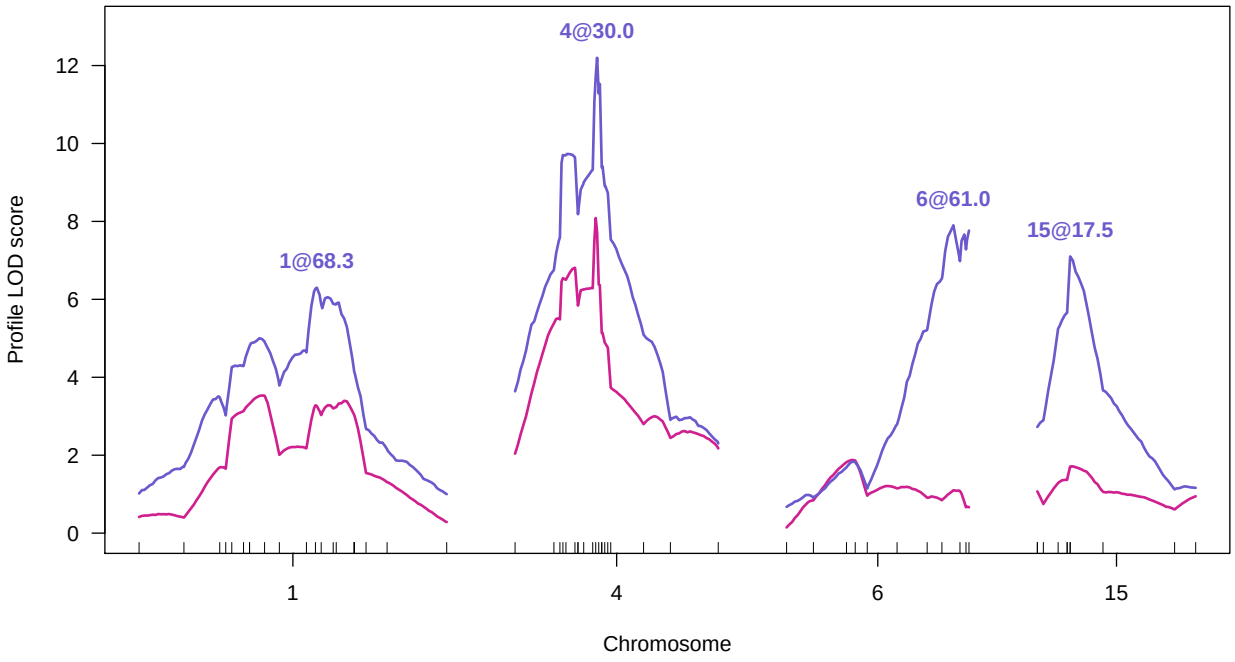
Results



$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Let's return to the example data. My exploratory analyses led to this four-QTL model with an interaction between the 6 and 15 loci. If we look at the drop in LOD score when we drop one QTL or drop the interaction, we find that each exceeds the penalties we have defined, so each piece has strong support.

Profile LOD curves



I find it useful to further depict the evidence for such a multiple-QTL model using profile LOD curves.

In pink are the results of the genome scan on this chromosome: at each point, we compare the single-QTL model with a QTL at that position to the null model with no QTL.

In blue are the profile LOD curves for the four-QTL model with an interaction between the chr 6 and 15 loci. For each chromosome, we compare the four-QTL model with the loci on other chromosomes fixed at their best positions and with the present QTL allowed to vary in position, and we compare the four-QTL locus to the three-QTL locus where we drop the given QTL.

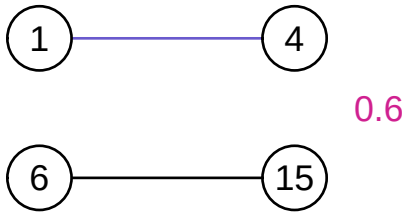
The height of each curve indicates the evidence for each QTL, and the curves also show the precision of localization of the QTL.

Drop-one-QTL table

	df	LOD	%var
1@68.3	1	6.30	11.0
4@30.0	1	12.21	20.1
6@61.0	2	7.93	13.6
15@17.5	2	7.14	12.3
6@61.0 : 15@17.5	1	5.68	9.9

The drop-one-QTL time is as I showed before; the values match the maximum height of each of the profile LOD curves on the previous slide.

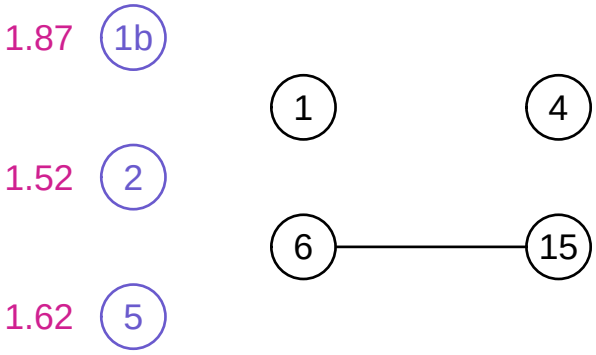
Add an interaction?



$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

If we take our four-QTL model and look at adding an interaction, none of them are interesting.

Add another QTL?



$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

We can also go through and look to add another QTL, or another QTL that interacts with one of the current ones, or add a pair of additional QTL. Adding a second QTL on chr 1 comes closest to being chosen, but still doesn't make it. Also interesting is the potential of a pair of linked loci on chromosome 3, but they also don't quite make it.

So my system of penalties results in the base model that I came to from exploratory analyses. But I view that as a benefit: it gives an automated procedure that anyone can use and that arrives at the model that my experience tells me is most trustworthy.

Summary

- QTL mapping is a model selection problem
- The problem is finding the major players, not minimizing prediction error
- The criterion for comparing models is most important
- We're focusing on a penalized likelihood method, with penalties derived from permutation tests with 1d and 2d scans
- Manichaikul et al., Genetics 181:1077–1086, 2009
  doi:10.1534/genetics.108.094565

QTL mapping is a variable selection problem, but it's somewhat unusual in that we're not selecting variables to give better predictions but rather we're interested in the major players. The associations among our covariates are also unusual.

The criterion for comparing models is the most important part of the problem. I discussed this penalized likelihood criterion that I like; it's described further in this 2009 Genetics paper.