

# Dissecting and fine-mapping *trans*-eQTL hotspots

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

[kbroman.org](http://kbroman.org)

[github.com/kbroman](https://github.com/kbroman)

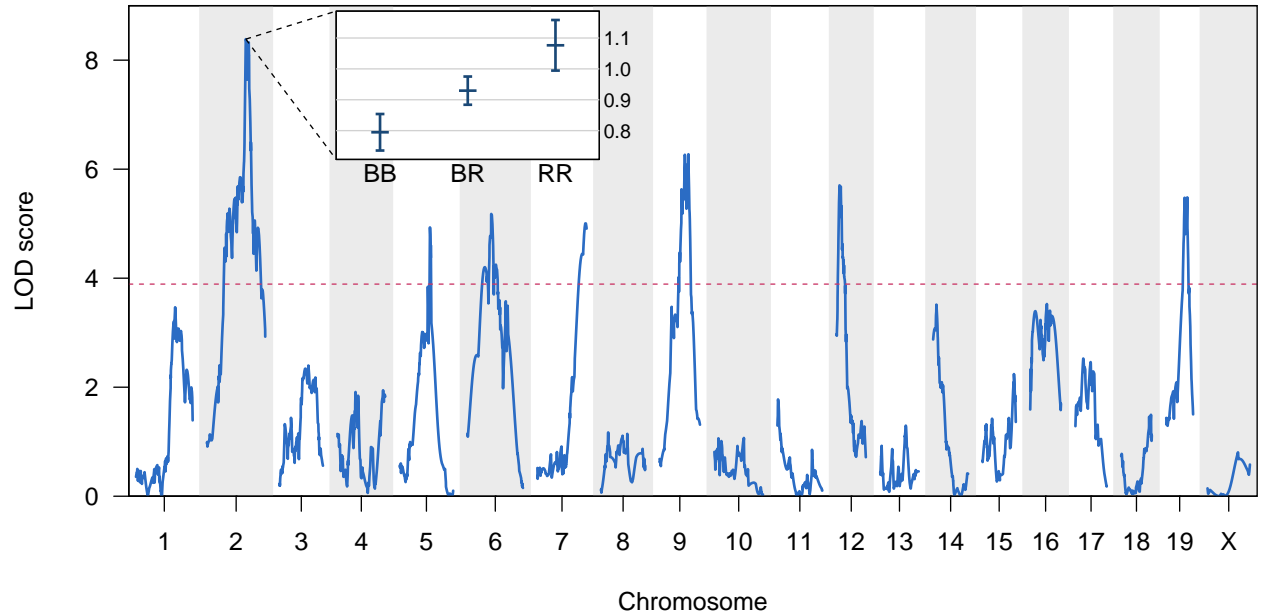
@kwbroman

Course web: [kbroman.org/AdvData](http://kbroman.org/AdvData)

In this lecture, we'll look at a case study on fine-mapping an eQTL, and determining where there is one or two eQTL in a region.

This is a case-study in the analysis of expression QTL (eQTL), where you're looking for loci that affect the level of mRNA of genes (gene expression).

# QTL mapping



2

Our goal is to identify quantitative trait loci (QTL): regions of the genome for which genotype is associated with the phenotype.

The basic analysis is to consider each locus, one at a time, split the mice into the three genotype groups, and perform analysis of variance.

We then plot a test statistic that indicates the strength of the genotype-phenotype association. For historical reasons, we calculate a LOD score as the test statistic: the  $\log_{10}$  likelihood ratio comparing the hypothesis that there's a QTL at that position to the null hypothesis of no QTL anywhere.

Large LOD scores indicate evidence for QTL and correspond to there being a difference in the phenotype average for the three genotype groups.

But a big disappointment in QTL analysis has been that the regions to which we can map them can be quite large and so include lots of potential candidate genes.

DNA → RNA → protein → phenotype

3

One approach to try to speed the process is to look at intermediate, molecular phenotypes.

The “central dogman” of molecular biology is that DNA is transcribed into RNA which is then translated into protein. And ultimately this is what leads to the observable phenotypes.

So there’s been a big effort to measure the RNA level of genes, called “gene expression.” Mapping loci that affect gene expression might help us to identify the loci that affect clinical traits.

One can also look at protein levels, and that might be more valuable, but it turns out to be much easier to measure RNA.

## Gene expression microarrays



4

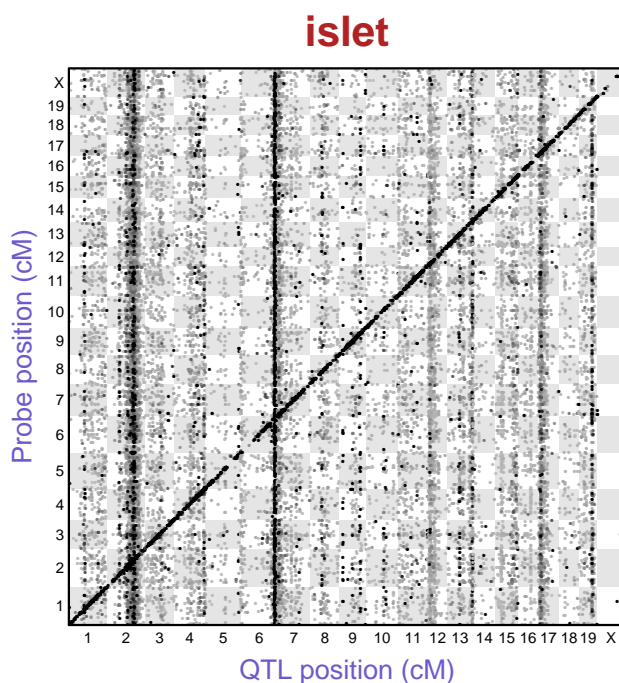
This is a picture of a pile of gene expression microarrays, which had classically been used to measure mRNA.

In collaboration with Alan Attie, we've been studying a B6×BTBR mouse intercross, with all mice knocked out for leptin (and so obese), in order to understand obesity-induced diabetes.

There are 500 intercross mice, phenotyped at a large number of clinical traits, and also with gene expression microarray data on 6 tissues. These were custom two-color Agilent arrays. In each tissue in each mouse, we measured the mRNA level in each of the 30,000 genes.

We're interested in identifying genes that control the expression of other genes.

## eQTL results

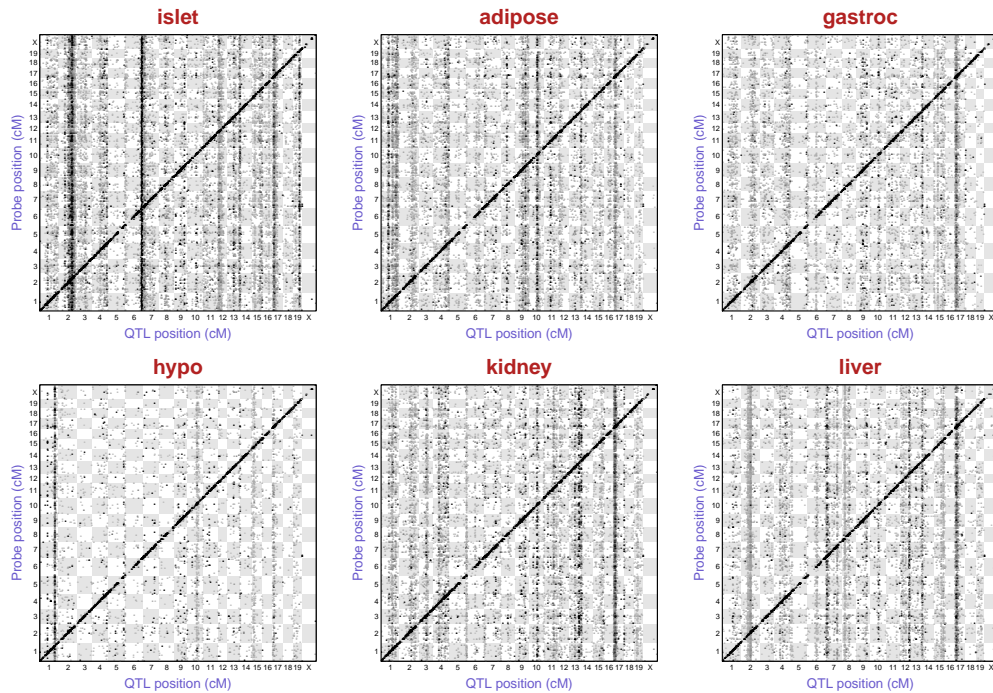


5

This figure shows the basic result of single-QTL genome scan for each expression trait, one at a time, in pancreatic islets. Each dot is an inferred QTL. The y-axis is the location of the corresponding gene, and the x-axis is the location of the QTL.

We see a prominent diagonal, of local-eQTL, where the mRNA level for a gene is affected by a QTL right at the gene. There are also several prominent vertical bands: “trans-eQTL hotspots” where genotype at a give region is associated with the mRNA expression of numerous genes across the genome.

## eQTL results



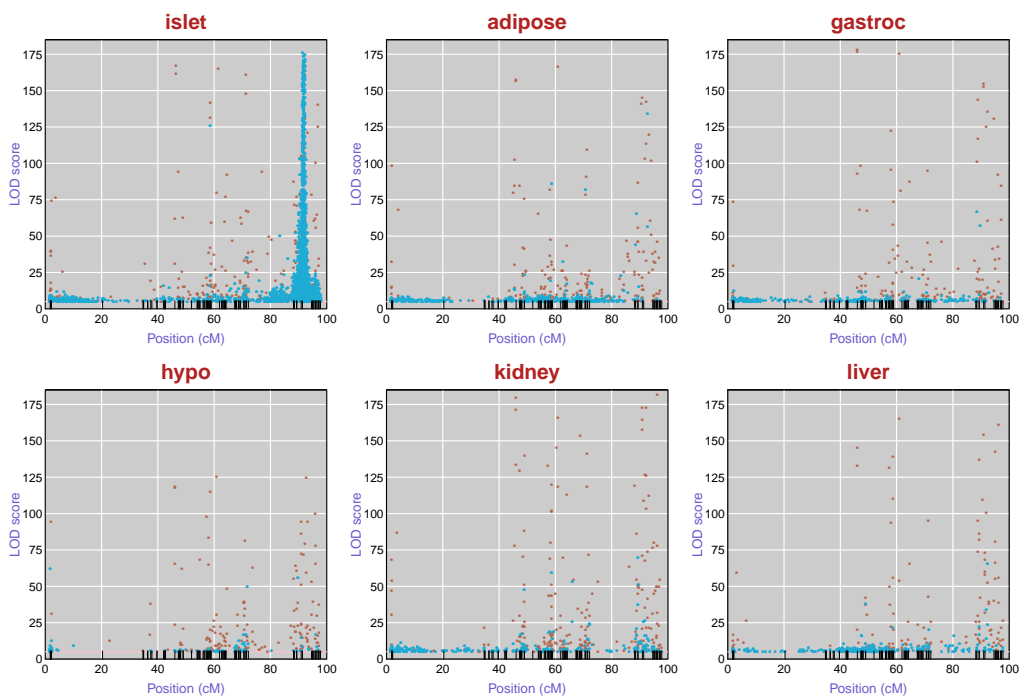
6

Here are the results for all six tissues.

There are numerous “trans-eQTL hotspots” (where genotype at a give region is associated with the mRNA expression of numerous genes across the genome). Some of these trans-eQTL hotspots are specific to a given tissue (e.g. islet chr 6) and some are seen in many tissues (e.g. chr 17).

We seek to fine-map these trans-eQTL hotspots, and to determine whether they involve one or multiple eQTL.

## Chr 6



7

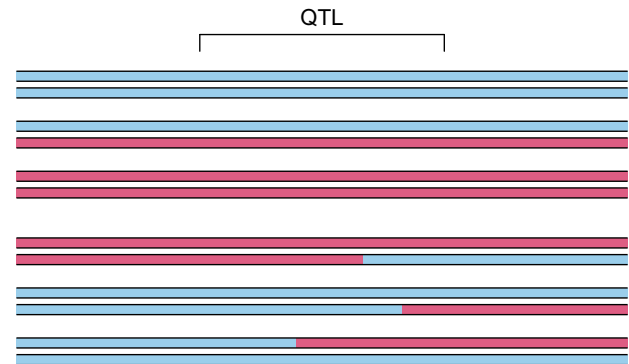
I'm particularly interested in the locus on chromosome 6, which affects like 8% of genes in pancreatic islets, but is entirely specific to islets.

Each dot corresponds to the peak LOD score of a single expression trait: the LOD score vs the position at which it occurred. The blue dots are for expression traits that exist on other chromosomes; the brown dots are for expression traits that reside on chromosome 6.

The local-eQTL are similar for the six tissues, but the hotspot on distal chr 6 in islets is seen only in islets.

Assuming this is a single gene, how can we define a precise interval for the gene? We'd like to refine the localization and actually determine the responsible gene.

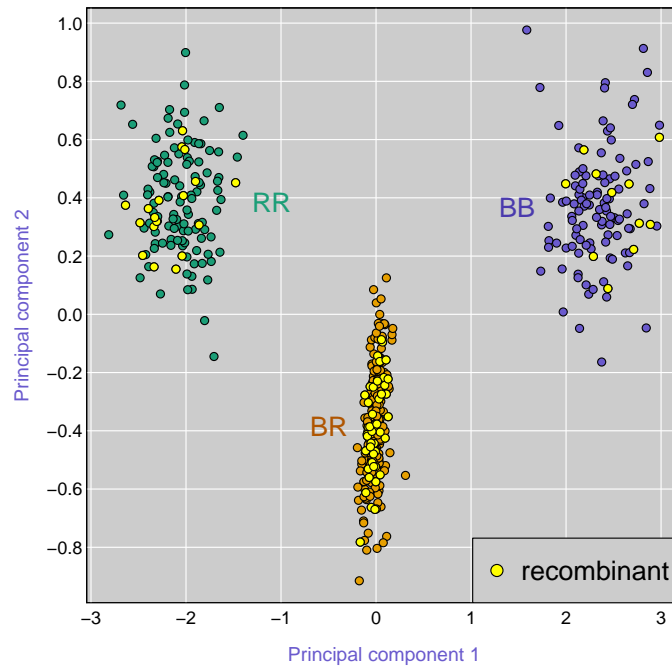
## Consider the non-recombinants...



Our key strategy is to focus on the non-recombinant mice (that is, those mice that have no recombination event in the region of this eQTL). For these mice, we know their QTL genotype. We can use this to establish the distribution of the multivariate expression phenotype in each genotype group.



## Islet c6 PCs



9

We focused on the 177 microarray probes that are not on chromosome 6 and that map to this region with LOD score  $> 100$ .

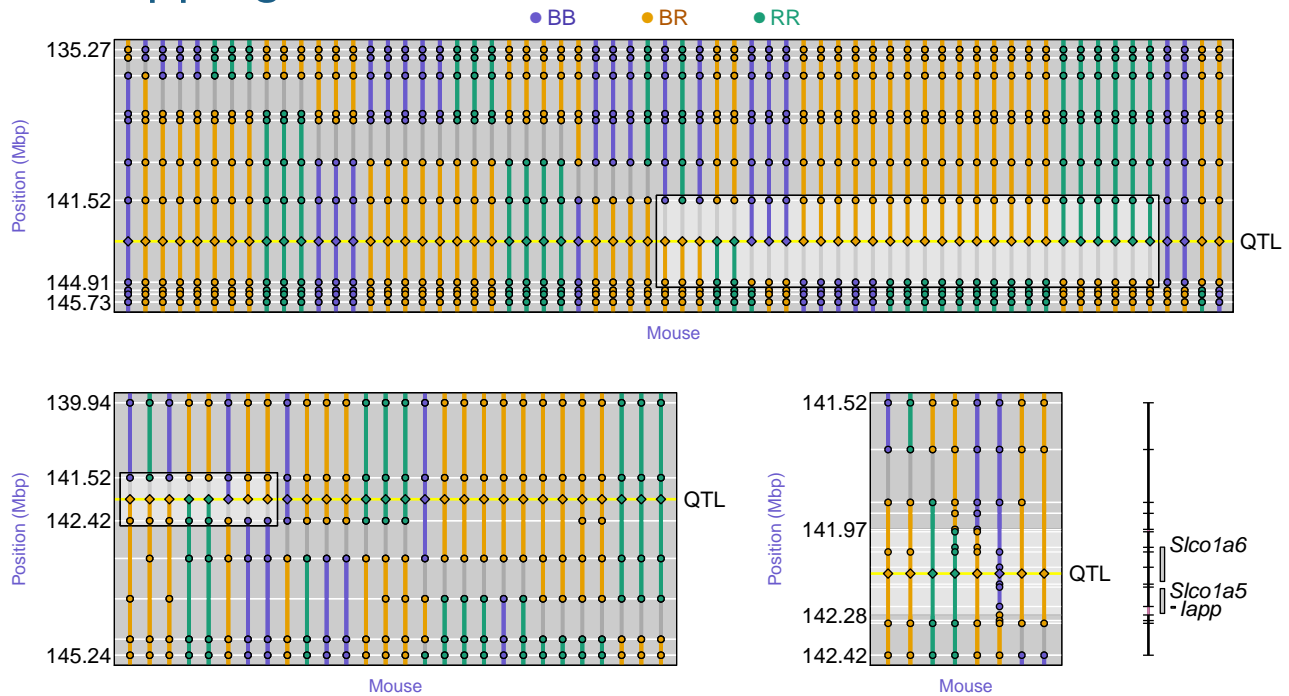
We exclude expression traits that reside on chromosome 6, thinking that they might be affected by separate, local-eQTL rather than the present locus of interest.

We first focus on mice that had no recombination event in the 10 cM interval surrounding the eQTL, apply principal components analysis, and make a scatter plot of the first two principal components. Each dot is a mouse. There are three clear clusters which correspond to the three possible QTL genotypes.

There were 74 mice that showed a recombination event in the interval; they all fall clearly into one of the three clusters. We can infer their eQTL genotype based on the cluster into which they fall.

In this manner, the multivariate gene expression phenotype is converted to a co-dominant Mendelian phenotype.

## Fine-mapping the c6 locus



10

Using these inferred eQTL genotypes, we can locate the QTL to a single 3.5 Mbp interval between two markers (top panel). There are 29 mice (highlighted) that showed a recombination event in that interval.

Genotyping an additional four markers in the interval in 28/29 of these mice (for one mouse, DNA was not available) reduced the QTL interval to 900 kbp (lower-left panel). There were eight mice (highlighted) that showed a recombination event in this interval.

Additional genotyping of these eight mice reduced the QTL interval to 298 kbp. This interval contains just three genes. Additional genotyping cannot exclude any of these genes.

Our best candidate is *Slco1a6*, which is a transporter of bile acids, some of which can have large effect on gene expression. B6 and BTBR show a number of coding variants in this gene, one of which is plausibly functional. Bruno Hagenbuch at the University of Kansas Medical Center has shown that the two variants differ in activity, but we've not completely proven that *Slco1a6* is responsible for the huge expression difference in pancreatic islets.

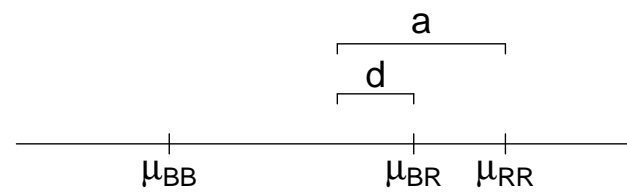
## Is it one QTL?

11

We now turn to the “dissection” part of the talk.

That chromosome 6 eQTL for islets looked like a single gene, but how can we tell? We’ve developed a number of graphical diagnostics, plus a formal statistical test of one vs two QTL for a trans-eQTL hotspot.

## Consider the QTL effects...

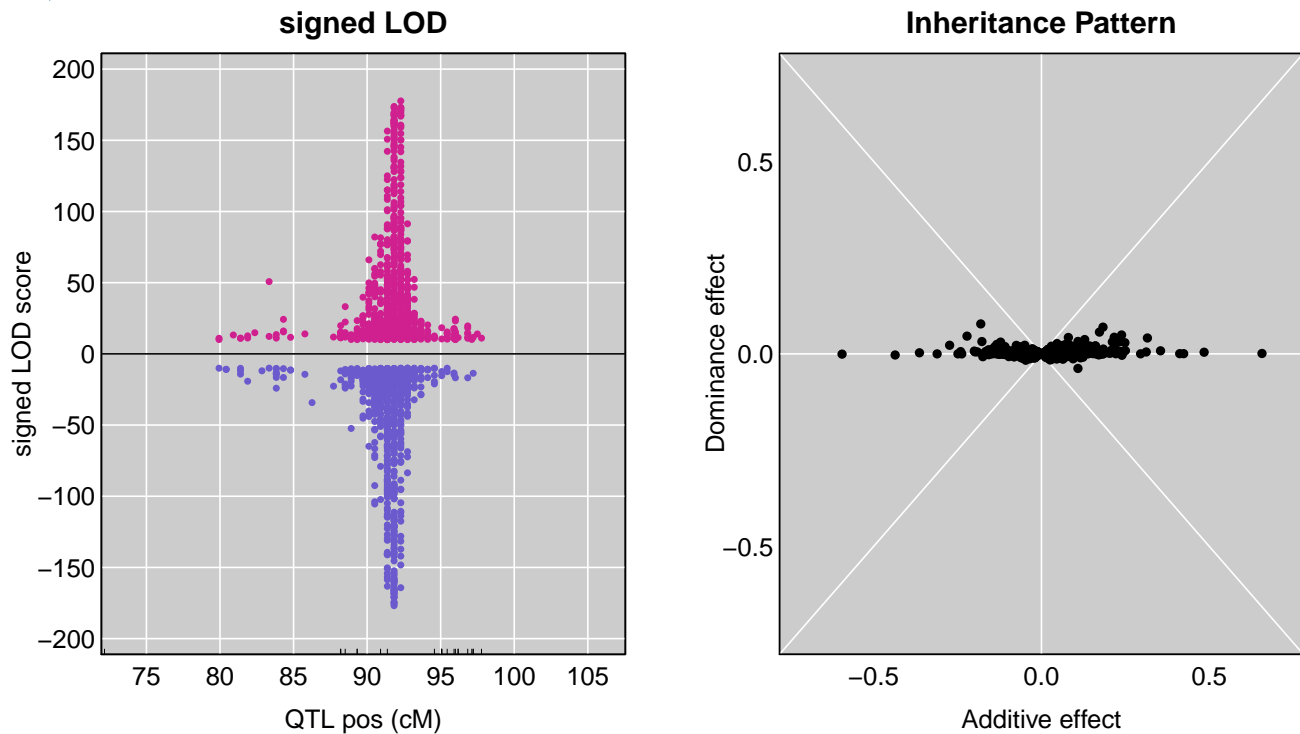


12

First, let's look at the effects of the QTL on the expression traits that map to the region.

The additive effect ( $a$ ) is half the difference between the two homozygotes' phenotype averages. The dominance effect ( $d$ ) is the difference between the heterozygote average and the midpoint between the two homozygotes.

## eQTL effects: Islet c6



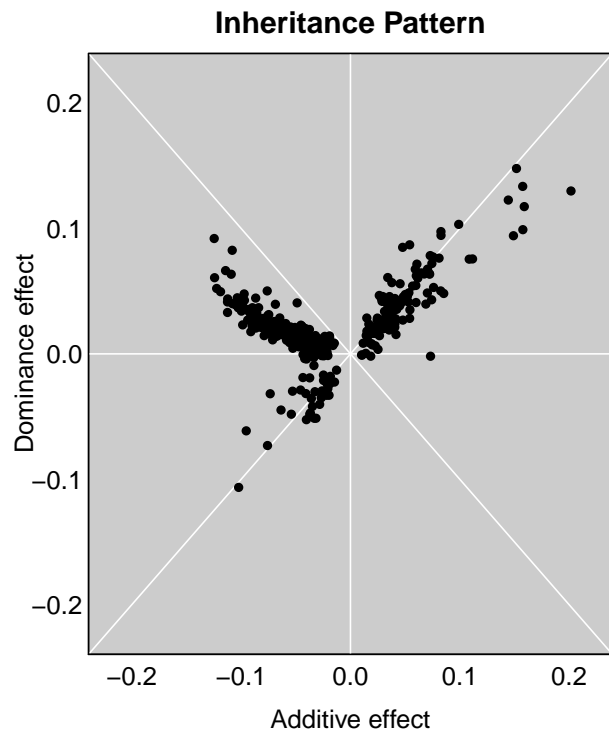
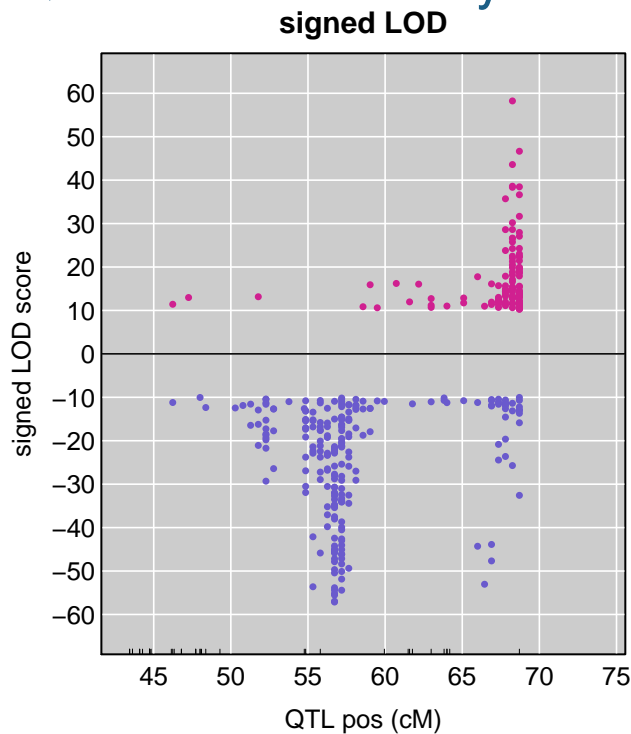
13

Here's the islet c6 locus we've been studying. On the left is a plot of LOD score vs QTL position; we're using a signed LOD score, with the sign taken from the estimated additive effect of the QTL on the corresponding expression trait. (Each dot is a single expression trait.)

On the right is a plot of the estimated dominance effect vs the estimated additive effect.

The QTL looks to be additive for all expression traits, and there are approximately equal numbers of traits where the B6 allele is associated with increased and decreased expression.

## eQTL effects: Kidney c13



14

Here's a trans-eQTL hotspot on chr 13, for kidney. We're considering a pretty wide interval here, but the pattern is interesting and instructive.

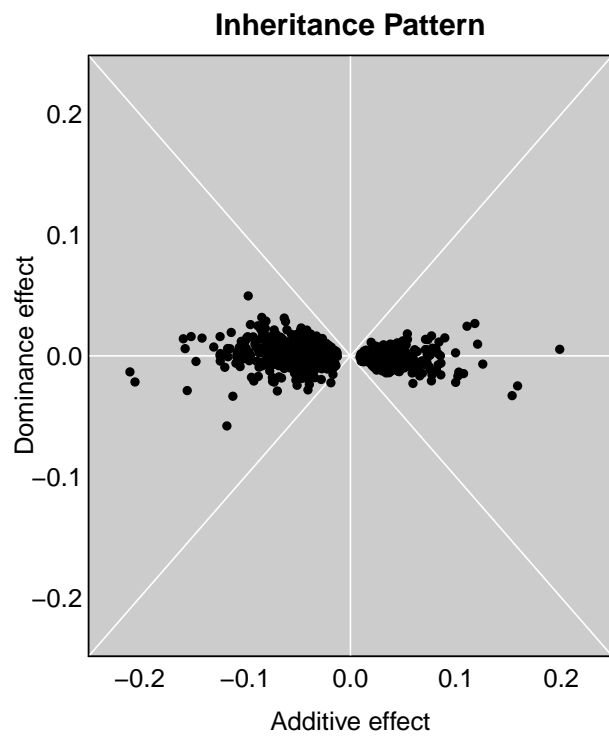
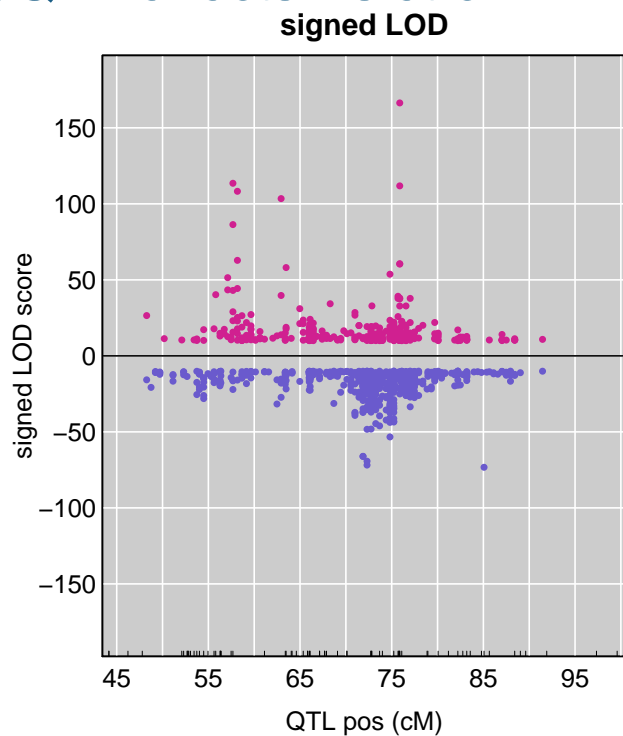
There looks to be a QTL at about 57 cM, and then another at about 68 cM. For the proximal locus, the B6 allele is associated with decreased expression for all traits. The distal locus has the opposite pattern, though there are some in each direction.

In the right panel, we see that there are a group of traits where the BTBR allele is dominant, and from the combination of the two figures, we can infer that these correspond to the proximal locus. At the distal locus, the B6 allele is dominant.

That the two loci show such distinctive inheritance patterns is strong evidence that they are in fact distinct.

Of course, they're rather far apart, and so we'd probably come to this conclusion anyway. But imagine the case that these were sitting just 5 cM apart. Consideration of the QTL effects could be useful.

## eQTL effects: Islet c2

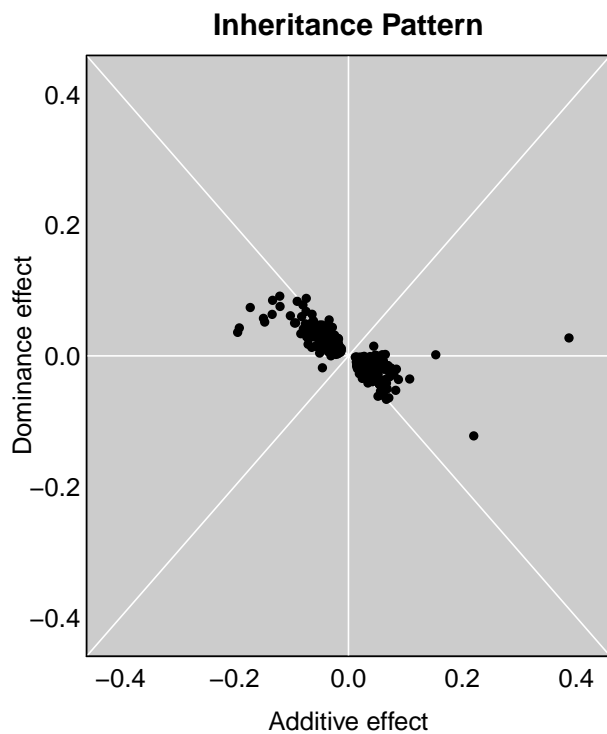
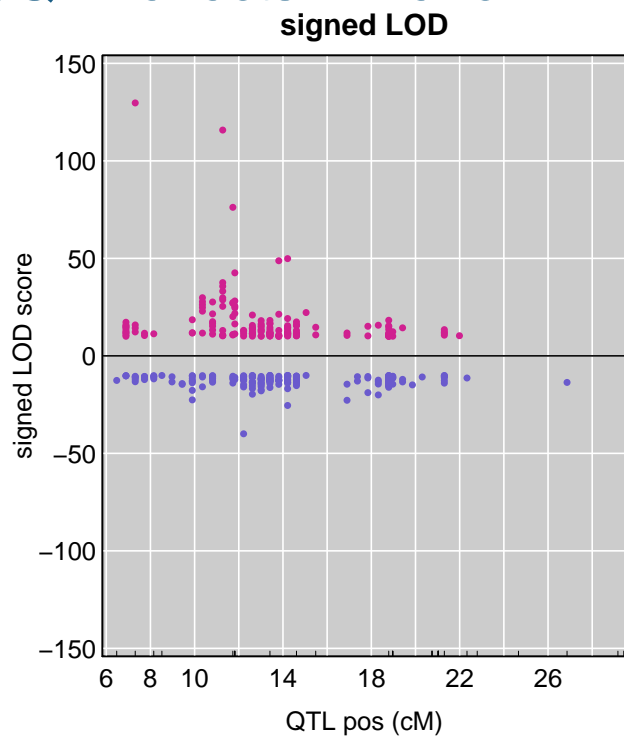


15

Here's another example: islet, chr 2. This is a bit of a muddle.

Note the additive effects, in both directions.

## eQTL effects: Liver c17

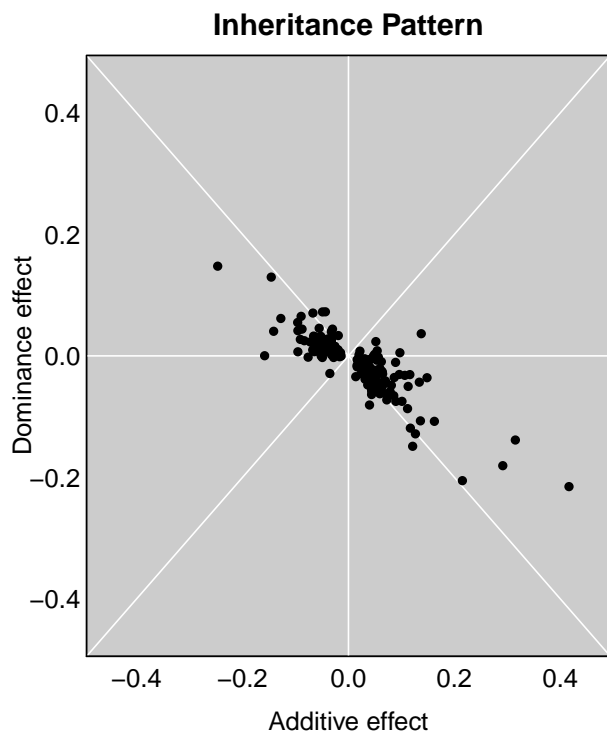
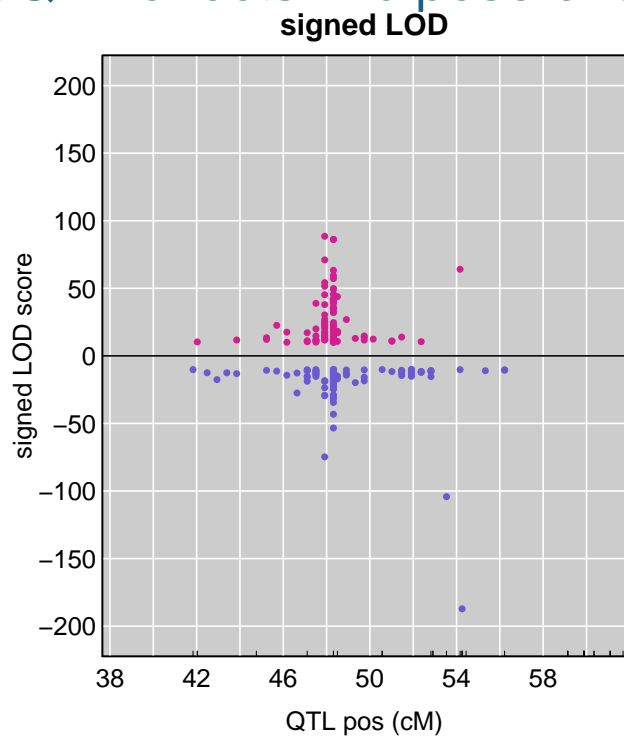


16

Here's liver, chr 17. Again, not clear evidence for two QTL. Here, BTBR looks to be dominant, but again with effects in both directions.



## eQTL effects: Adipose c10



17

One last example: adipose, chr 10. Again, no clear evidence for two QTL, but note the two expression traits with large effects, mapping to 54 cM.

Again, BTBR looks to be dominant, and again with effects in both directions.

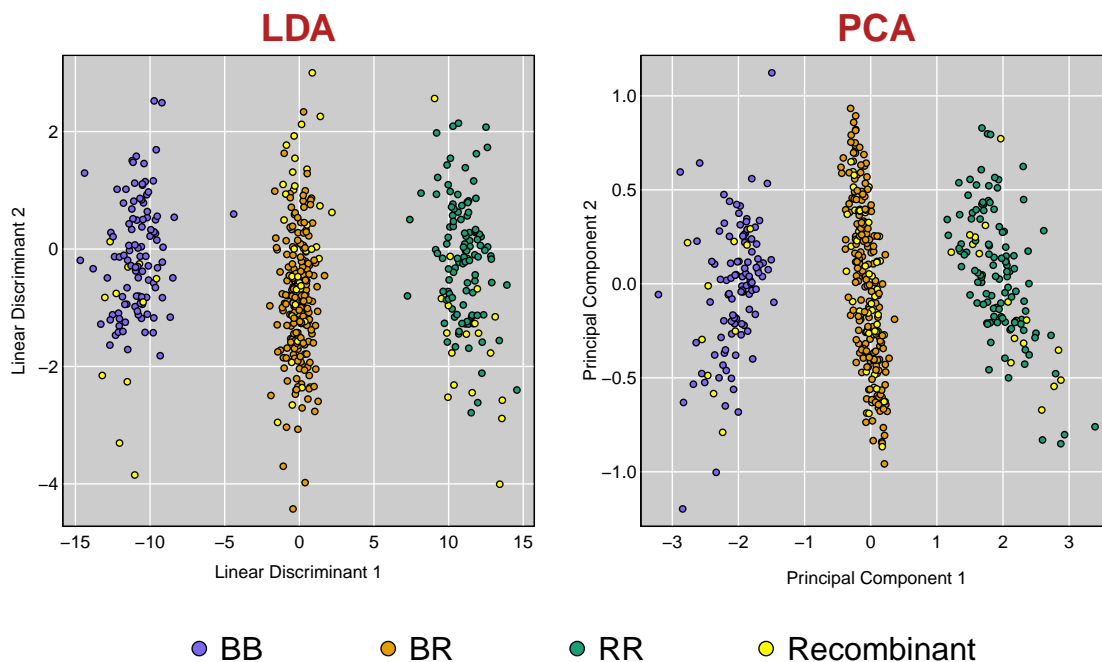
## Compare the recombinants and non-recombinants.

18

A second strategy is to compare the recombinants and non-recombinants, much as we'd done in fine-mapping the islet chr 6 locus.

We can use the non-recombinants to establish the relationship between eQTL genotype and the multivariate expression phenotype. If the effects are strong, we should be able to discern three clear clusters. If there's a single QTL, the recombinants should fit clearly within these three clusters.

## LDA & PCA: Islet c6



19

Here's the islet chr 6 locus again.

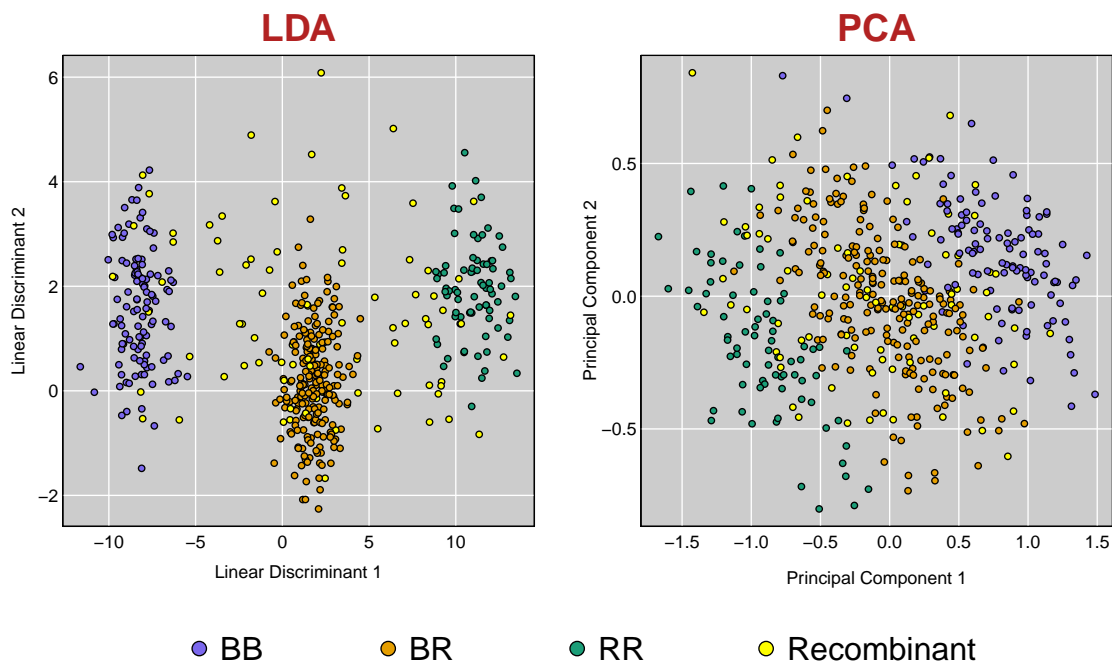
On the left, we use linear discriminant analysis to form a classifier of genotype based on expression phenotype, using just the non-recombinant mice. On the right, we use principal component analysis, again just with the non-recombinant mice.

We look to see whether the recombinant mice fall into the clusters defined by the non-recombinant mice.

In this case, the recombinants look just like the non-recombinants, which is consistent with there being a single eQTL.

Note that the PCA figure (on the right) is a bit different than the one I'd shown earlier. We're focusing on the top 50 expression traits here; before we looked at all traits mapping with  $\text{LOD} > 100$ .

## LDA & PCA: Islet c2

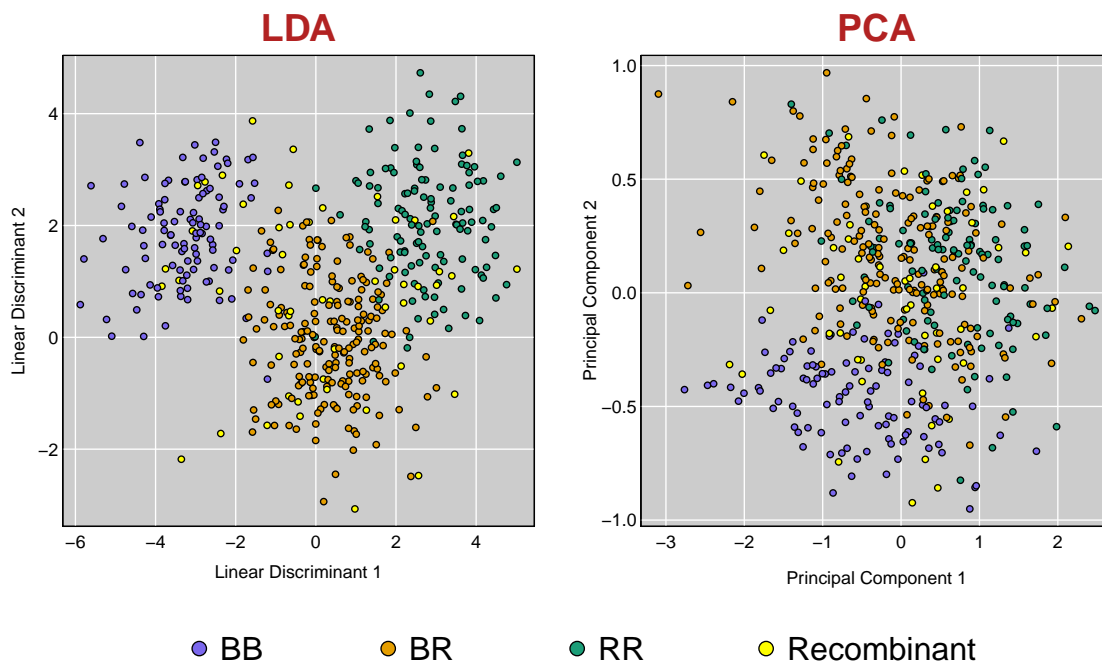


20

Here's islet chr 2. The LDA plot shows distinct clusters for the non-recombinant mice. PCA is a bit of a mess. As we'll see, the PCA figure is always a bit of a mess. We prefer LDA for this.

Most interesting: the recombinant mice (in yellow) **don't** all fall within the clusters defined by the non-recombinant mice. This is good evidence for there being multiple QTL in the region.

## LDA & PCA: Kidney c13

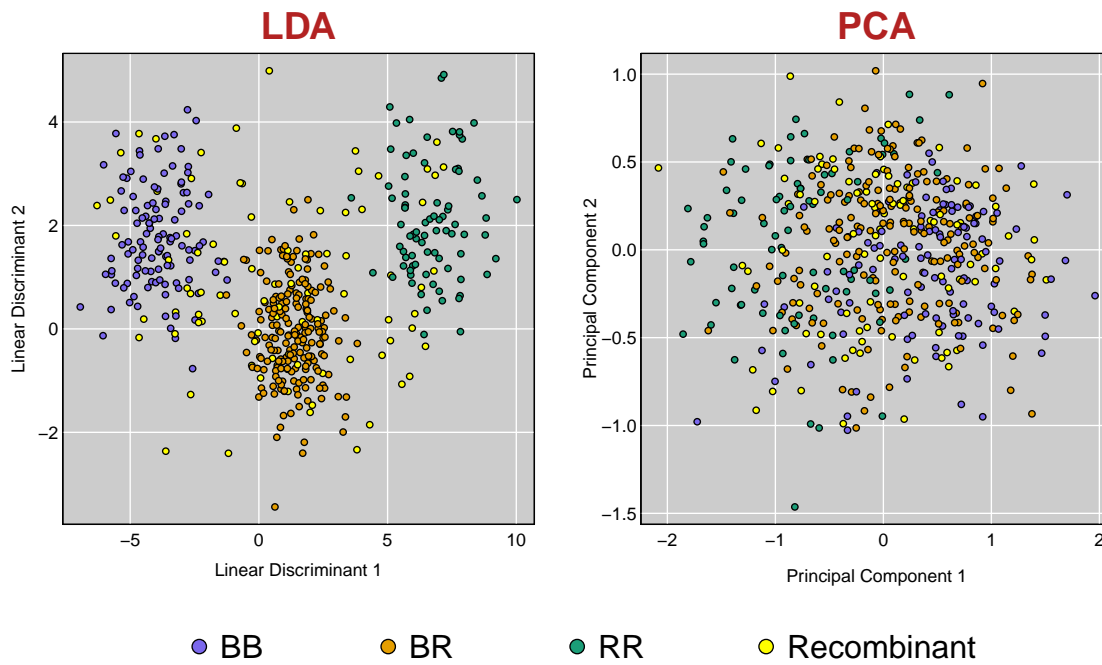


21

Here's kidney chr 13.

The three clusters are not so clearly defined. The recombinants don't look too different from the non-recombinants, but it's all a bit of a mess. These ideas aren't always helpful.

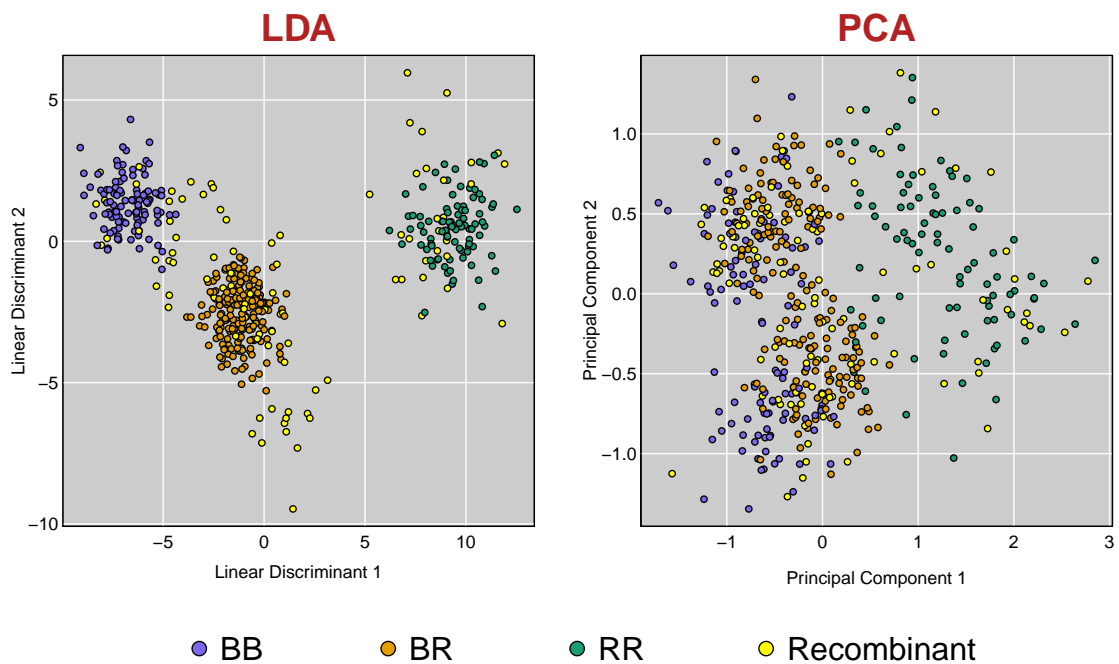
## LDA & PCA: Liver c17



22

Here's liver chr 17. It's not as clear as the previous example, but it does seem that the recombinant mice are different from the non-recombinant mice.

## LDA & PCA: Adipose c10



23

Here's adipose chr 10. The non-recombinant clusters are quite tight, and some of the recombinant mice fall outside of those clusters, indicating multiple QTL.

## Formal test for 1 vs 2 QTL

- ▶ Consider a set of traits mapping to common eQTL
- ▶ Multivariate QTL analysis with 1 or 2 QTL
- ▶ With 2-QTL model, each trait affected by one or the other QTL
  - Order traits by estimated QTL location when considered separately
  - Consider cut points of the list, assign first group to one QTL and second group to other.
- ▶ P-value: parametric bootstrap or stratified permutation

24

As a formal test of whether a trans-eQTL hotspot is due to one or multiple QTL, we use multivariate QTL analysis with one- and two-QTL models. In the two-QTL model, each expression trait is affected by one or the other QTL.

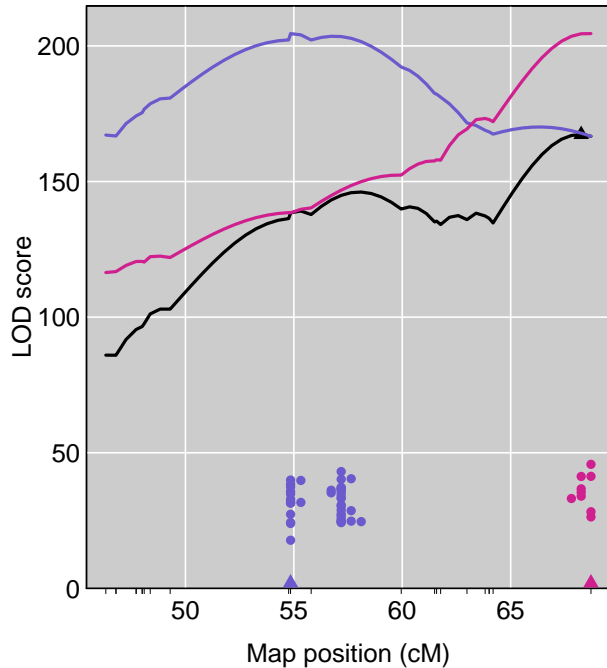
A key technical difficulty is that the two-QTL model requires consideration of all possible partitions of expression traits between the two QTL. As an approximation, we sort the traits by their estimated QTL locations, when considered individually, and then consider only cut-points on this list:  $p - 1$  partitions rather than  $2^{p-1}$ . For each cut-point, we perform a two-dimensional scan for the pair of QTL locations.

We use a parametric bootstrap (simulate data from the estimated single-QTL model) or a stratified permutation test (permute mice within strata defined by genotypes at the estimated QTL location under the single-QTL model).

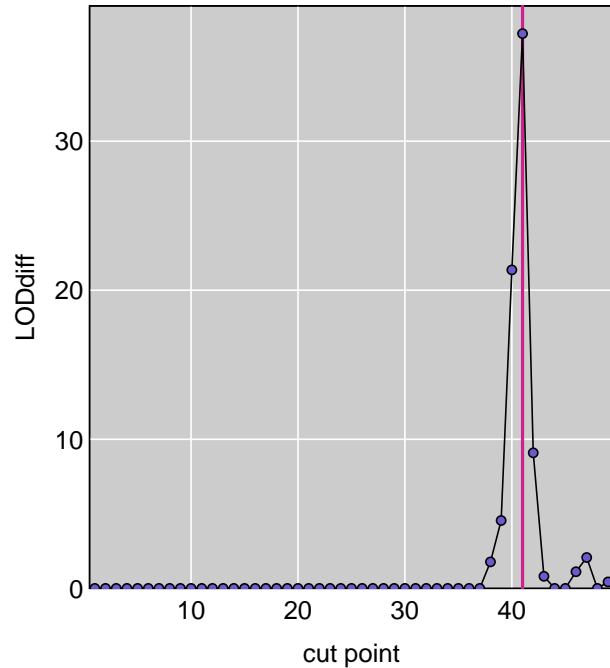


# 1 vs 2 QTL: Kidney c13

## LOD profile



## LOD diff by cutpoint



25

I'll start with a locus with clear evidence for two QTL.

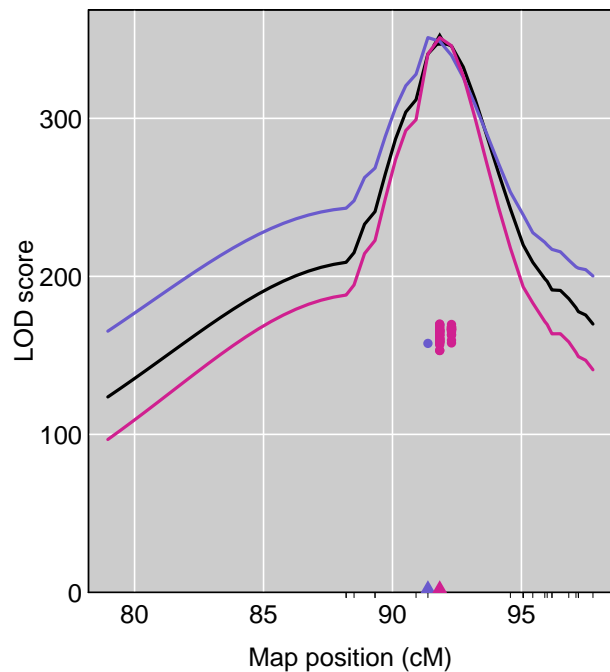
On the left: The blue and red curves are LOD profiles for the location of the two QTL (for the split that gave the best fit): slices through the 2d LOD surface, keeping the other QTL location fixed. The black curve is the LOD profile for the single-QTL model. The dots at the bottom are the estimated QTL locations for the expression traits, considered individually. The triangles at the bottom are the estimated locations of the two QTL, under the two-QTL model.

On the right: the difference between the LOD score for the two-QTL model for a given cut-point and that for the single-QTL model, for each cut point.

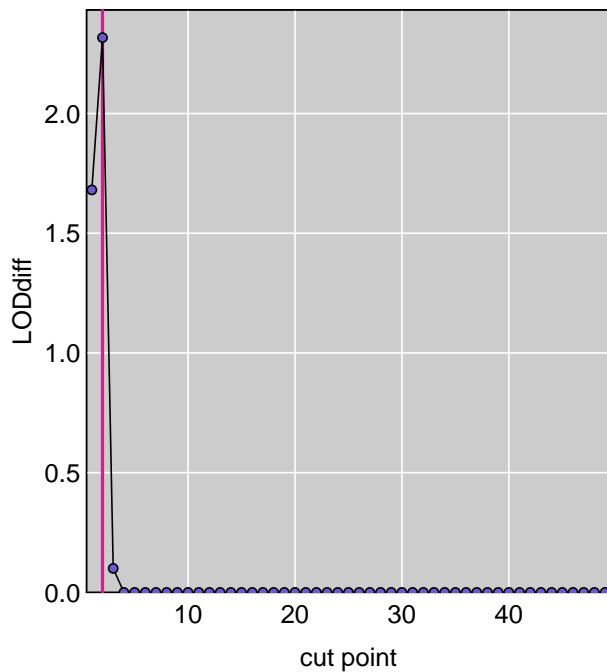
There's very strong evidence for two QTL here (LOD near 40), and a clear inference of which expression traits are affected by the left and the right QTL.

## 1 vs 2 QTL: Islet c6

### LOD profile



### LOD diff by cutpoint

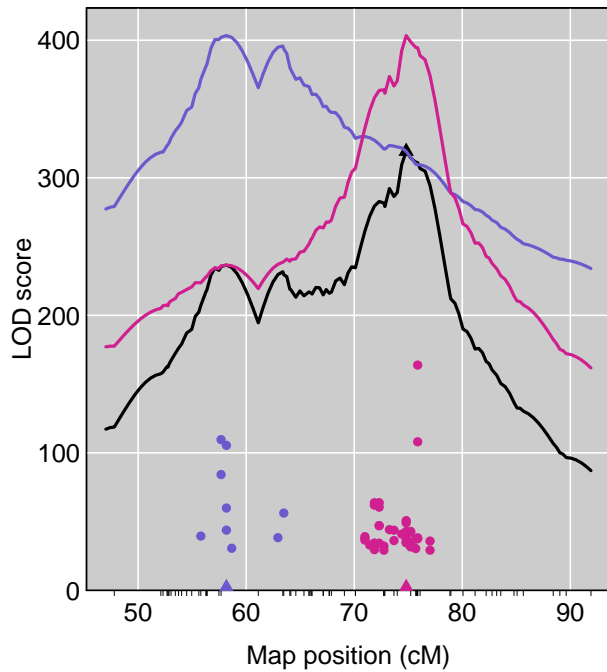


26

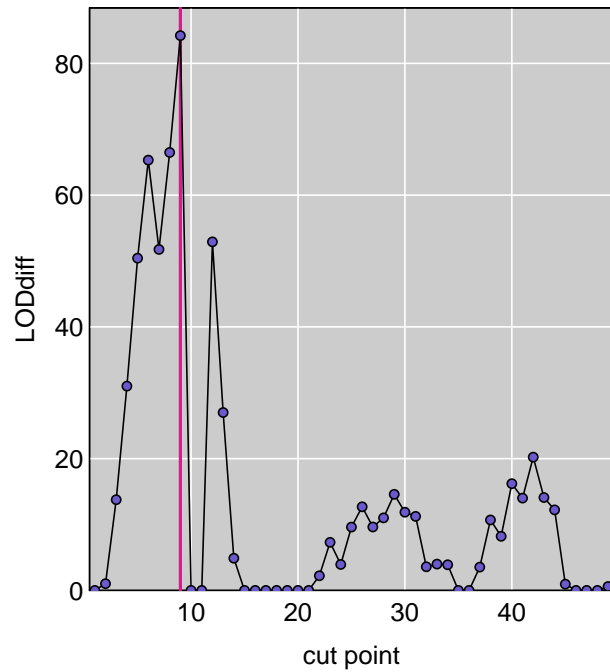
Here's the islet chromosome 6 locus. The two-QTL models are never much better than the single-QTL better. We'd conclude that there's a single QTL.

# 1 vs 2 QTL: Islet c2

## LOD profile



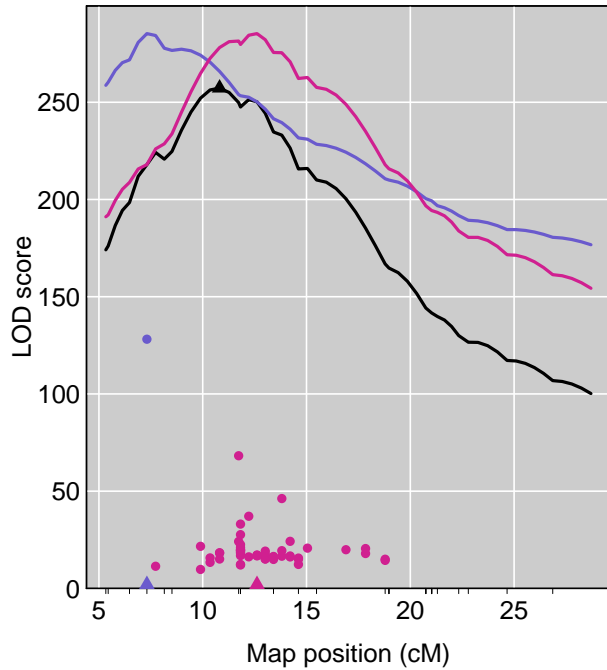
## LOD diff by cutpoint



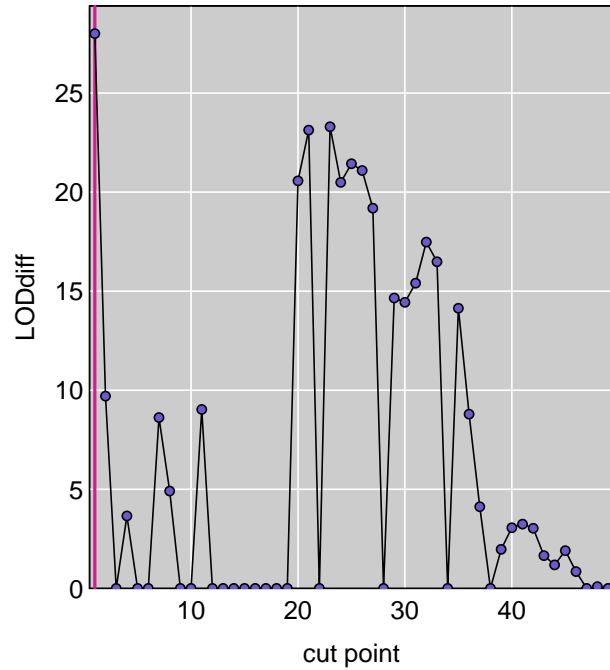
Here's islet chromosome 2. Again, very strong evidence for two QTL. It would be interesting to split off the proximal locus and study the distal locus on its own: it's likely that we'll find evidence for three QTL in this region.

# 1 vs 2 QTL: Liver c17

## LOD profile



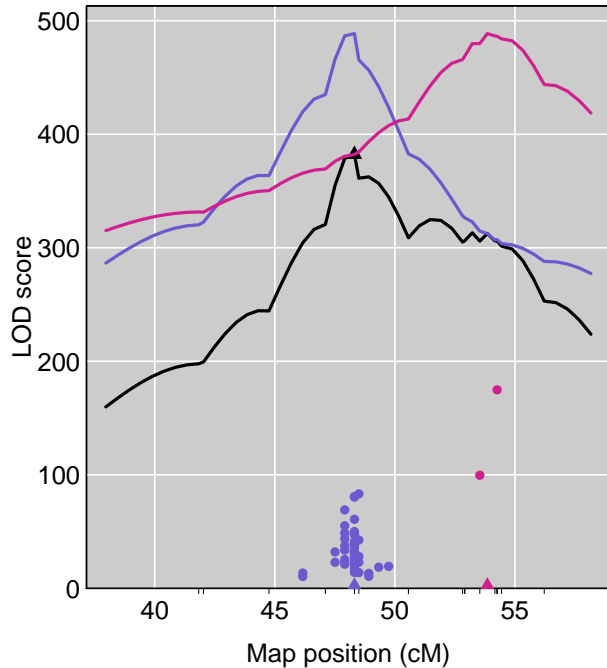
## LOD diff by cutpoint



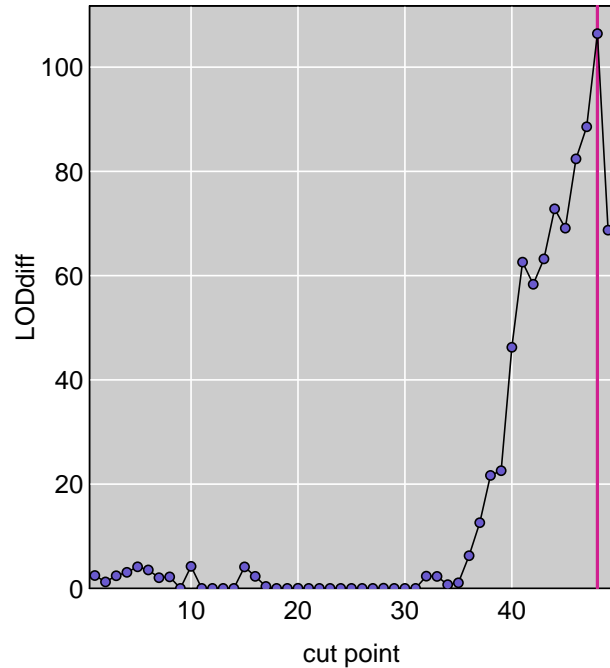
Liver chromosome 7: good evidence for two QTL, but it's splitting off just the first expression trait. We should omit this one and look for evidence for multiple QTL among the rest. The high LOD differences for the split at 23 suggests there would still be evidence for multiple QTL, among those traits.

# 1 vs 2 QTL: Adipose c10

## LOD profile



## LOD diff by cutpoint



Last one: good evidence for two QTL, but it's pulling off just a couple of expression traits. Again, it seems we might use a more focused interval.

There is, of course, room for improvement in this method. I think the key issue is how to pick which set of expression traits to focus on. The results are quite dependent on this choice. Here we've been picking out a QTL interval in a relatively arbitrary way and then choosing the 50 traits that map to that region with the largest LOD score.

## Summary

- ▶ Fine-mapping a *trans*-eQTL hotspot
  - Consider the non-recombinants
  - Predict QTL genotype of recombinants
    - Mendelian trait
  - Fine-map by traditional means
- ▶ Large-effect locus on chr 6
  - Affects expression of ~8% of genes
  - Effects specific to pancreatic islets
  - Looks to be *Slco1a6*
- ▶ Dissecting a *trans*-eQTL hotspot
  - Sign of eQTL effect
  - Degree of dominance
  - Compare recombinants and non-recombinants
  - Formal statistical test

It's always good to include a summary slide.

## Acknowledgments

### Univ. Wisconsin–Madison

Jianan Tian

Alan Attie

Mark Keller

Aimee Teo Broman

Christina Kendziorski

Brian Yandell

Angie Oler

Mary Rabaglia

Kathryn Schueler

Donald Stapleton

### Univ. Kansas Medical Center

Bruno Hagenbuch

Wen Zhao

### Merck & Co., Inc.

Amit Kulkarni

### Mt. Sinai

Eric Schadt

NIH: R01 GM074244, R01 DK066369

31

Lots of people were involved in this work.

Jianan did the bulk of the method development and all of the software development and data analysis.

The data come from Alan and Mark; Angie, Mary, Kathryn, and Donald did the work there. Aimee, Christina, and Brian are collaborating with me on the analysis.

Amit worked out the annotation information for this custom microarray.

The project was a collaboration with Eric Schadt, whose group did the microarrays and genotyping when he was at Rosetta.

Bruno and Wei worked to validate the *Slco1a6* gene.

## References

- ▶ Tian J et al. (2015) Identification of the bile acid transporter *Slco1a6* as a candidate gene that broadly affects gene expression in mouse pancreatic islets. *Genetics* 201:1253–1262  
[doi:10.1534/genetics.115.179432](https://doi.org/10.1534/genetics.115.179432)
- ▶ Tian J et al. (2016) The dissection of expression quantitative trait locus hotspots. *Genetics* 202:1563–1574  
[doi:10.1534/genetics.115.183624](https://doi.org/10.1534/genetics.115.183624)

Here are the two papers on the work.