Exploratory data analysis

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Slides: kbroman.org/BMI773/eda.pdf

This lecture concerns exploratory data analysis. Techniques for the creative investigation of data, to identify problems and generate ideas.

What is exploratory data analysis?

What is exploratory data analysis? The term comes from John Tukey. For that matter the term "data analysis" itself is from Tukey.

I think he would contrast it with say "confirmatory" data analysis. Exploratory data analysis is all about creative investigation to generate new ideas. Confirmatory data analysis is about answering specific questions.

What is exploratory data analysis?

Tukey: Looking at data to see what it seems to say.

It is important to understand what you can do
before you learn to measure how well you seem to have done it.

Here is what Tukey says in the preface of his book. He defines exploratory data analysis as "looking at data to see what it seems to say."

Uses of EDA

- Get a sense of things
- Data diagnostics (quality control)
- Hoping for an “a-ha” moment
- Following up “huh” moments

What is exploratory data analysis good for?

Personally, I’m either trying to get a sense of things (as Tukey said, figure out what is it that you can do with the data), or I’m trying to identify potential problems in the data (data cleaning).

I’m usually hoping that my explorations will lead some new insight that I wouldn’t otherwise have achieved. But in practice, I’m usually following up on some puzzling aspect of the problem.

Data diagnostics: principles

- What might have gone wrong?
- How could it be revealed?
- Make lots of plots
  - scatterplots
  - plots against time
  - consider taking logs
- Check consistency between files
- Re-calculate derived variables and check that they match
- Outliers
  - Real or error?
  - Are the results affected?
- Don't trust anyone, including yourself

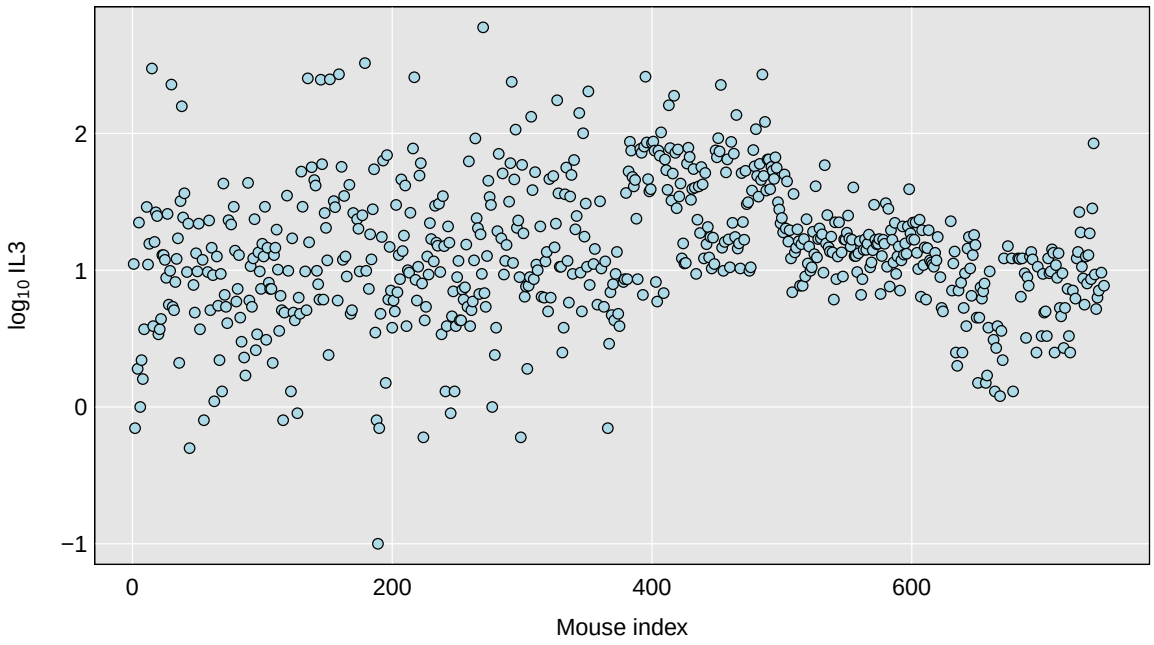
Let's start by looking at data diagnostics, sometimes called data cleaning. Our goal is to identify problems in the data, and I feel the best way to do that is to anticipate the problems and target them specifically: what might have gone wrong? how can we tell?

But further, just make lots of plots. For high-dimensional data it can be tricky. Think about how to summarize the results in ways that can reveal the sorts of odd problems. But just make lots of scatterplots and plots of variables against time. For measurements that span multiple orders of magnitude, you usually want to take logs.

Also, check consistency between files. If subjects are present in one file but missing from another, is that as expected or could part of a file be missing? If measurements are repeated in multiple files, do they match? Re-calculate any derived variables and check that they match.

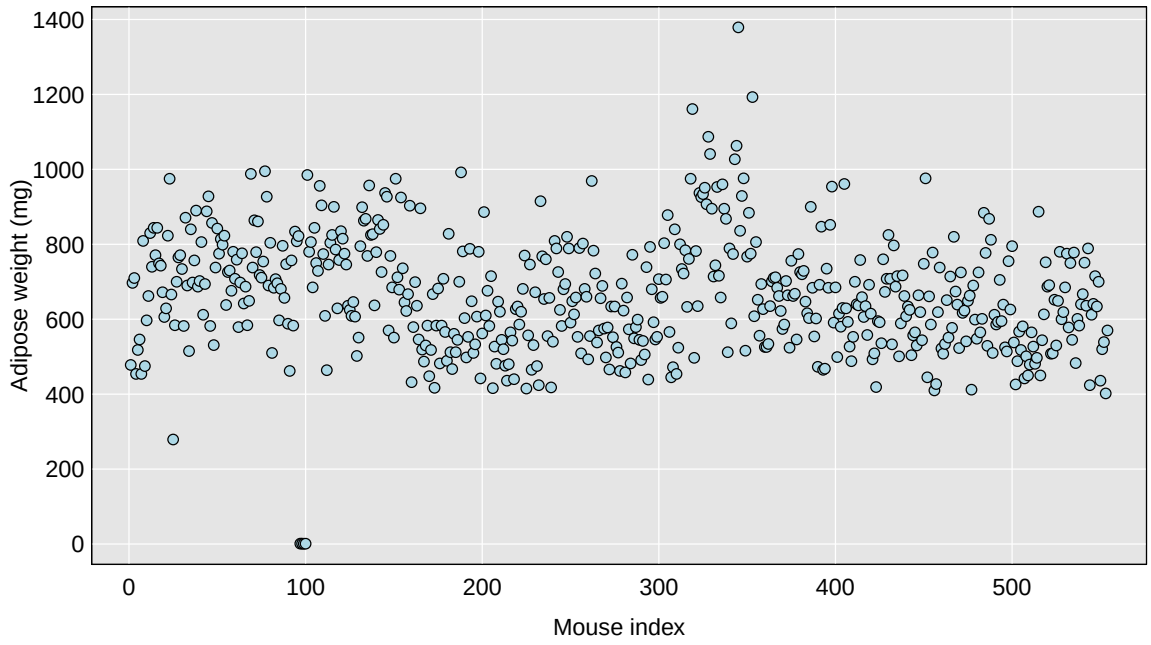
For outliers, you want to figure out if they are real or an error. Do they affect the results? If they're errors, fix them. If they're real but don't affect the results, no worries. If they're real and affect the results, worry.

Batch effect



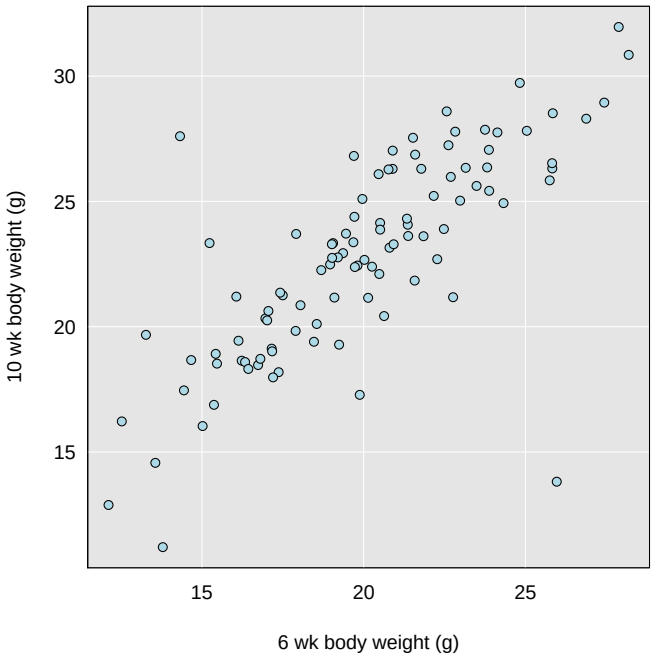
Here's an example of a clear batch effect. You can really only tell if you plot the variable by the order of measurement, and it's much more clear if you take logs.

Messed up units



Here's a case where a variable was recorded in the wrong units (g rather than mg) for a few individuals)

Outliers



In this particular case, it turned out that the day 10 weights for two subjects got swapped.

When you look at this sort of situation, ask yourself how you might find this problem if you have 20 weight measurements and 1500 individuals.

Weird stuff I've seen

- 500 worksheet excel file where the middle 100 worksheets have the variables arranged in a different order
- Weird rounding patterns
- Missing values that shouldn't be, because derived values are not missing
- Categorical data with inconsistent categories
- Missing value codes that weren't mentioned and that could be real values (e.g., 999)
- OMG dates

All kinds of things get messed up in data files. It's hard to find it if you don't look for it; you have to check.

When it comes to the order of variables in multiple files, never assume consistency; always check.

Weird rounding

37.5	89	404.04308	6.55818503449434	146.9497
41.9	90	218.343	9.55324086763758	101.9179
36	88	287.62704	4.65914900117792	91.0011
22.8	79	114.2122	32.46127	70.38872
20.8	75	166.4504	8.211126	60.96332
27.2	84	202.51284	13.1384923833842	105.07665
20.8	77	313.51314	11.1372217899707	93.32436
12.6	65	199.61718	16.7719514987531	66.61461
12.1	64	429.33954	18.9643060968415	49.52037
27.4	81	512.34846	4.31272238159915	101.51535
25.3	79	591.4965	9.70506442962546	186.98655
22	78	142.6692	14.9913480181089	53.79393
22.9	80	349.70889	17.0824838559225	180.93234
24.2	77	425.96127	5.77571495445421	151.72968
25.7	82	248.36079	14.3881991417965	99.37857
23.9	79	441.8874	17.1454129445892	70.17591
26.6	93	359.8437	11.3140598977232	152.79807
37.1	87	445.14312	10.4517	87.77684
35.3	85	183.7356	7.32103	67.86024
37.9	88	471.54792	11.8114	166.35688

Here's an example of some weird rounding in an excel file. The fonts aren't even consistent. This indicates that some copy and pasting went on, which makes me question whether there is some other master file that I should really be looking at.

Identifiers

- Are the subject IDs unique?
- Are there subject or gene IDs that don't fit the typical pattern?
  - `1e5` vs `100000`
  - hyphens turned into periods
  - IDs that became dates
- Subjects in one file but not in another and vice versa
  - Real, or messed up IDs?

IDs are really important but they can be screwed up in all kinds of ways. R can mess them up; Excel can mess them up. They can get messed up repeatedly by your collaborators, no matter how hard you work to preserve them.

Missing values

- As intended?
- Below detection limit?
- Telling you something about sample quality?
- Introducing bias?

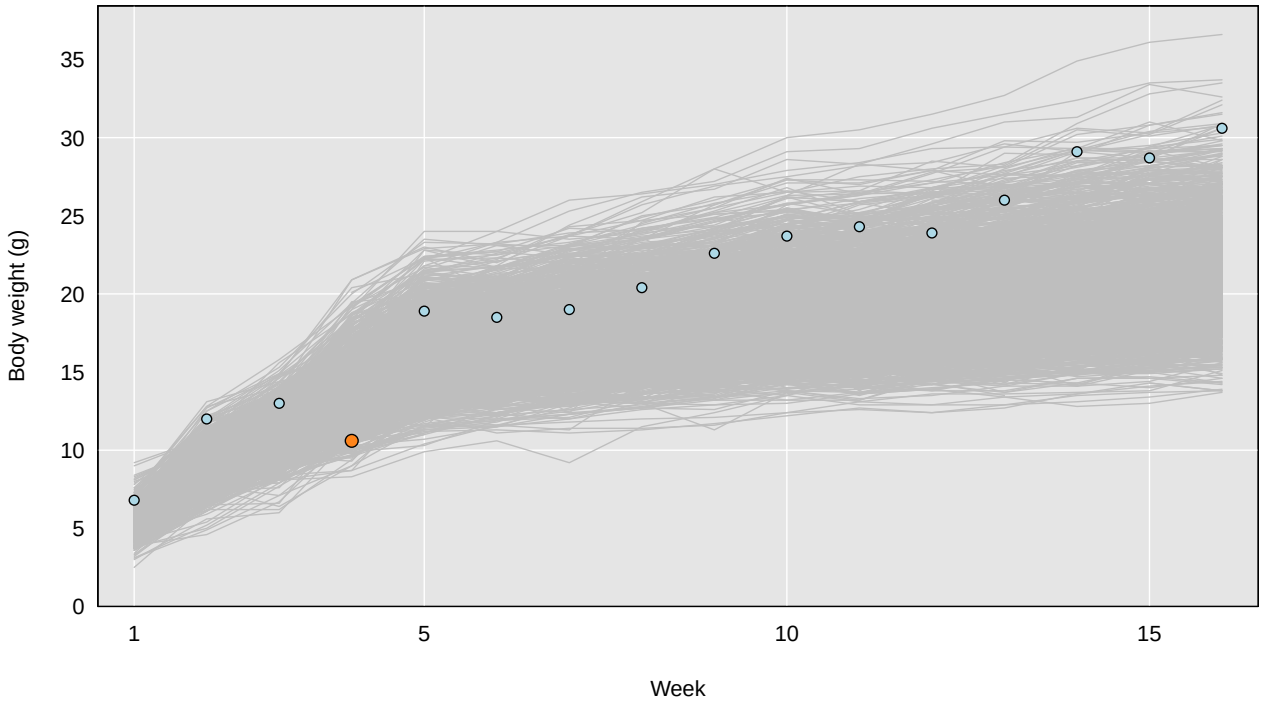
It can be important to look at the pattern of missing data. For genotyping and sequencing assays, a high rate of missing data often indicates poor quality samples.

But also, are the missing data really as intended?

Could they maybe be values below the detection limit of the assay? And does that mean that they should be just treated as small values, or omitted?

Is the nature of the missing data going to bias your conclusions?

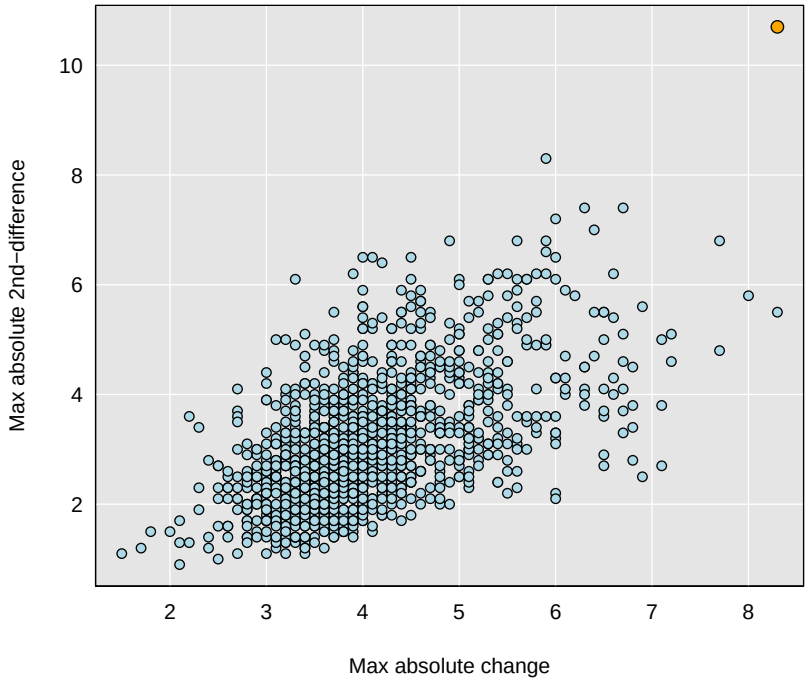
Fitting a model can be useful



Sometimes, it's helpful to fit some sort of model. Particularly if you have very large quantities of data, you could then better identify problem samples or data points, for example by looking for large residuals.

Here, we have an apparent outlier in the body weight data for a mouse. Maybe it's real; maybe it's an error. But it would be hard to even find it in the midst of data on 1200 mice.

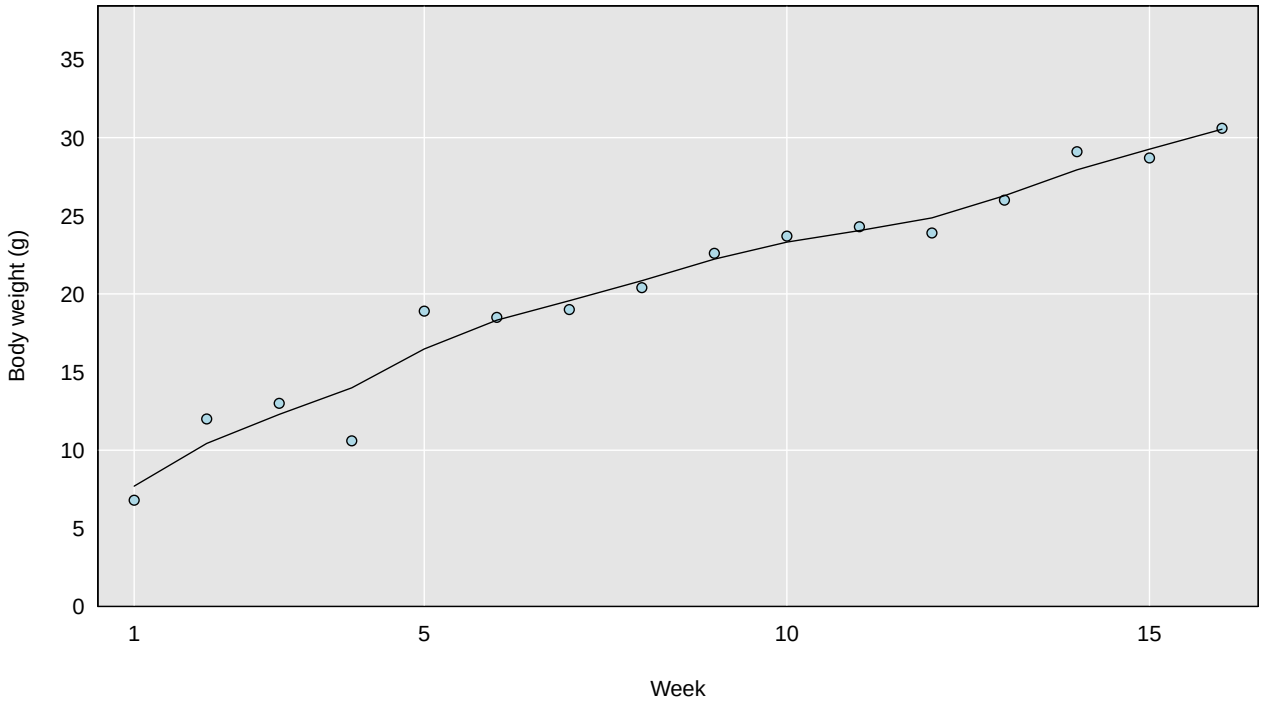
Biggest change vs 2nd difference



We can look for this sort of thing directly. For each curve, calculate the maximum absolute change in body weight and the maximum absolute 2nd-difference. Make a scatterplot of those values.

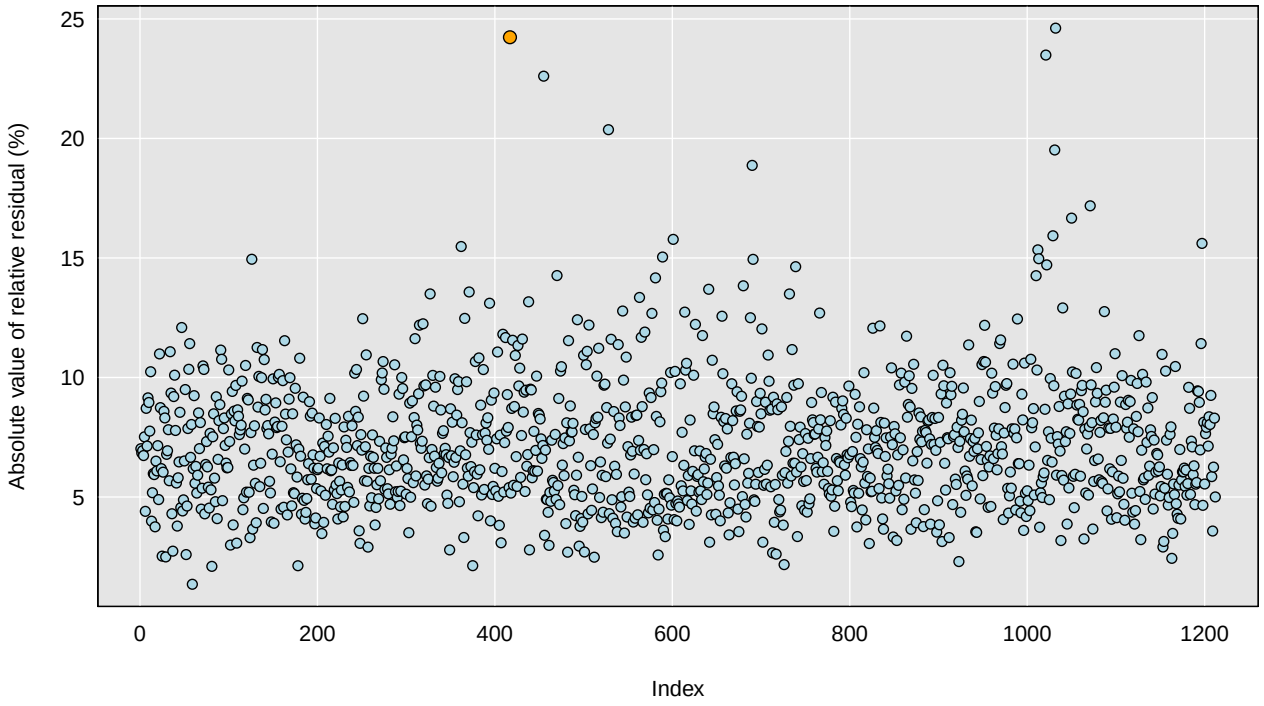
The particular individual we were looking at stands out as extreme in both ways.

Fit a smooth curve



Alternatively, we could fit a smooth curve to the data, and then look for points that deviate from the smooth, by calculating the residuals.

Residuals



For each subject, find the maximum absolute residual, relative to the fitted value.

The particular individual we were looking at is highlighted in orange, and is one of just five subjects with residuals $> 20\%$.

Follow up artifacts

They might be the most interesting results

A solid lesson I've learned is the importance of following up on artifacts.

Attie project

∼500 B6 × BTBR intercross mice, all ob/ob

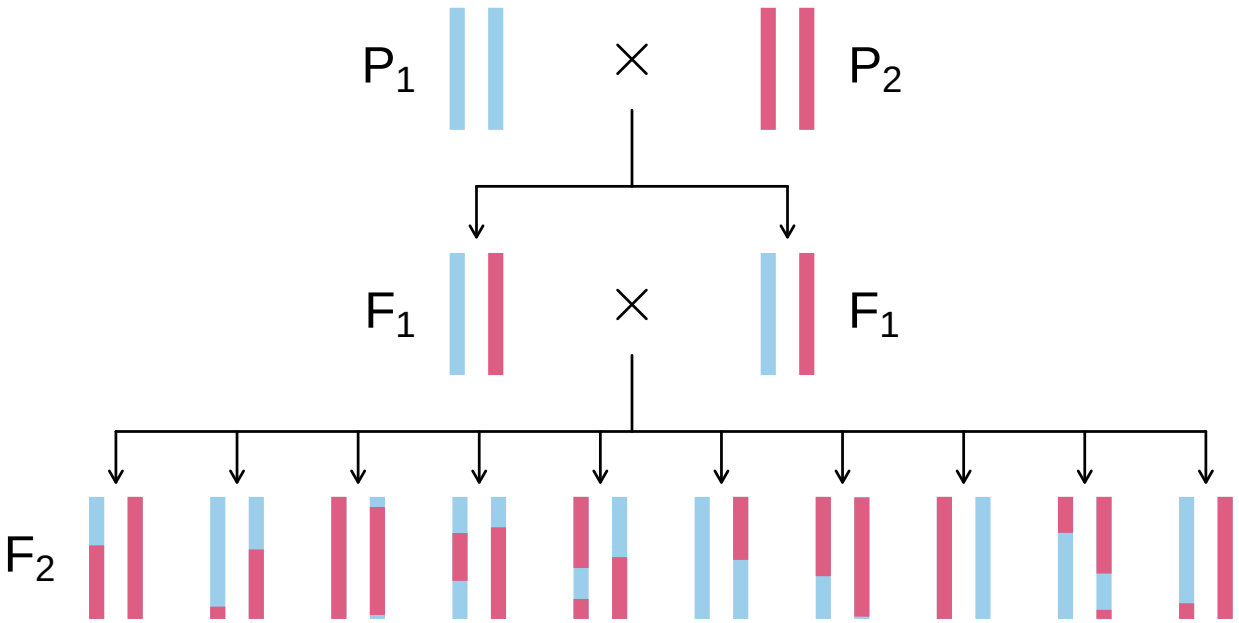
- Genotypes at 2057 SNPs (Affymetrix arrays)
- Gene expression in six tissues (Agilent arrays)
  - adipose
  - gastrocnemius muscle
  - hypothalamus
  - pancreatic islets
  - kidney
  - liver
- Numerous clinical phenotypes

  (e.g., body weight, insulin and glucose levels)

When I first got to UW-Madison, I joined a collaboration that was carrying out a very large QTL mapping experiment that included about 500 mice with dense genotype data and numerous clinical phenotypes, but also with gene expression data in six different tissues.

I had mostly been in the back of the room, heckling. But a couple of years into the project, I agreed to write the first paper.

Intercross

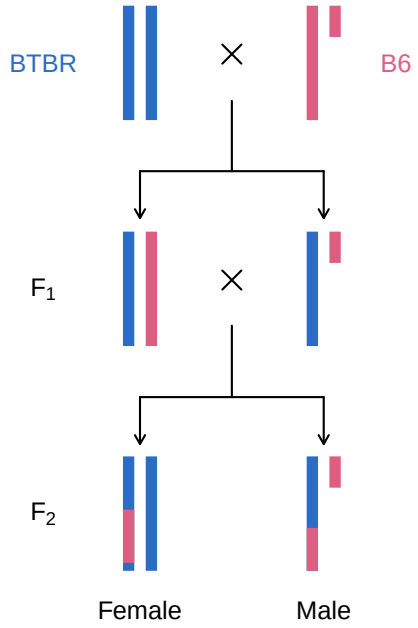


QTL mapping is an effort to identify the genetic loci that contribute to variation in some quantitative trait, like blood pressure. Such loci are called quantitative trait loci (QTL).

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.

Sex and the X chr



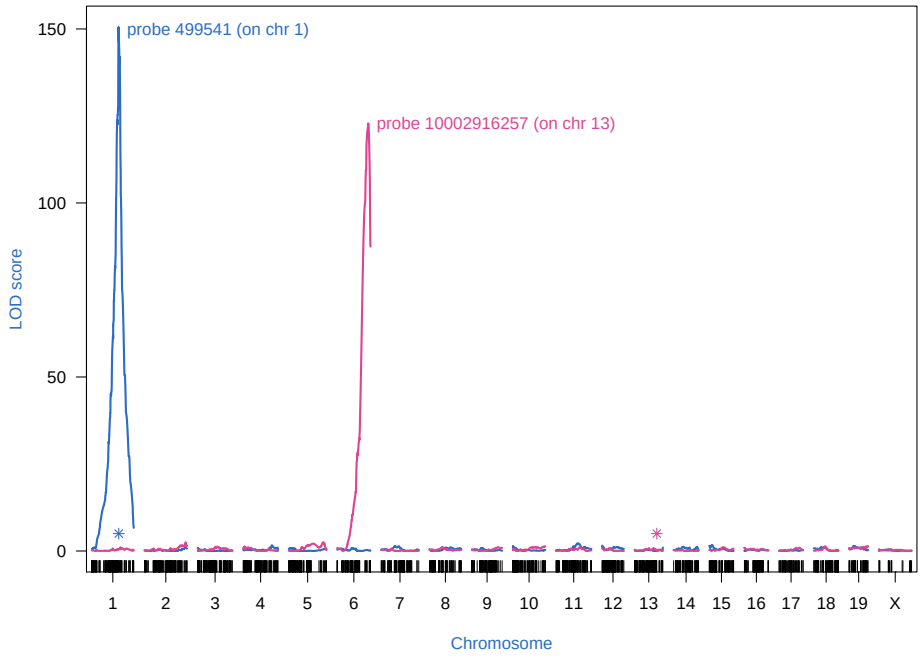
In getting ready to prepare that first paper, I decided to go back to the basics and really check that all of the data were in good order, starting from the raw genotype files.

I noticed that there were a number of mice whose X chromosome genotype data did not match their sex. The way the cross was carried out, female F_2 mice will be homozygous BTBR or heterozygous, and male F_2 mice will be hemizygous (and so look like homyzogous). But there were a number of females who were homozygous B6 on the X, and a number males who were heterozygous. (Previously, these incompatible genotypes had just been omitted.)

The number of mice with this problem ($\sim$16 out of 500) was not large, but it was more than I'd expected, and I sat and pondered how to figure out which was correct: sex or genotype.

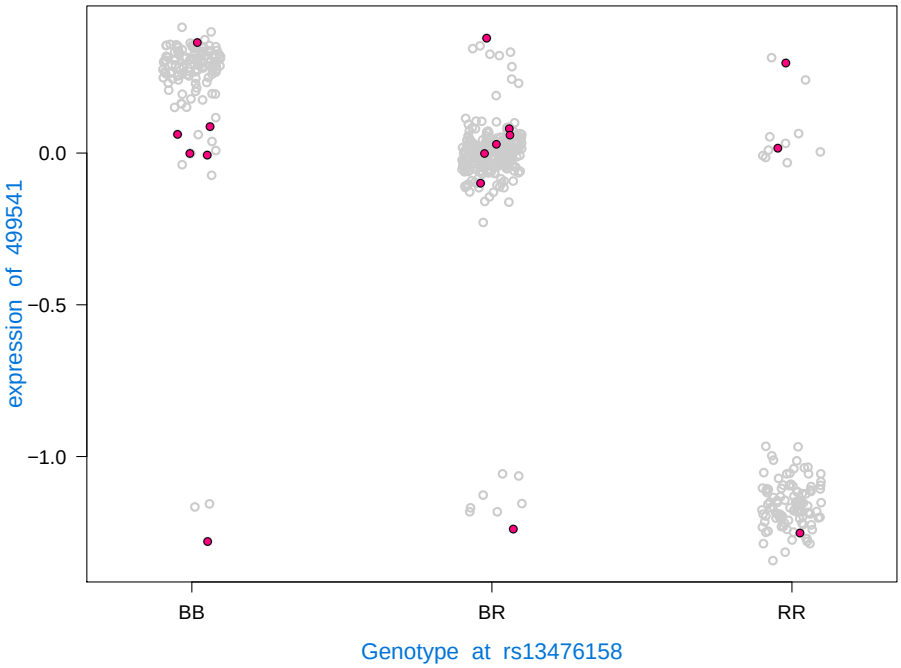
I realized that I could maybe use the gene expression data to help.

Strong eQTL



In many cases the gene expression traits have very strong genetic effects. In particular, for many genes the expression level is strongly affected by genotype right at the location of the gene. For other genes, expression is strongly affected by genotype at some other location. A locus that effects gene expression is called an expression QTL or eQTL.

E vs G



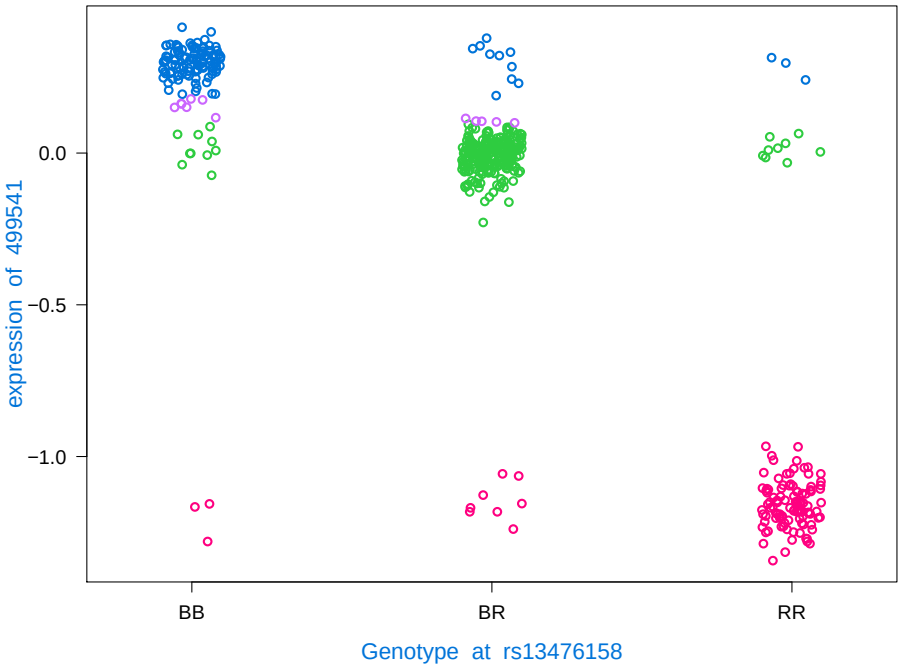
I looked at the gene expression versus genotype at one of these eQTL and saw a very strange pattern. There was a very strong association, but there were also a lot of mice whose gene expression seemed to not match their genotype.

I mean, there are basically three kinds of mice, expression-wise: low, high, or very high. And the low-expression mice are mostly RR, while the very-high mice are mostly BB, with the high-expression mice being BR. Except there are a bunch of mice that seem to be in the wrong ball, expression-wise. And the 16 six-swapped mice include 9 that are in the wrong ball.

It's like the sex-swapped mice had been assigned to a random genotype. If the genotypes are in the proportions 1:2:1, then we'd expected 3/8 to be correct just by chance, which is very similar to the 7/16 we see in these data.

And note that there are 43 mice that look to be in the wrong ball. If they are all being assigned genotypes at random, that would suggest that there are like $43 \times (8/3) \approx 115$ problem mice.

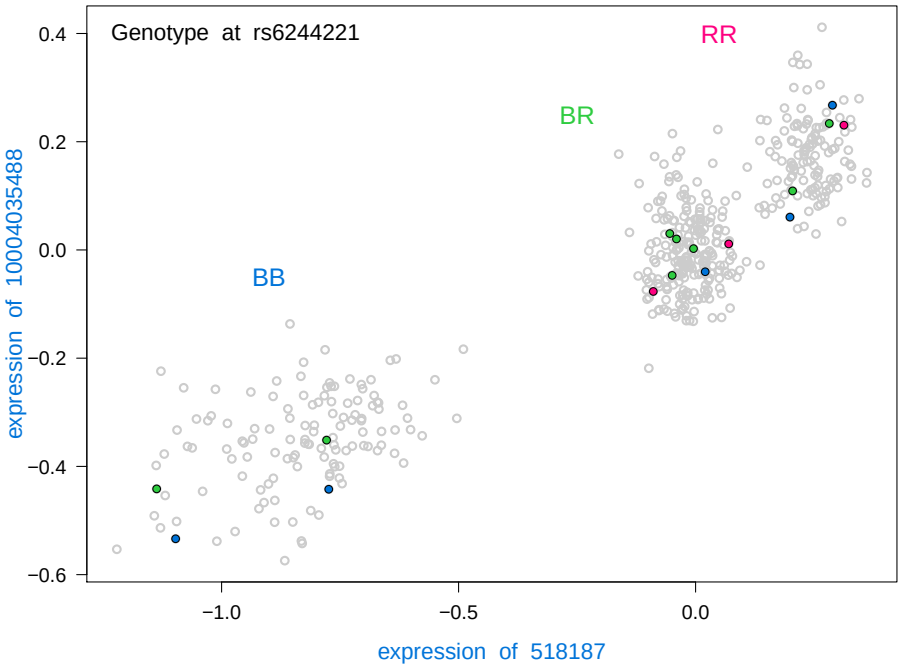
kNN classifier



But we can use the gene expression data to figure out what we think each mouse's genotype at this location really is. For example, we can create a k-nearest-neighbor classifier, for predicting genotype from gene expression.

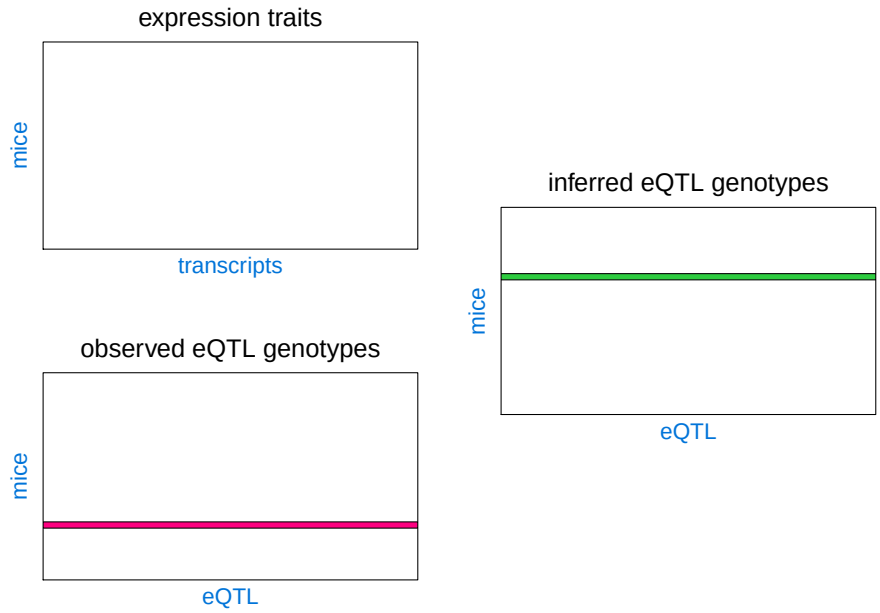
If we do this at many strong eQTL, we could potentially reconstruct the true genotypes for each mouse, from their expression data.

E vs G



Many times there will be two different genes whose expression maps to a common location. We can look at their expression jointly. In many cases, the gene expression clusters are even more clear. And again the sex-swapped mice are seen in the wrong ball with frequency like 9/16.

Basic scheme

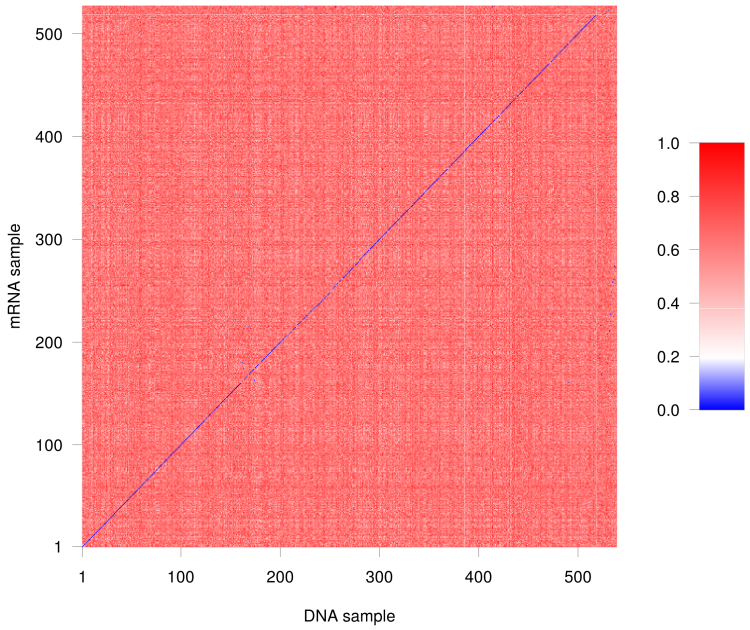


So this leads to our basic scheme for identifying (and correcting) the sample mix-ups.

We first identify a set of expression traits with very strong eQTL. We use the expression and corresponding eQTL genotypes to form classifiers for predicting eQTL genotype from gene expression. This gives us a matrix of inferred eQTL genotypes.

We then compare the inferred eQTL genotypes to the observed eQTL genotypes. If a sample's observed eQTL genotypes don't match its inferred eQTL genotype, we conclude that the labels for one or the other are incorrect. And we might be able to find another row in the inferred eQTL genotypes that matches its observed genotypes.

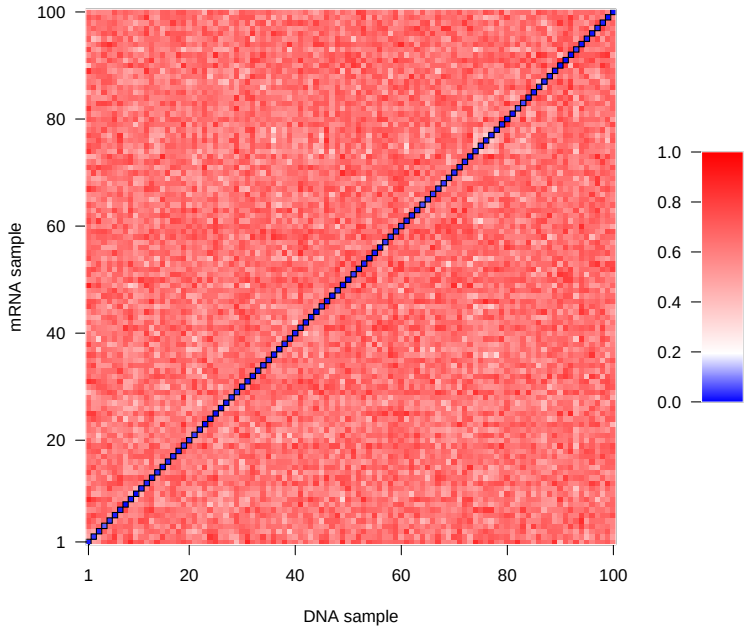
Prop'n mismatches



For each pair of samples, one DNA (genotype) sample and one RNA (gene expression) sample, we get a measure of distance as the proportion of mismatches between the observed eQTL genotypes and the inferred eQTL genotypes.

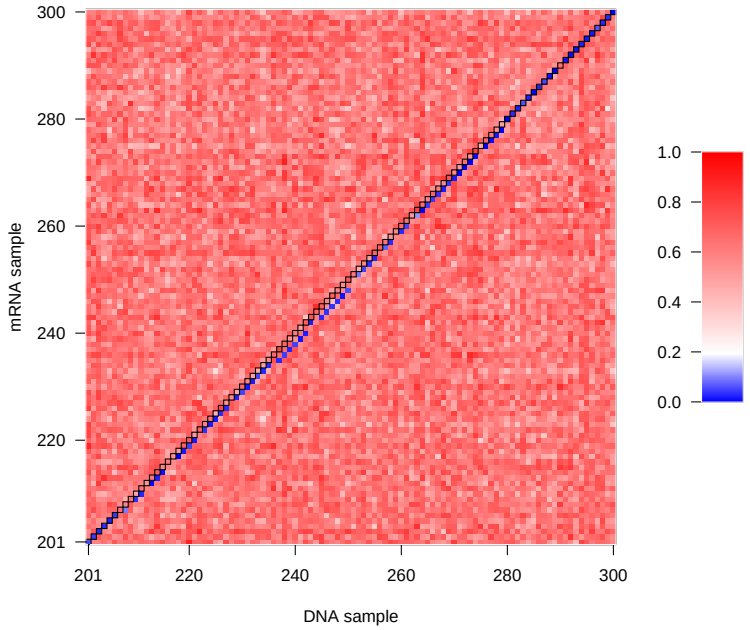
Here's a picture of this distance matrix. It should be blue along the diagonal and red everywhere else.

Prop'n mismatches



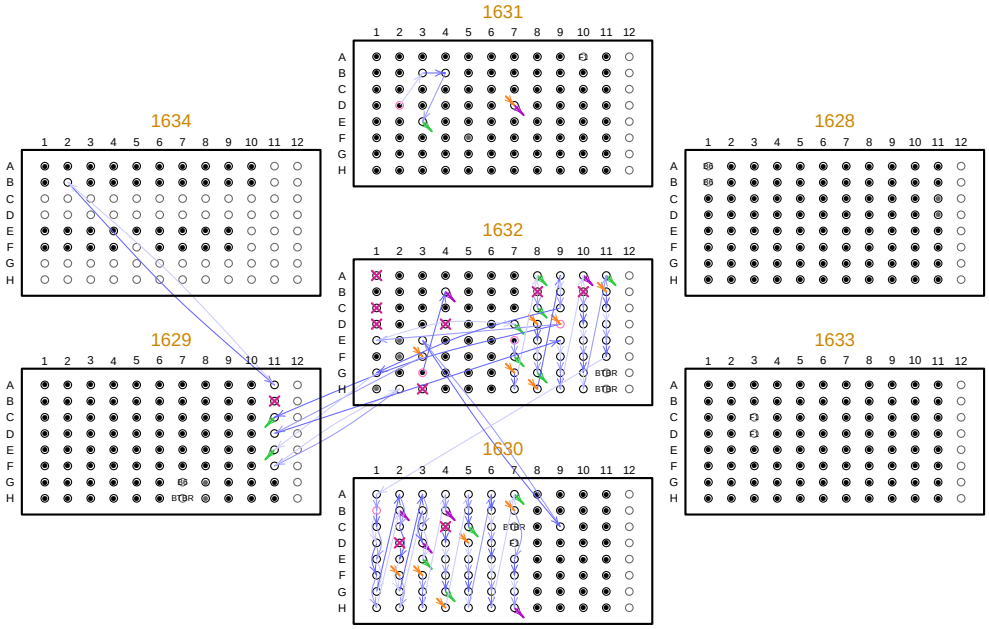
And if we look at the first 100 samples, that's exactly what we see: the samples are close to themselves and not to anyone else.

Prop'n mismatches



But if we look at the middle 100 samples, we find a whole bunch of off-by-one and off-by-two errors. The samples are quite different from the corresponding one, but their close to the one next to it or two over.

Genotype mix-ups



Even more incriminating, though, is the information about the locations of the DNA samples. DNA samples were arrayed in a set of six 8×12 plates. In this figure, the black dots indicate the correct DNA sample was placed in the correct well, while the arrows point from where a DNA sample should have been to where it actually ended up.

Two of the plates look fine, while half of each of two plates are entirely messed up.

Plate 1631

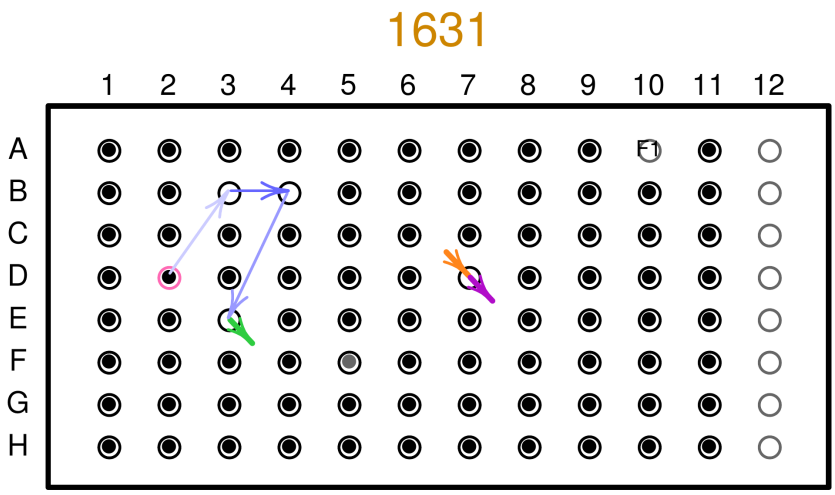


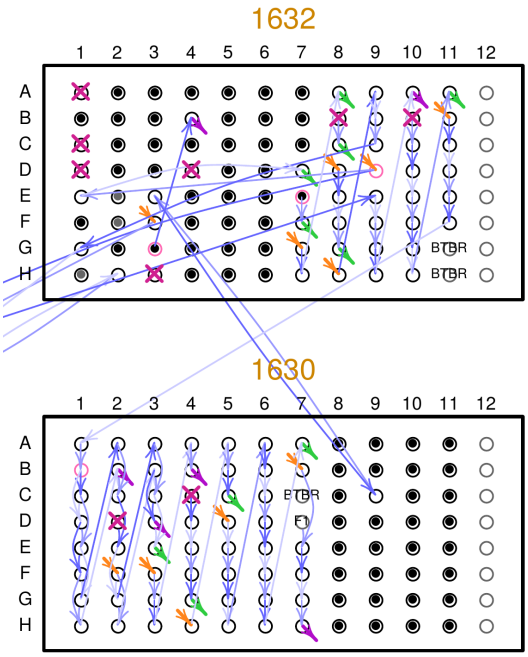
Plate 1631 is a good example. Again, black dots indicate that the correct DNA was placed in the correct well.

The little orange and purple arrow heads indicate that sample in well D7 is of unknown origin, and the sample that should have been there was lost.

The pink circle around D2 indicates that that sample was duplicated: it was placed in the correct well (the black dot), but it was also placed in well B3. The sample that was supposed to be in B3 was placed in B4, the sample that was supposed to be in B4 was in E3, and the sample that was supposed to be in E3 was lost.

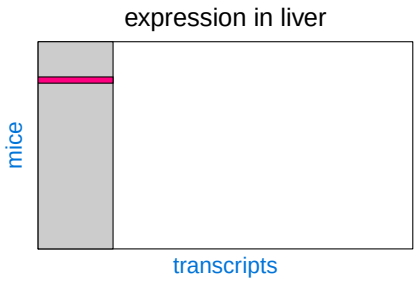
(The purple arrow head for D7 means that the DNA was lost but that there is expression data for that sample, while the green arrow head for E3 means that the DNA was lost but there is no expression data for that sample.)

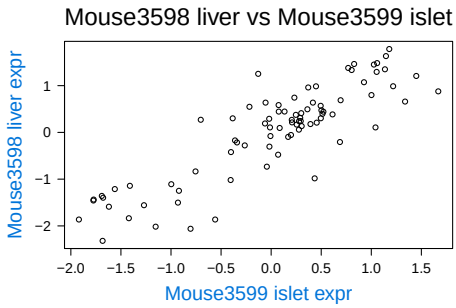
Plates 1632 and 1630



Plates 1632 and 1630 are where most of the problems are. There are some long-range swaps and other misplacements of samples, but most of the problems are due to a series of off-by-one and off-by two errors. Note that the red X's indicate DNAs that were omitted due as being of bad quality (possibly mixtures).

E vs E





We can use the same trick to look for mix-ups among the gene expression data sets.

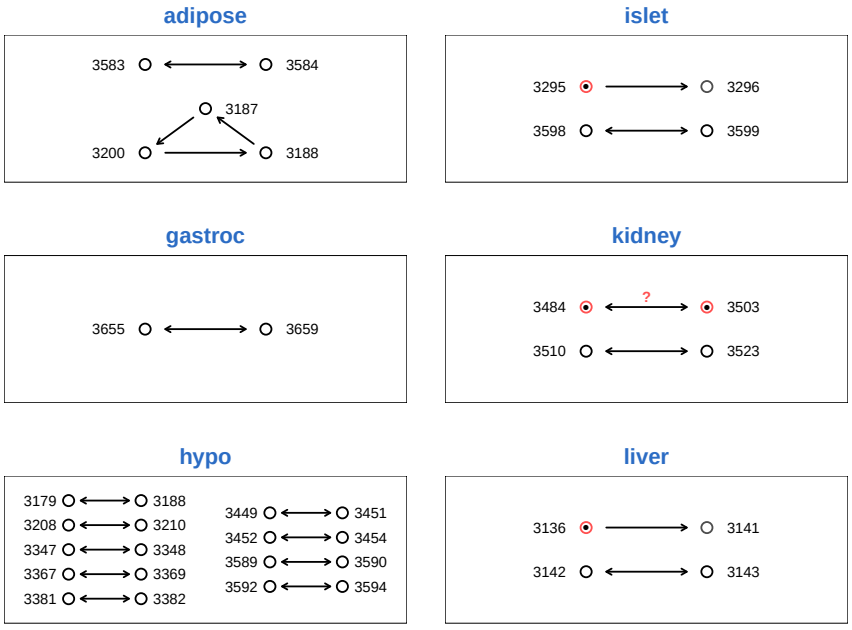
The basic scheme is to first identify a subset of expression traits that are highly correlated between two tissues.

Then look at the correlation between samples, using just that subset of expression traits.

When a sample is correctly labeled in both tissues, the expression values should be correlated. If not, we may find another sample in one tissue that is correlated, to indicate the true label.

Again, we make use of the multiple tissues to figure out the truth. If we had just two tissues we could see that they were mixed up but not which was the correct label.

Expression mix-ups



Here are the set of mix-ups I found in the expression data. The arrows point from the correct label to how it appeared.

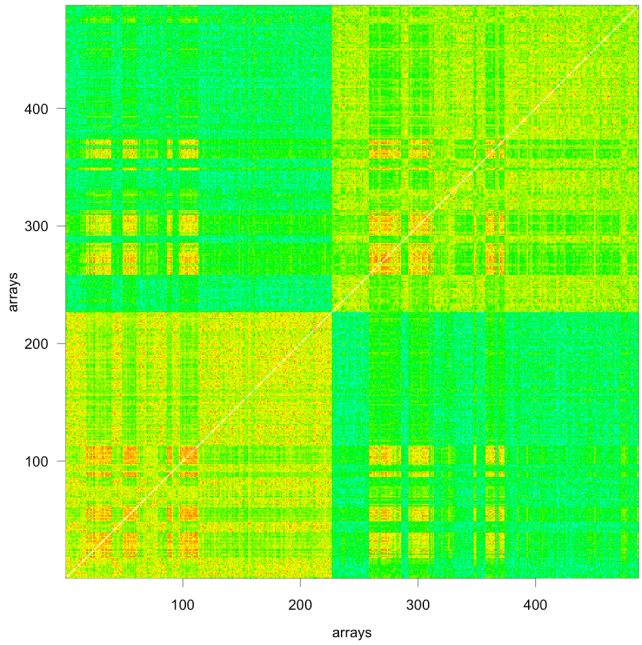
Each tissue had some mistakes; hypothalmous was the worst. The pink circles indicate a sample duplicate. So, for example, in islet sample 3295 was correctly labeled but also appeared in duplicate with one sample labelled as 3296. The 3296 islet sample was lost.

Adipose had a 3-way swap. 3187 was labelled as 3200 which was labelled as 3188 which was labelled as 3187. Note that most of the problems concern sample numbers that are close (but not necessarily immediately adjacent) in number.

The general idea here has wide application for high-throughput data, generally. If you have mutiple rectangles of data whose rows are supposed to correspond, you should check to see if they do correspond. The strategy we used for aligning two expression datasets could work with little change in much broader contexts.

Remember: all of these mistakes, including the 20% sample mix-ups in the DNA, were discovered by following up on a set of just 16 samples (out of about 500) whose sex didn't match their X chromosome genotype.

Another example

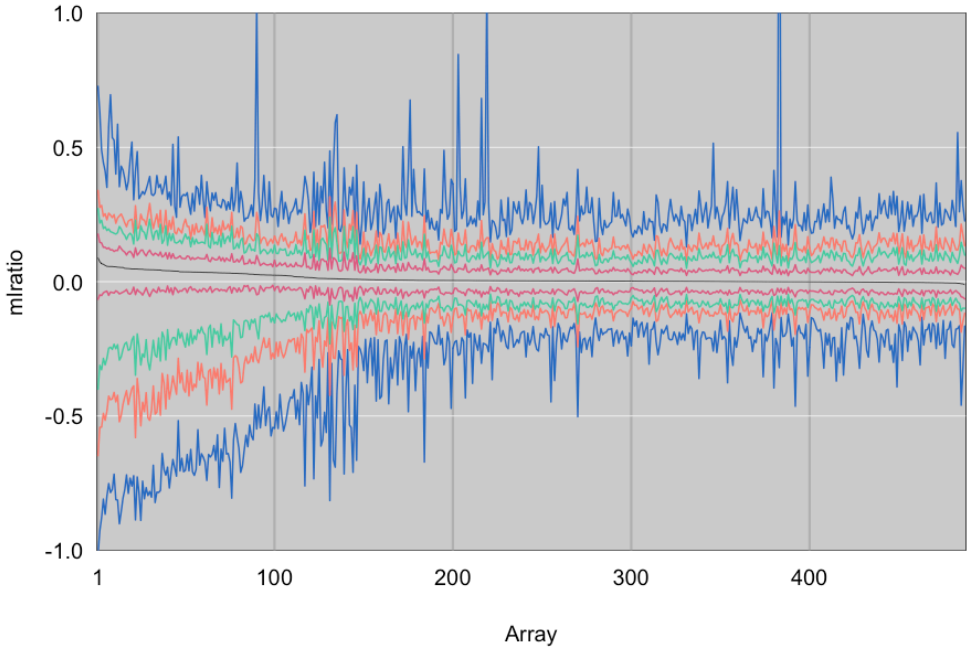


kbroman.org/blog/2012/04/25/microarrays-suck

This is a correlation matrix for a set of microarrays. What the heck is going on?

A problem here is that we hadn't really done much QC on the arrays. For example, we hadn't really examined the distributions of values on each array. It's hard to look at 500 histograms. If we had 50, we'd have looked at all of them, but since we had 500, we didn't look at any.

Dense box plots



This is like a set of 487 boxplots of the array data, sorted by their median. The black line is at the median. The pink lines are at the 25th and 75th percentiles. The green, orange and blue lines are at the 10, 5, and 1 percentiles.

It turned out that there were a batch of 120 badly-behaved arrays.

Follow up artifacts

They might be the most interesting results

Another story to emphasize the importance of following up on artifacts.

Am. J. Hum. Genet. 63:861–869, 1998

Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination

Karl W. Broman,[1] Jeffrey C. Murray,[2,3] Val C. Sheffield,[2,4] Raymond L. White,[5] and James L. Weber[1]

[1]Marshfield Medical Research Foundation, Marshfield, WI; Departments of [2]Pediatrics and [3]Biology, University of Iowa, and [4]Howard Hughes Medical Institute, Iowa City; and [4]Eccles Institute for Human Genetics, University of Utah, Salt Lake City

Summary

Comprehensive human genetic maps were constructed on the basis of nearly 1 million genotypes from eight CEPH families; they incorporated >8,000 short tandem-repeat polymorphisms (STRPs), primarily from Généthon, the Cooperative Human Linkage Center, the Utah Marker Development Group, and the Marshfield Medical Research Foundation. As part of the map building process, 0.08% of the genotypes that resulted in tight double recombinants and that largely, if not entirely, represent genotyping errors, mutations, or gene-conversion events were removed. The total female, male, and sex-averaged lengths of the final maps were 44, 27, and 35 morgans, respectively. Numerous (267) sets of STRPs

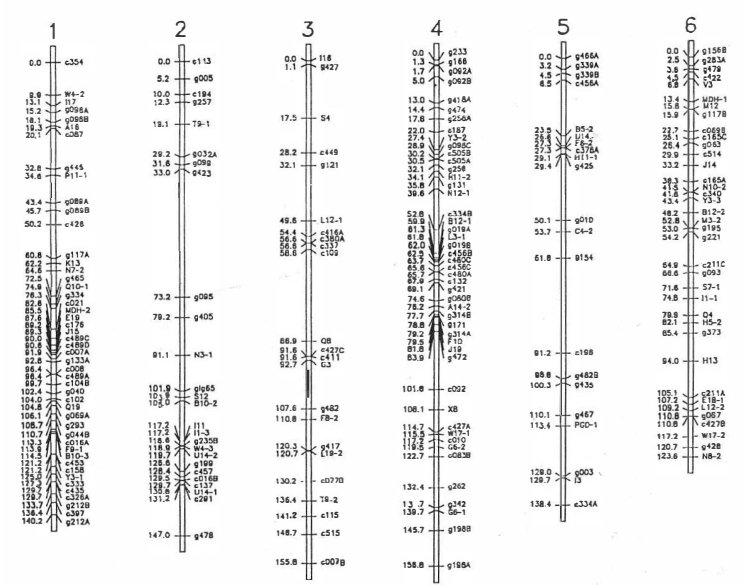
Introduction

Polymorphic DNA markers and their corresponding maps are an essential resource for localization of genes via linkage analysis, for characterization of meiosis, and for providing a foundation for the construction of physical maps. Although physical maps, including genome sequences, can provide the order of tightly linked polymorphisms, the physical maps do not provide genetic distances or other recombination data.

The era of human genome-scale genetic-map construction was heralded by the landmark paper by Botstein et al. (1980), in which both the use of DNA polymorphisms, as opposed to protein polymorphisms or other measurable phenotypes, in linkage mapping and an ef-

After I finished my PhD, I did a postdoc with a geneticist, Jim Weber, at the Marshfield Clinic. My central project was to develop new human genetic maps.

Eucalypt genetic map



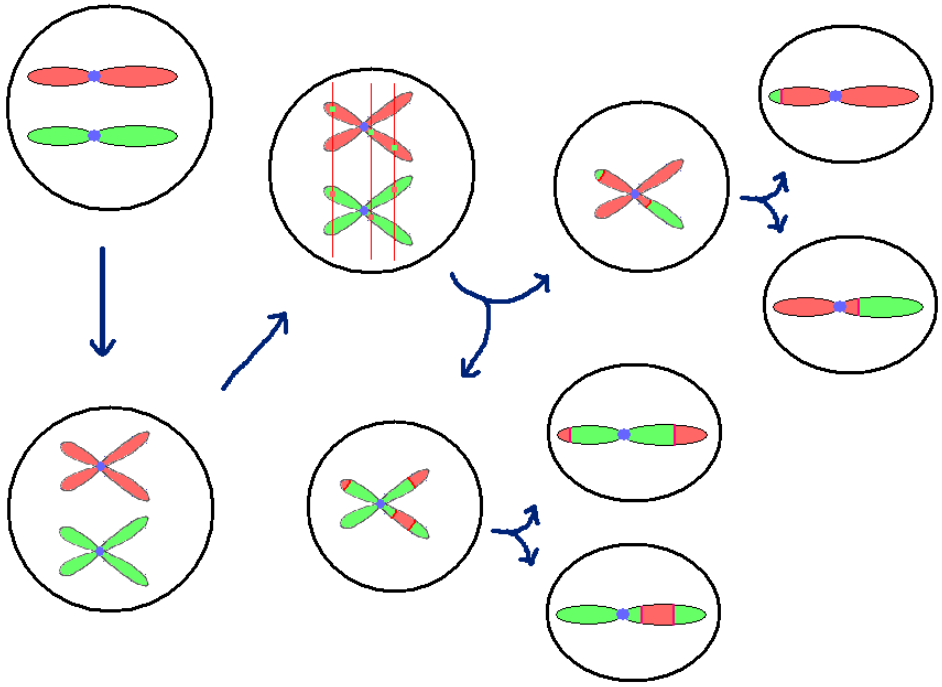
Byrne et al., Theor Appl Genet 91:869–875, 1995

A genetic map specifies the order of a set of markers along chromosomes.

This is part of a genetic map for eucalyptus trees. It is the first map that I had looked at in detail.

The original genetic maps were for observeable mutations, in Drosophila (fruit flies). Later markers were more directly DNA-based, and really chosen due to the convenience of measurement.

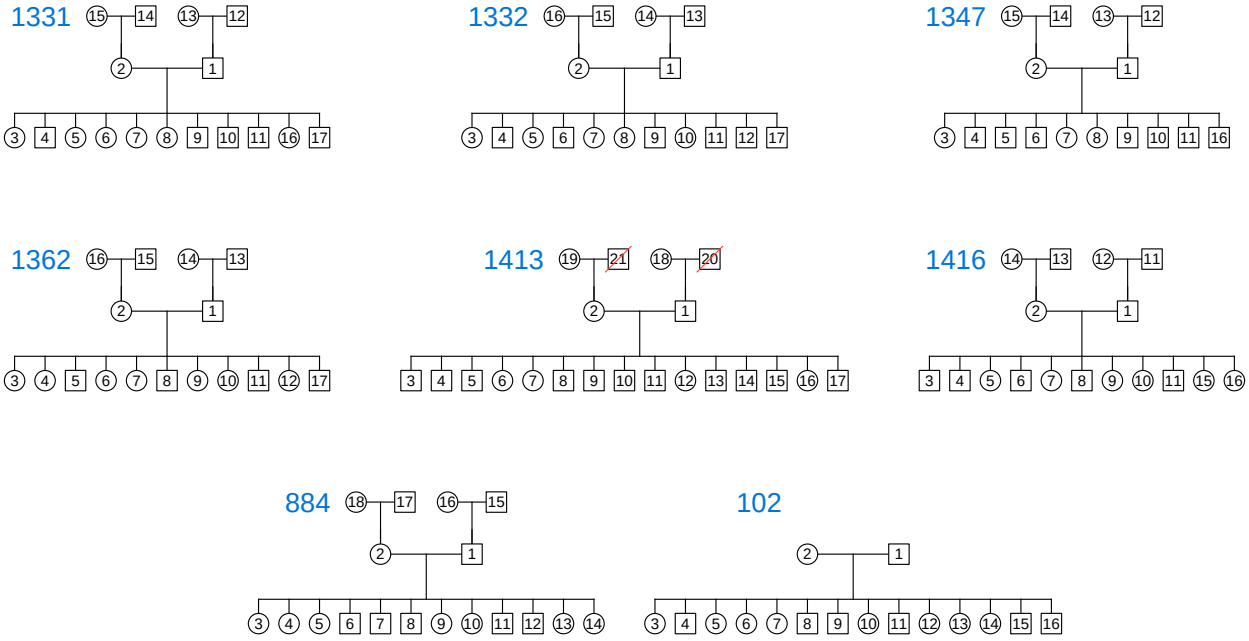
Meiosis



Distances on a genetic map are according to recombination at meiosis. Meiosis is the cell division process that produces sperm and egg cells. DNA duplicates, and then homologous chromosomes find each other and become intimately associated with each other and then actually exchange material at locations called chiasmata. Two cell divisions later you have gametes with one copy of each chromosome, which will generally be mosaics of the original chromosomes, with the points of exchange called crossovers.

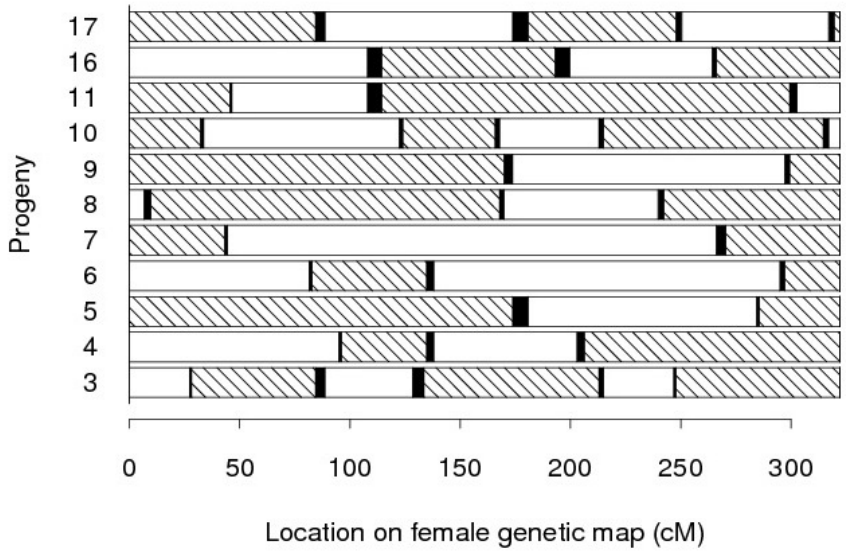
Distance on a genetic map is measured by the frequency of crossovers. Two points are d cM apart if there is an average of d crossovers in the interval per 100 meiotic products.

CEPH pedigrees



In my postdoc, I focused on data on a set of large 8 human families. A mother/father pair with 10-15 offspring. Most of the families also included data on the grandparents.

Crossover locations



Broman and Weber, Am J Hum Genet 66:1911–1926, 2000

What I was really interested in was crossover interference: the tendency of the crossovers to not be too close together on chromosomes. The open and hatched segments here are the grandmother's and grandfather's DNA, and the black bars are the intervals in which crossovers occurred.

I wanted to look at was this dependence in crossover locations.

Am. J. Hum. Genet. 66:1911–1926, 2000

Characterization of Human Crossover Interference

Karl W. Broman and James L. Weber

Marshfield Medical Research Foundation, Marshfield, WI

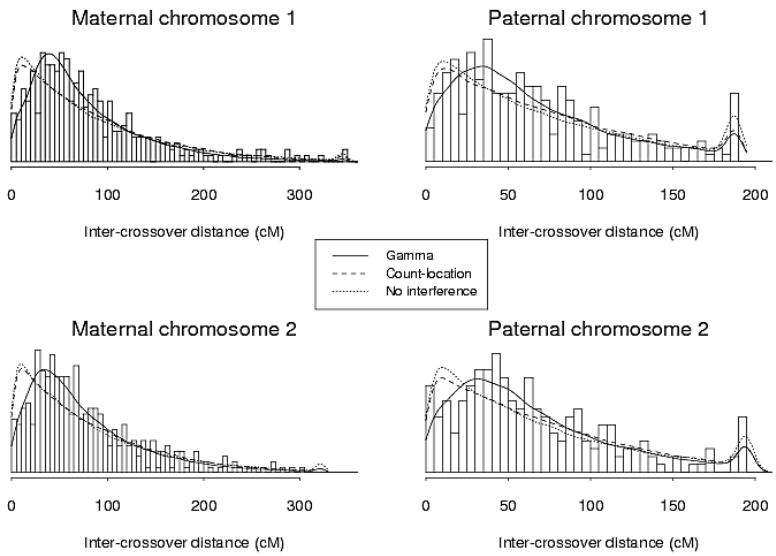
We present an analysis of crossover interference over the entire human genome, on the basis of genotype data from more than 8,000 polymorphisms in eight CEPH families. Overwhelming evidence was found for strong positive crossover interference, with average strength lying between the levels of interference implied by the Kosambi and Carter-Falconer map functions. Five mathematical models of interference were evaluated: the gamma model and four versions of the count-location model. The gamma model fit the data far better than did any of the other four models. Analysis of intercrossover distances was greatly superior to the analysis of crossover counts, in both demonstrating interference and distinguishing between the five models. In contrast to earlier suggestions, interference was found to continue uninterrupted across the centromeres. No convincing differences in the levels of interference were found between the sexes or among chromosomes; however, we did detect possible individual variation in interference among the eight mothers. Finally, we present an equation that provides the probability of the occurrence of a double crossover between two nonrecombinant, informative polymorphisms.

Introduction

Crossover interference may be defined as the nonrandom placement of crossovers along chromosomes in meiosis. Interference was identified soon after the development of the first working models for the recombination process (Sturtevant 1915; Muller 1916). Strong evidence for matid interference is a dependence in the choice of strands involved in adjacent chiasmata. There is little consistent evidence for the presence of chromatid interference in experimental organisms (Zhao et al. 1995*a*), and any inference with regard to chromatid interference generally requires that data be available for all four products of meiosis (so-called "tetrad data");

I did then get to my analysis of crossover interference (the tendency of crossovers to not be too close together).

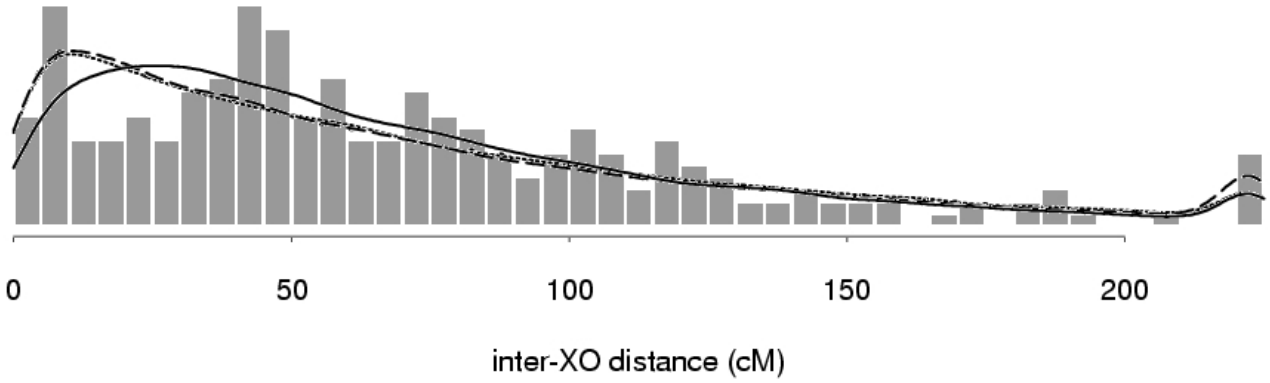
Crossover interference



Broman and Weber, Am J Hum Genet 66:1911–1926, 2000

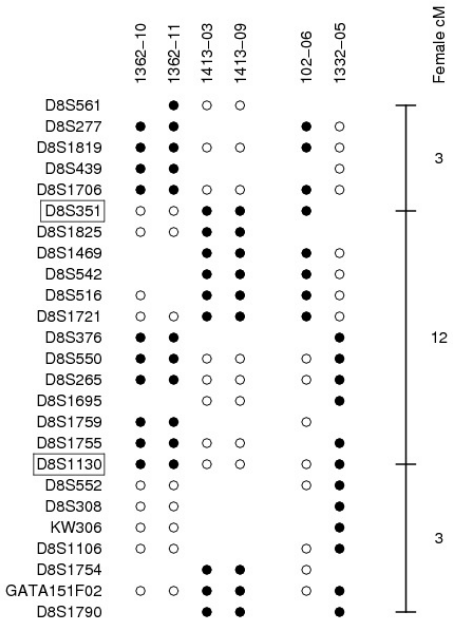
A main part of the result concerned fitting different models to the inter-crossover distance data. One model fit much better than others.

Maternal chr 8



But on one particular chromosome (maternal chromosome 8), my favorite model really didn't fit well at all.

Apparent triple XOs



Broman et al., In: *Science and Statistics: A Festschrift for Terry Speed*, 2003 45

I could have just left it at that, but I was curious about what was going on, and in studying the problem, I found that there were two families that showed an apparent triple-crossover event in a small region. This really shouldn't happen.

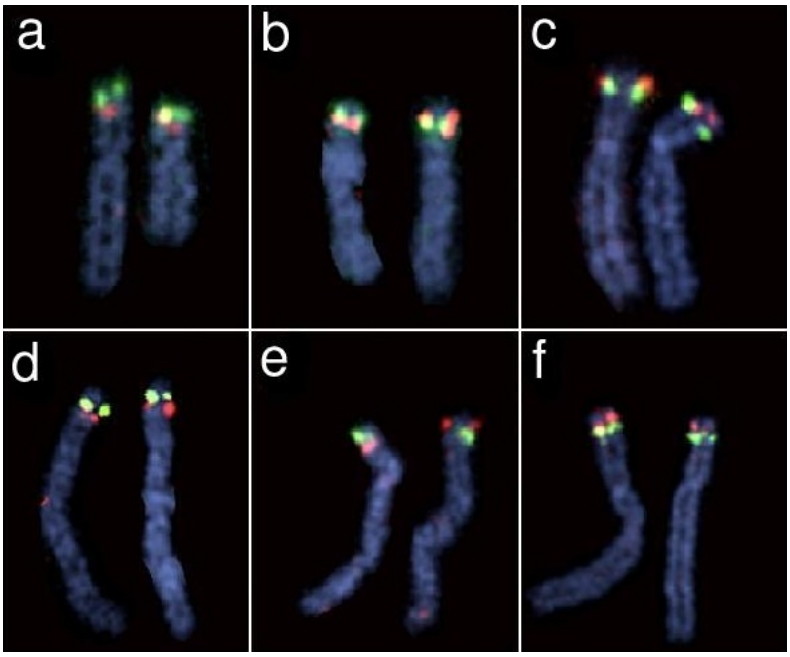
My initial reaction was that I had the marker order messed up; if I were to invert this region, the triple crossovers would become single crossovers.

But there were other families that showed a crossover in the region. If I invert the region, these single crossovers will become triple crossovers.

So then I thought: suppose the region is inverted in these two families but not in the other families? This was a pretty crazy idea, because the region is quite large (12 cM, which turned out to be about 5 Mbp), and we would need individuals to be homozygous for each of the two orientations to have recombination occur.

So a crazy idea: a very long inversion polymorphism where the two orientations were each reasonably common.

Chr 8p inversion



Broman et al., In: *Science and Statistics: A Festschrift for Terry Speed*, 2003

I posed the hypothesis to my postdoc advisor, who talked to a friend whose lab had the ability to investigate this sort of thing, and sure enough, we had discovered the largest common inversion polymorphism in the human genome.

This picture shows chromosome 8 with the green and red lighting up the two ends of the region. On the left, green is above red on both chromosomes. On the right, red is above green on both chromosomes, and in the middle green is above red on one chromosome and red is above green on the other.

So this is the best possible example of the importance of following up artifacts. Lack of model fit for a particular chromosome led me to investigate the cause of the problem, which led me to postulate this idea of an inversion polymorphism, which really seemed kind of crazy at the time. But it turned out to be real, and it's the coolest thing I've discovered in all my work as a data scientist.

Capturing EDA

- what were you trying to do?
- what you're thinking about?
- what did you observe?
- what did you conclude, and why?

We want to be able to capture the full outcome of exploratory data analysis.

But we don't want to inhibit the creative flow. How to capture this stuff?

Avoid

- “How did I create this plot?”
- “Why did I decide to omit those six samples?”
- “Where (on the web) did I find these data?”
- “What was that interesting gene?”

I’ve said all of these things to myself.

Basic principles

Step 1: slow down and document.
Step 2: have sympathy for your future self.
Step 3: have a system.

I can't emphasize these things enough.

If you're not thinking about keeping track of things, you won't keep track of things.

One thing I like to do: write a set of comments describing my basic plan, and then fill in the code afterwards. It forces you to think things through, and then you'll have at least a rough sense of what you were doing, even if you don't take the time to write further comments.

Capturing EDA

- copy-and-paste from a script
- grab code from the log (e.g., `.Rhistory`)
- Write an informal report (R Markdown or Jupyter)
- Write code for use with the KnitR function `spin()`

  Comments like `#' This will become text`
  Chunk options like so: `#+ chunk_label, echo=FALSE`

The creative flow in data exploration is something I don't want to stifle, but it's really important to capture the work so that it can be later reproduced.

There are a number of techniques you can use to capture the EDA process. You don't need to save all of the figures, but you do need to save the code and write down your motivation, observations, and conclusions.

I usually start out with a plain R file and then move to more formal R Markdown. knitr::spin() seems an interesting alternative, when you're writing more code than text.

If you torture the data long enough,
it will confess to anything.

– Tukey

When you do find something interesting, it's important to keep in mind the set of things that you looked at. Don't jump in with a statistical test at the end; this will be especially hard to do in an exploratory context.

The more things you explore, the greater the chance that you'll find something interesting that is really just chance association.