Exploratory data analysis

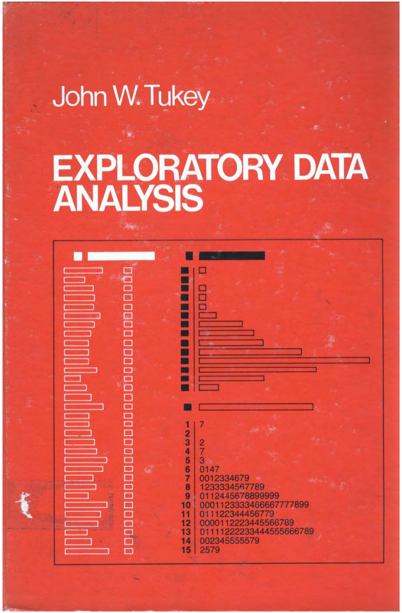
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

`kbroman.org`
`github.com/kbroman`
`@kwbroman`

Slides: `kbroman.org/BMI773/eda.pdf`

What is exploratory data analysis?



John W. Tukey

EXPLORATORY DATA ANALYSIS

```
 1 | 7
 2 |
 3 | 2
 4 | 7
 5 | 3
 6 | 0147
 7 | 0012334679
 8 | 1233334567789
 9 | 0112445678899999
10 | 00011233333466667777899
11 | 011122344456779
12 | 0000112223445566789
13 | 0111122222334445556666789
14 | 002345555579
15 | 2579
```

What is exploratory data analysis?

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

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.

Uses of EDA

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

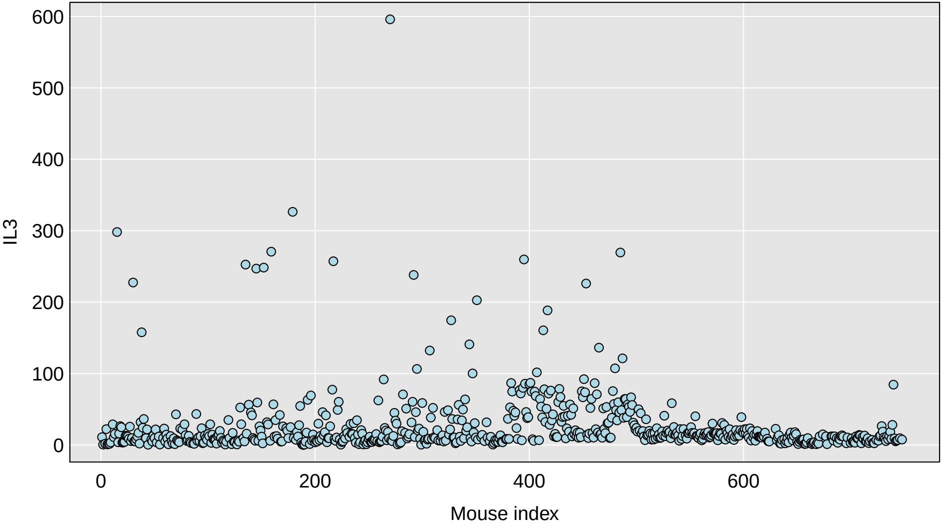
Data diagnostics: principles

- 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?

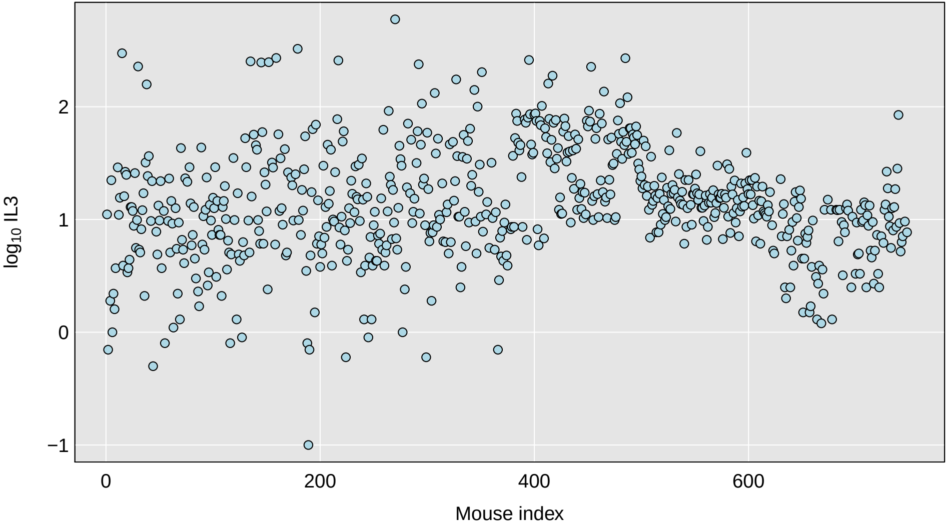
Data diagnostics: principles

- 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

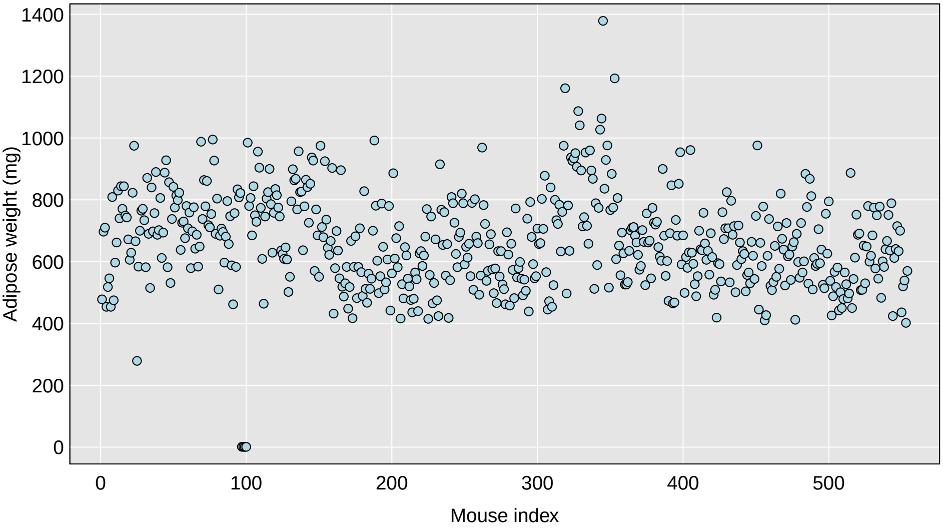
Batch effect



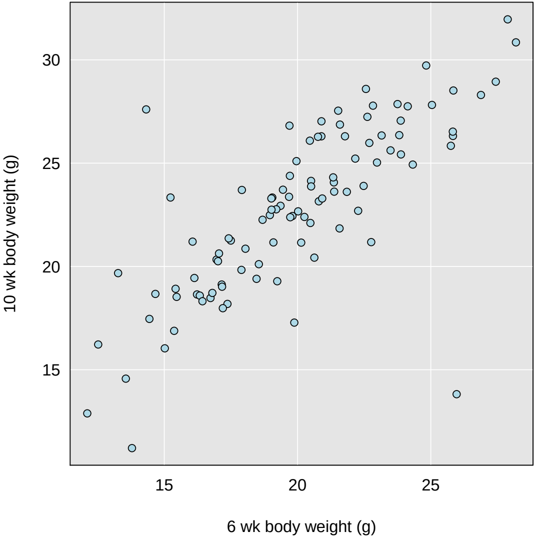
Batch effect



Messed up units



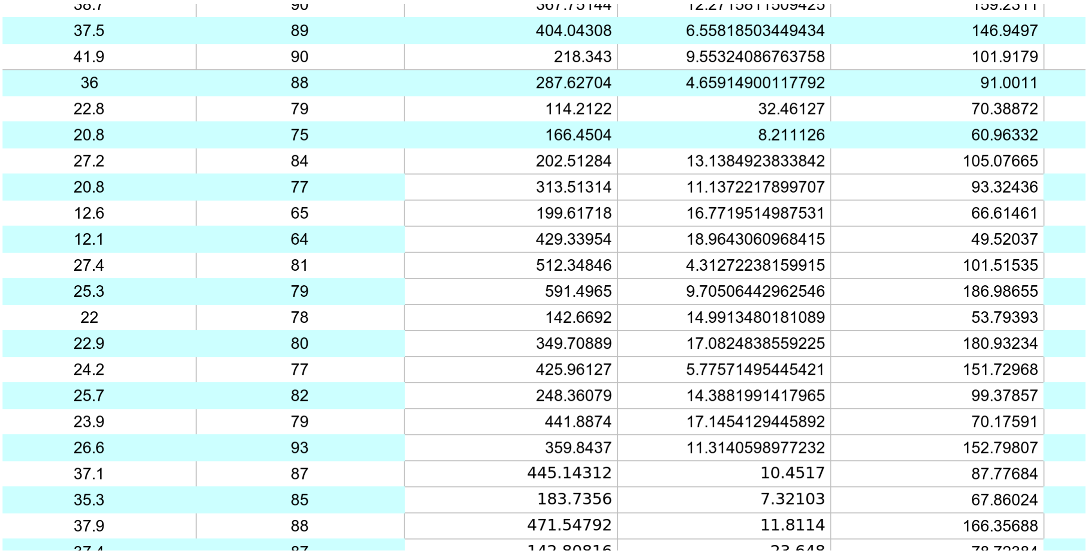
Outliers



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

Weird rounding



38.7	90	367.75144	12.2715811509425	159.2311
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
37.4	87	142.80816	23.648	78.72384

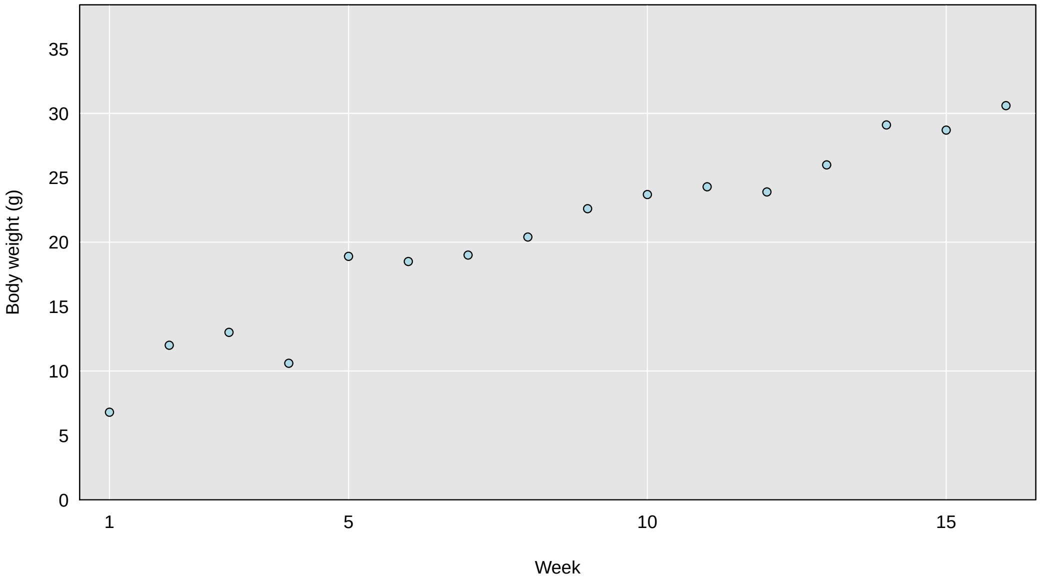
Identifiers

- Are the subject IDs unique?
- Are there subject or gene IDs that don’t fit the typical pattern?
  - `1e6` 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?

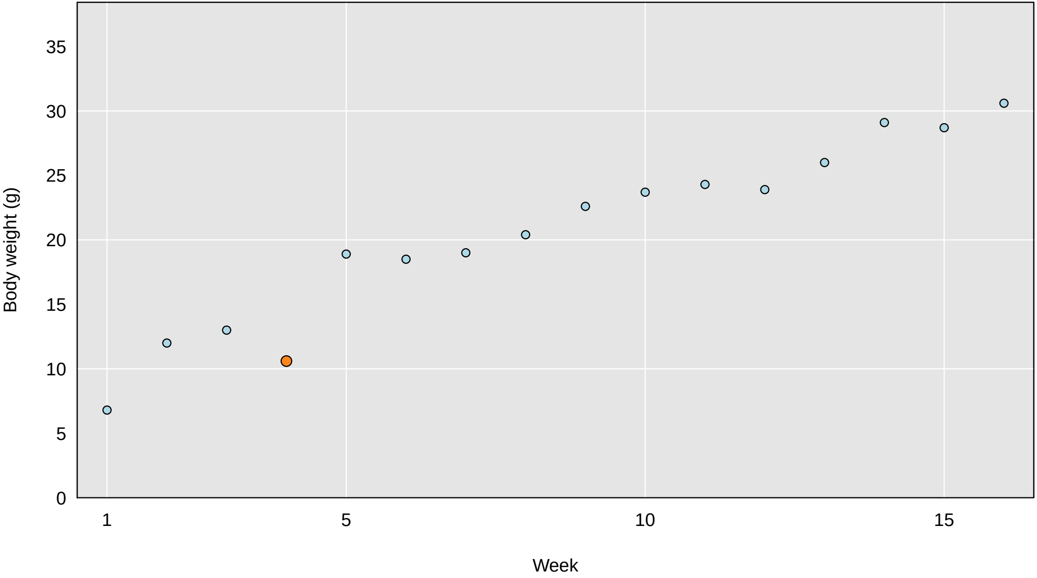
Missing values

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

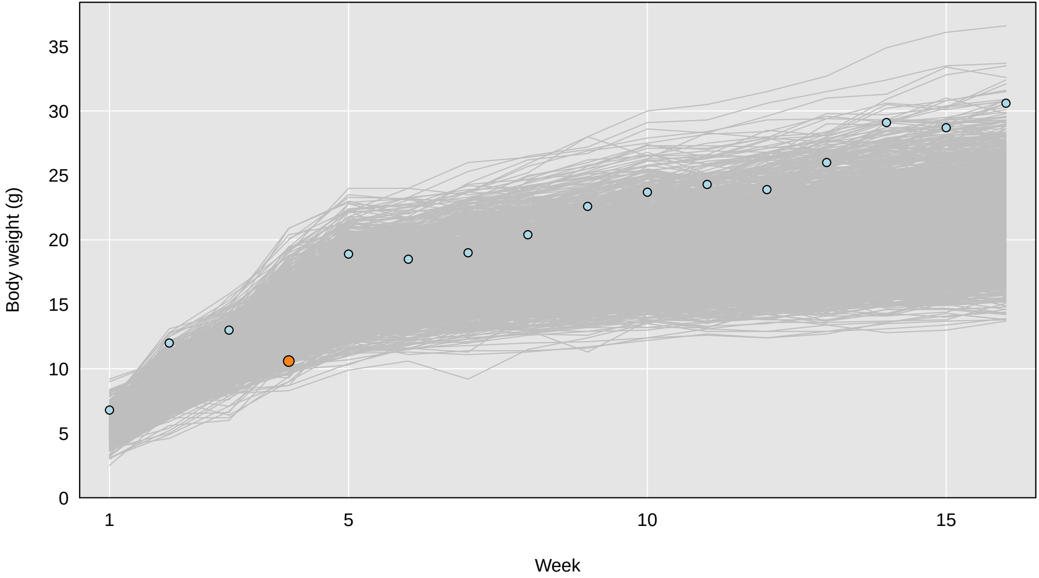
Fitting a model can be useful



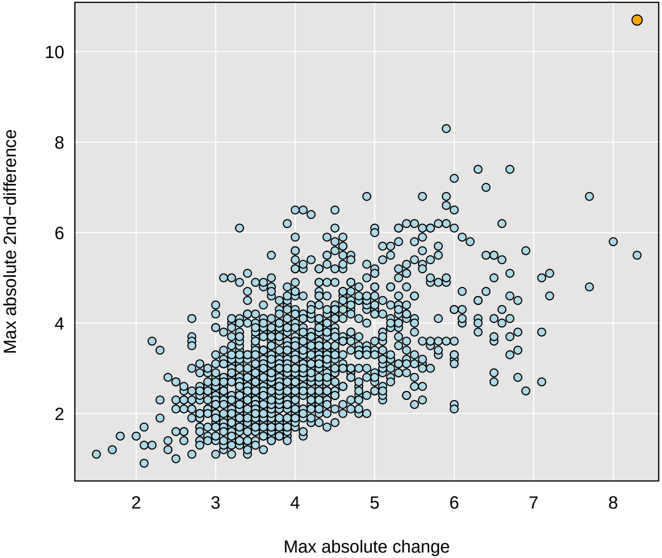
Fitting a model can be useful



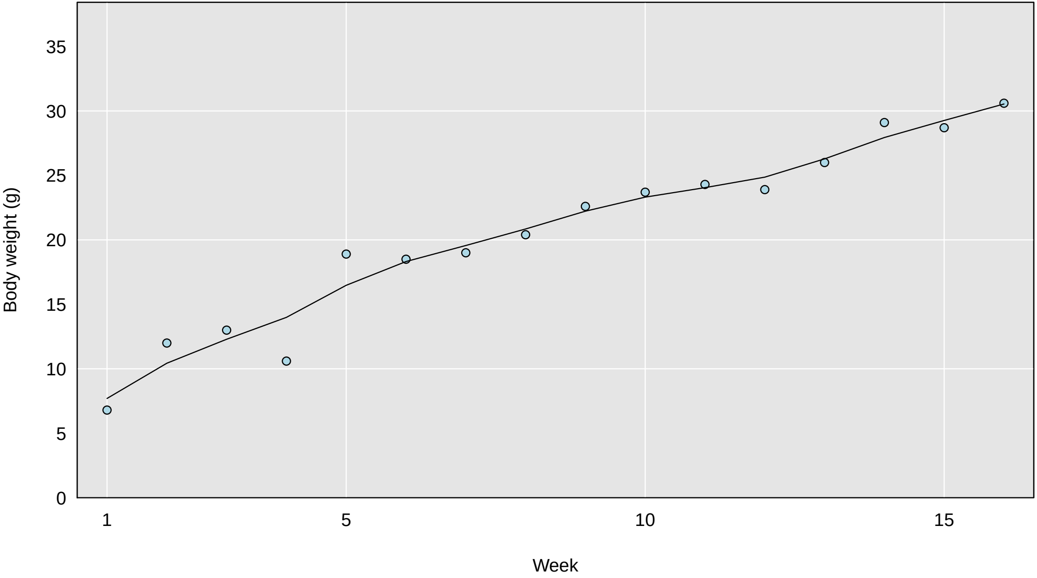
Fitting a model can be useful



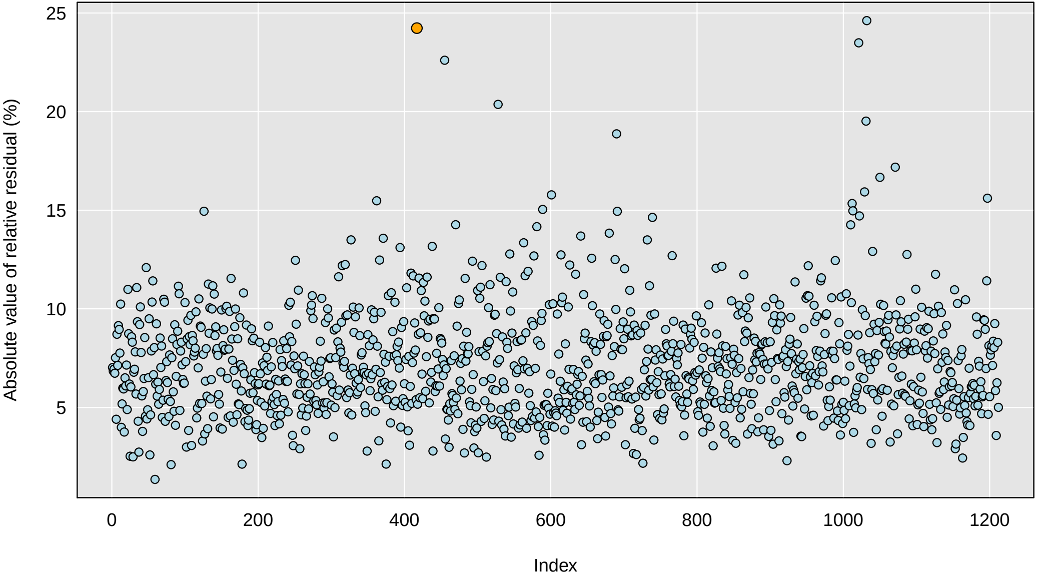
Biggest change vs 2nd difference



Fit a smooth curve



Residuals


Absolute value of relative residual (%)
25
20
15
10
5
0
200
400
600
800
1000
1200
Index

Follow up artifacts

They might be the most interesting results

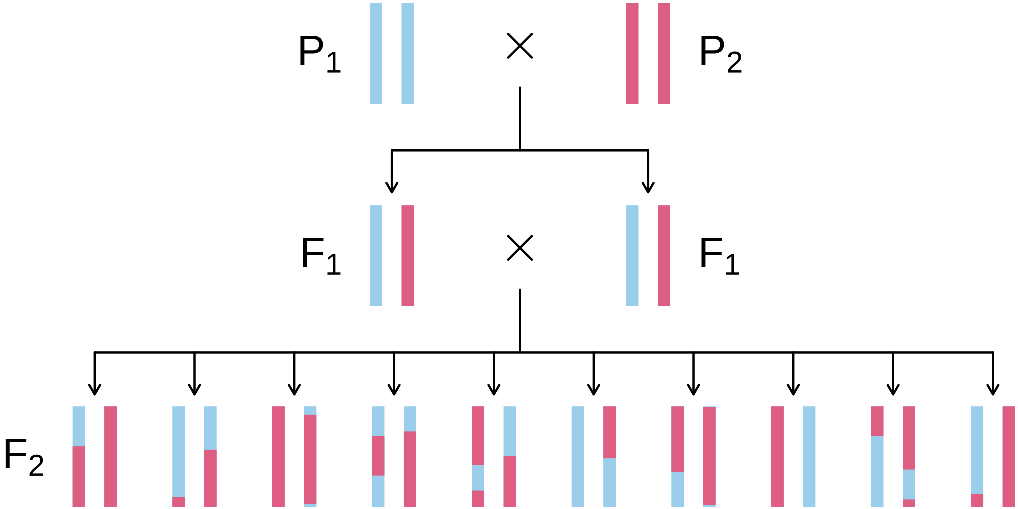
Attie project

$\sim$500 B6 × BTBR intercross mice, all ob/ob

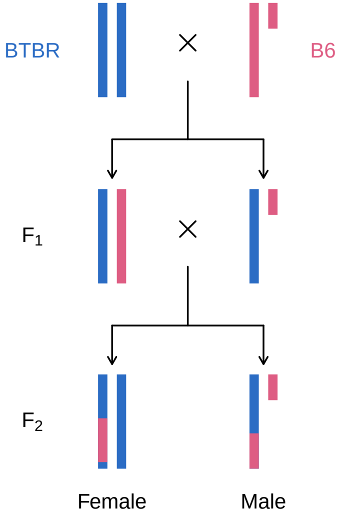
- Genotypes at 2057 SNPs (Affymetrix arrays)
- Gene expression in six tissues (Agilent arrays)
  - adipose
  - gastrocnemius muscle
  - hypothalamus
  - pancreatic islets
  - kidney
  - liver
- Numerous clinical phenotypes

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

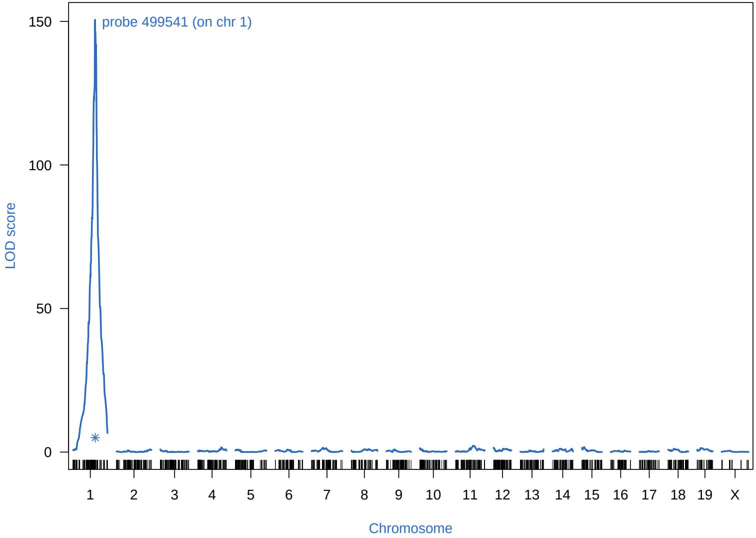
Intercross



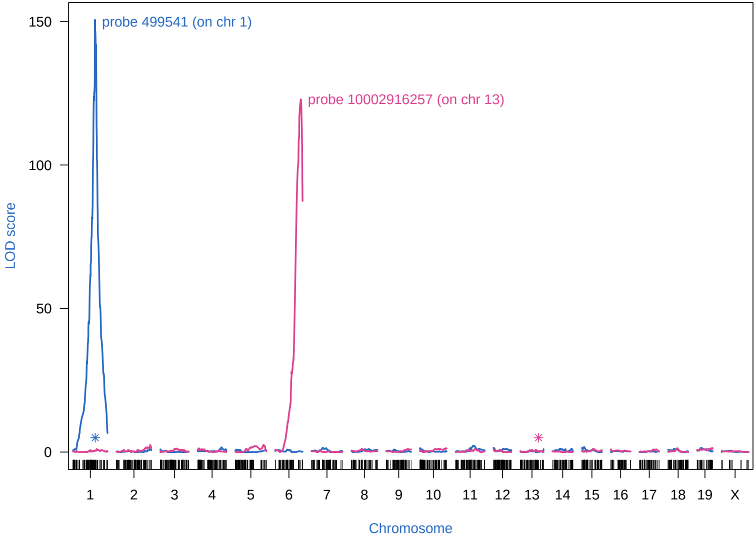
Sex and the X chr



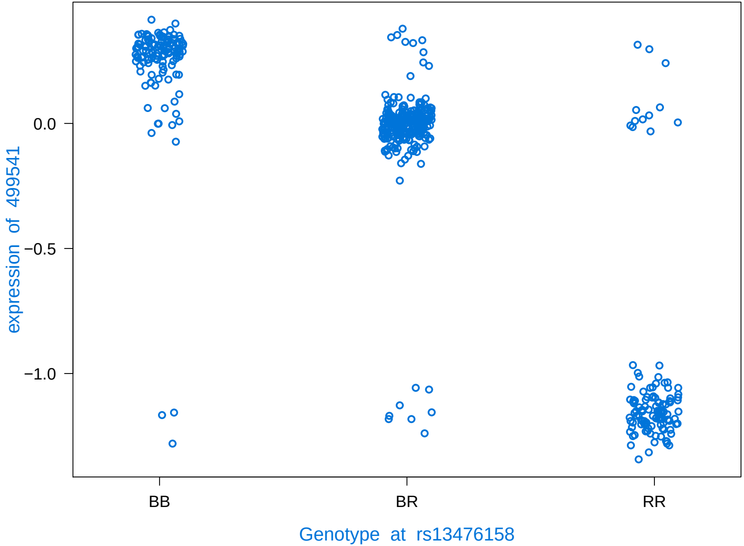
Strong eQTL


probe 499541 (on chr 1)
150
100
50
0
LOD score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

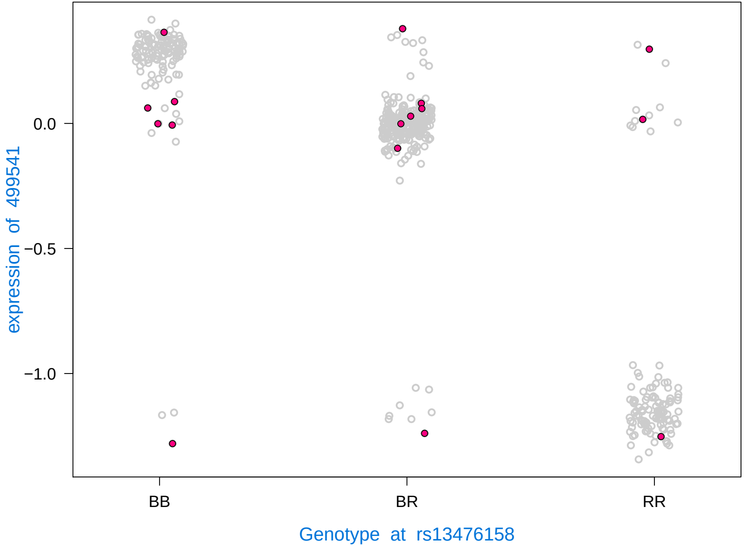
Strong eQTL


probe 499541 (on chr 1)
probe 10002916257 (on chr 13)
LOD score
150
100
50
0
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X

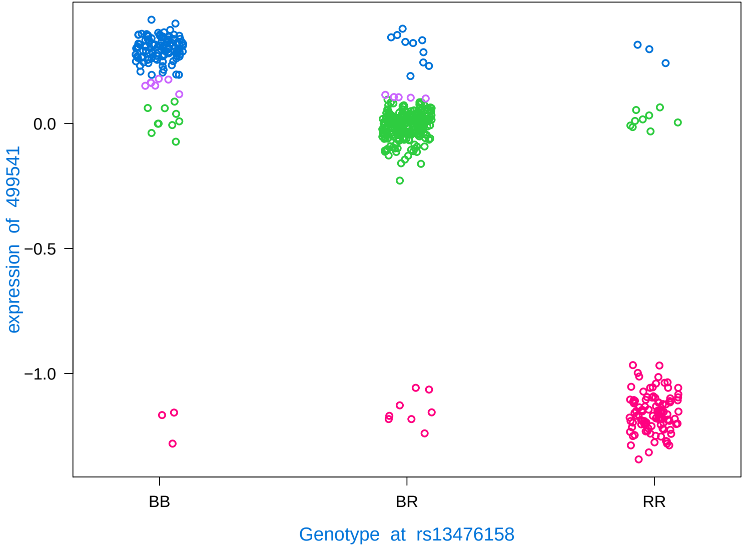
E vs G


0.0
−0.5
−1.0
expression of 499541
BB
BR
RR
Genotype at rs13476158

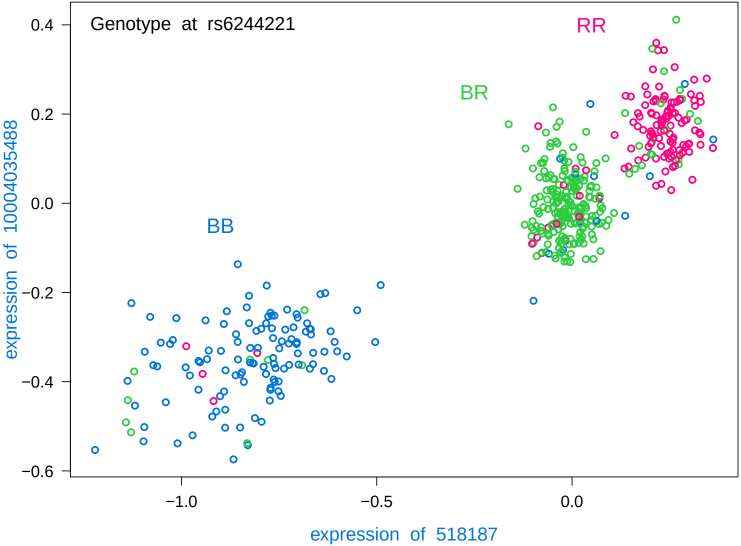
E vs G


0.0
−0.5
−1.0
expression of 499541
BB
BR
RR
Genotype at rs13476158

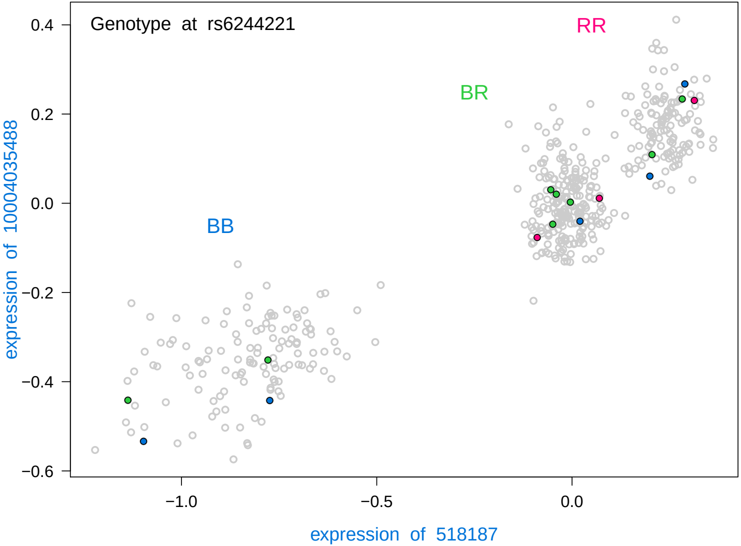
kNN classifier



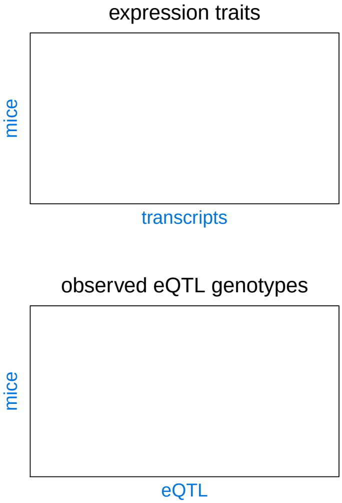
E vs G


Genotype at rs6244221
RR
BR
BB
0.4
0.2
0.0
−0.2
−0.4
−0.6
−1.0
−0.5
0.0
expression of 10004035488
expression of 518187

E vs G


Genotype at rs6244221
RR
BR
BB
0.4
0.2
0.0
−0.2
−0.4
−0.6
−1.0
−0.5
0.0
expression of 100040335488
expression of 518187

Basic scheme



expression traits

mice

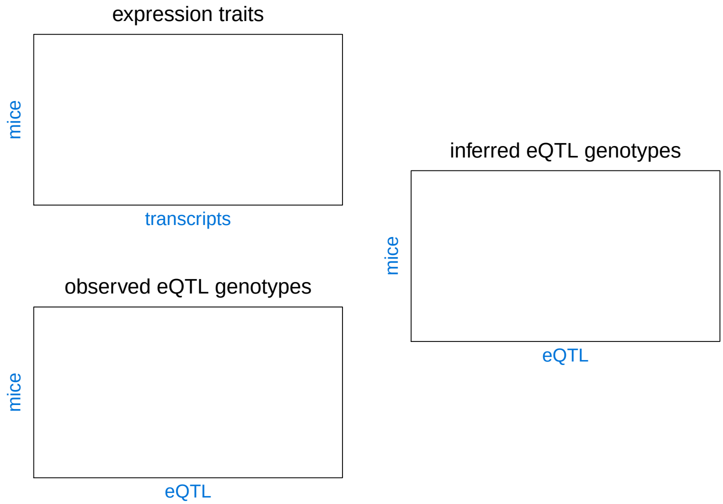
transcripts

observed eQTL genotypes

mice

eQTL

Basic scheme



expression traits

mice

transcripts

observed eQTL genotypes

mice

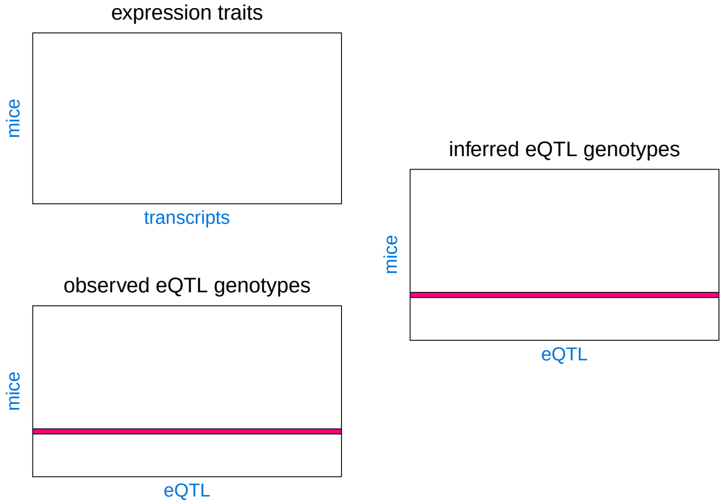
eQTL

inferred eQTL genotypes

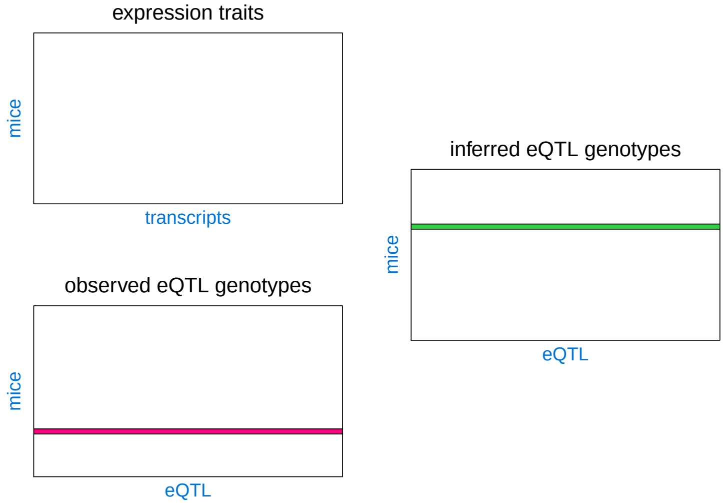
mice

eQTL

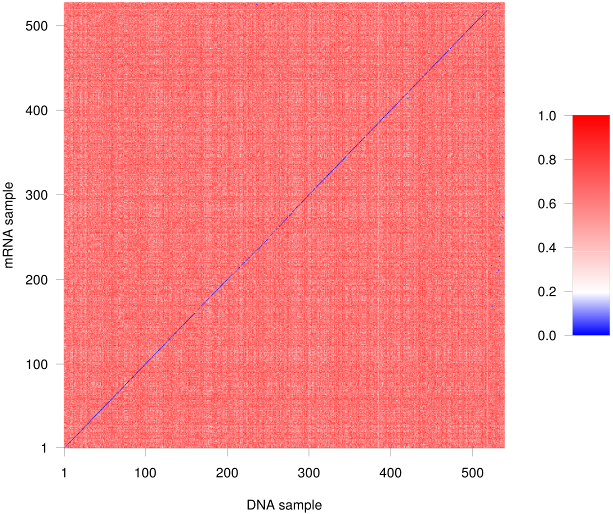
Basic scheme



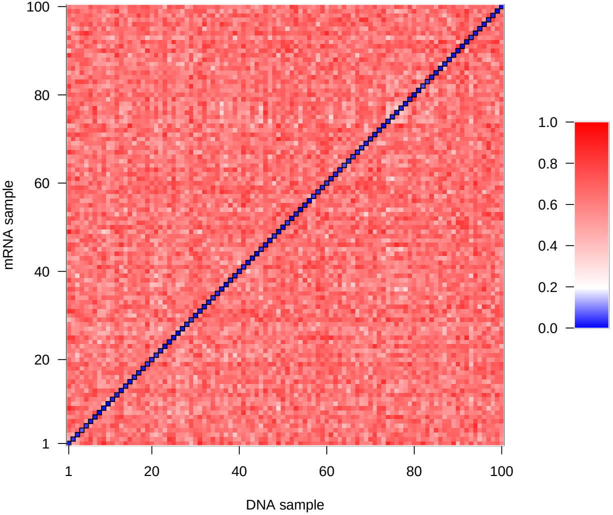
Basic scheme



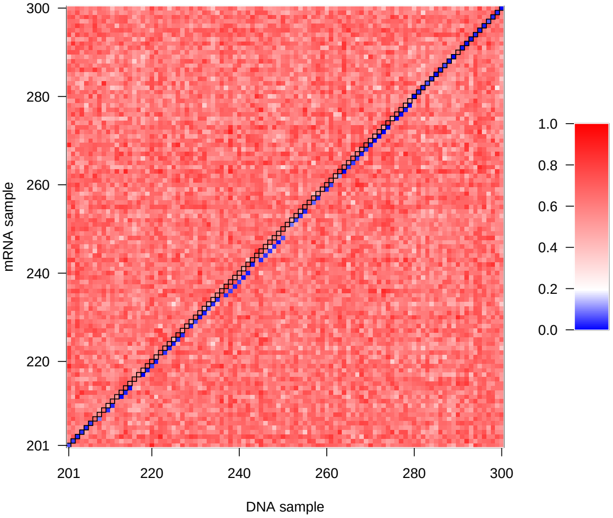
Prop’n mismatches



Prop’n mismatches



Prop’n mismatches



Genotype mix-ups

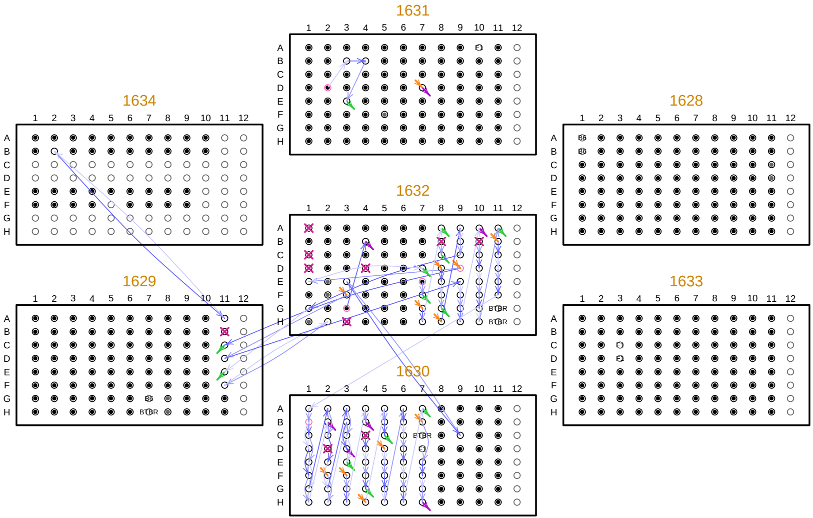
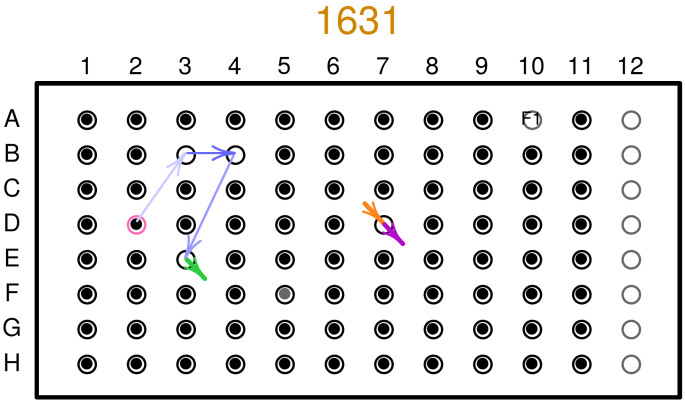
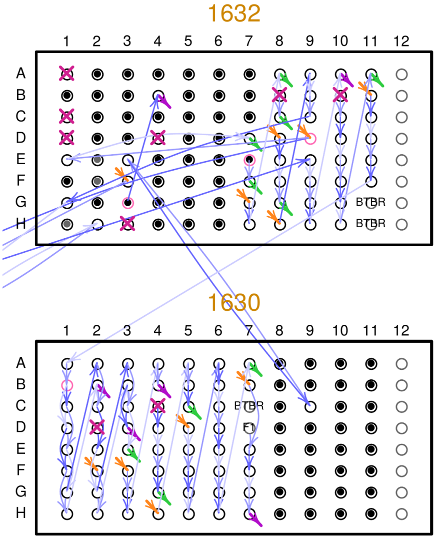


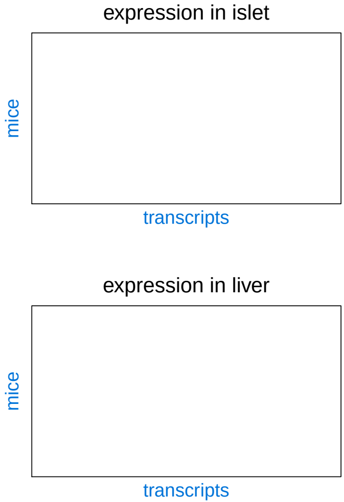
Plate 1631



Plates 1632 and 1630



E vs E



expression in islet

mice

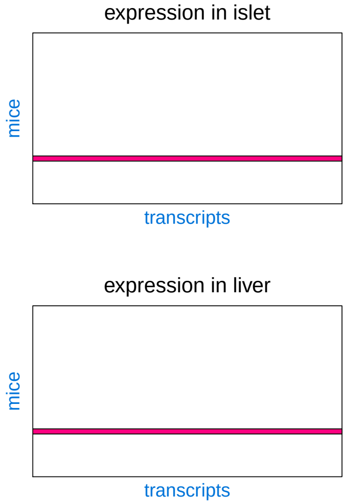
transcripts

expression in liver

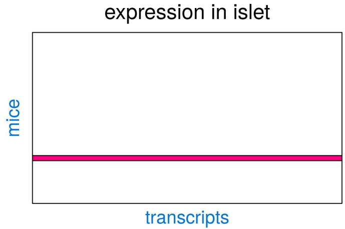
mice

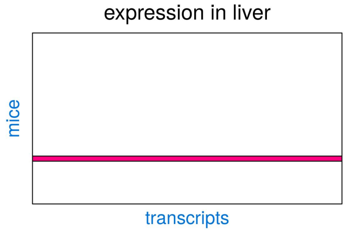
transcripts

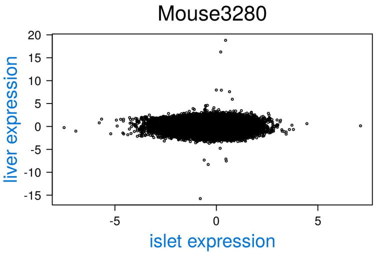
E vs E


expression in islet
mice
transcripts
expression in liver
mice
transcripts

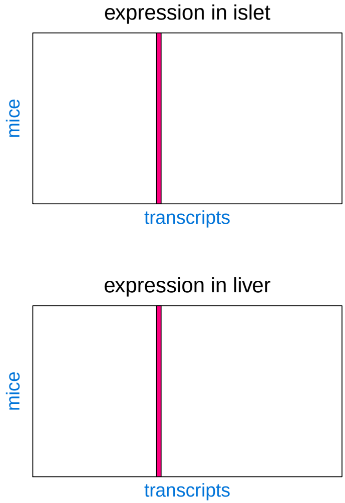
E vs E

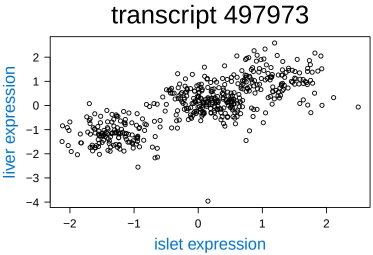

expression in islet
mice
transcripts


expression in liver
mice
transcripts

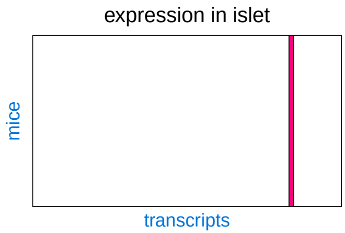

Mouse3280
liver expression
20
15
10
5
0
-5
-10
-15
-5
0
5
islet expression

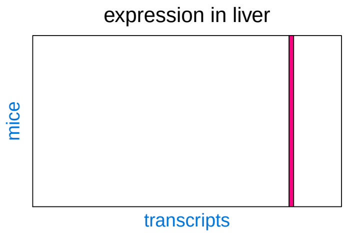
E vs E

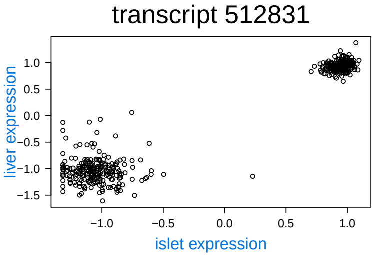

expression in islet
mice
transcripts
expression in liver
mice
transcripts


transcript 497973
liver expression
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
islet expression

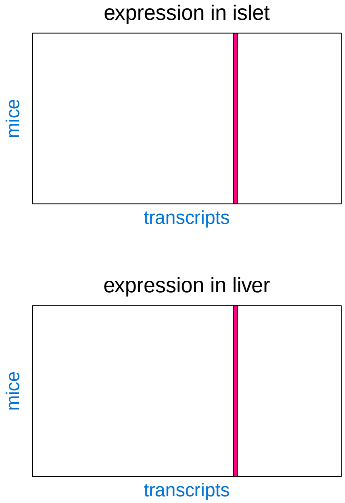
E vs E


expression in islet
mice
transcripts


expression in liver
mice
transcripts

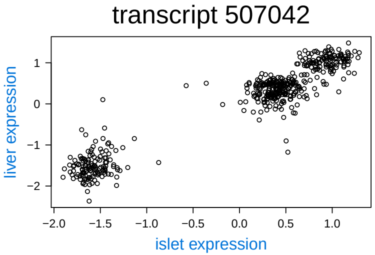

transcript 512831
liver expression
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
islet expression

E vs E



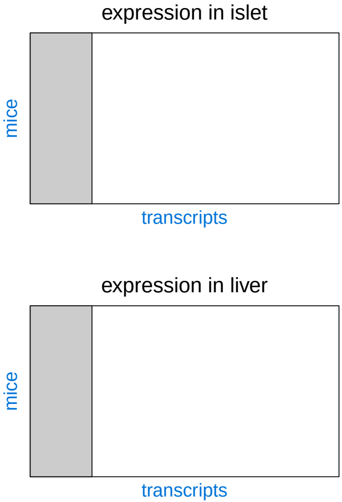
expression in islet

expression in liver



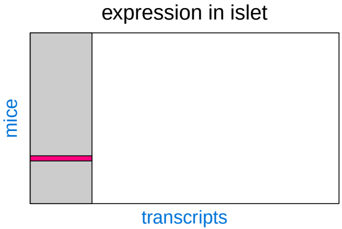
transcript 507042

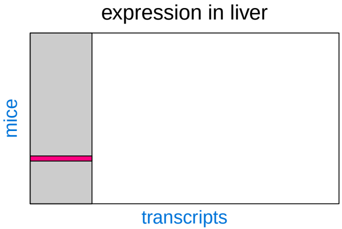
E vs E

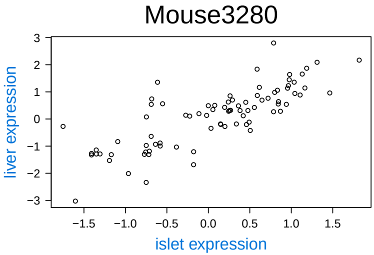

expression in islet

expression in liver

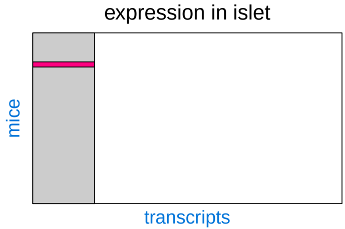
E vs E

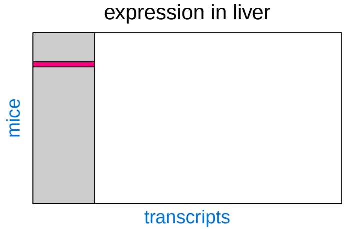

expression in islet
mice
transcripts

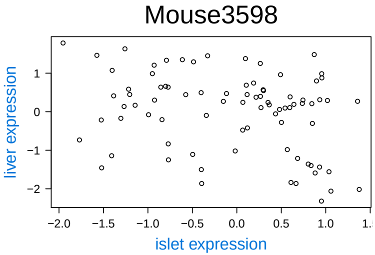

expression in liver
mice
transcripts


Mouse3280
liver expression
3
2
1
0
−1
−2
−3
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
islet expression

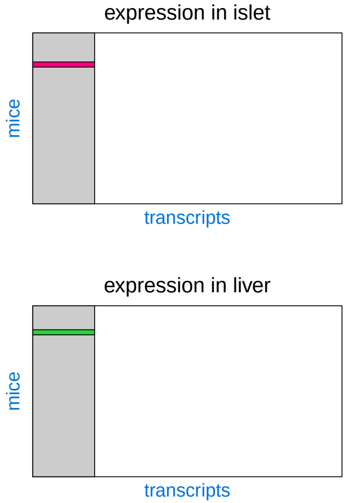
E vs E


expression in islet
mice
transcripts

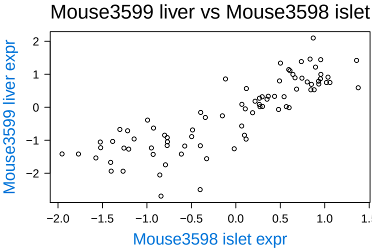

expression in liver
mice
transcripts


Mouse3598
liver expression
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.
islet expression

E vs E

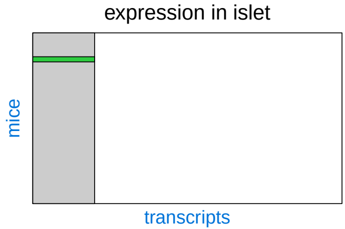


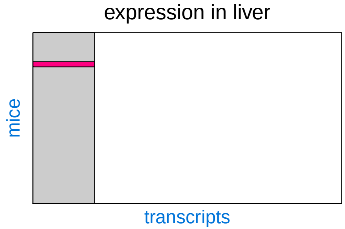
expression in islet
mice
transcripts
expression in liver
mice
transcripts

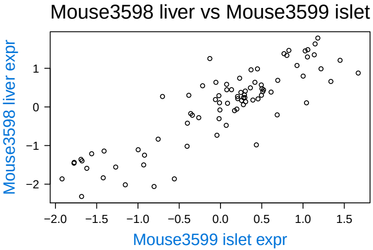


Mouse3599 liver vs Mouse3598 islet
Mouse3599 liver expr
2
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.
Mouse3598 islet expr

E vs E

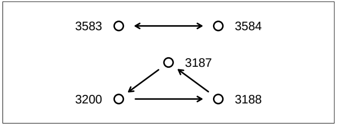

expression in islet
mice
transcripts


expression in liver
mice
transcripts


Mouse3598 liver vs Mouse3599 islet
Mouse3598 liver expr
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Mouse3599 islet expr

Expression mix-ups

adipose



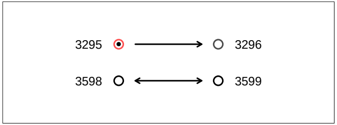
3583 ⟷ 3584

3187

3200

3188

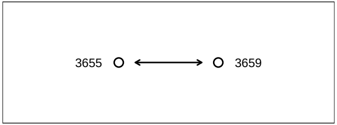
islet



3295 → 3296

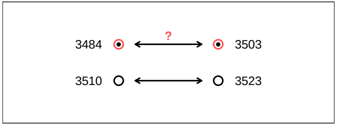
3598 ⟷ 3599

gastroc



3655 ⟷ 3659

kidney

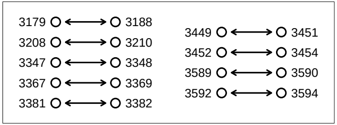


?

3484 ⟷ 3503

3510 ⟷ 3523

hypo



3179 ⟷ 3188

3208 ⟷ 3210

3347 ⟷ 3348

3367 ⟷ 3369

3381 ⟷ 3382

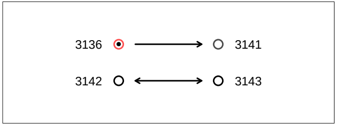
3449 ⟷ 3451

3452 ⟷ 3454

3589 ⟷ 3590

3592 ⟷ 3594

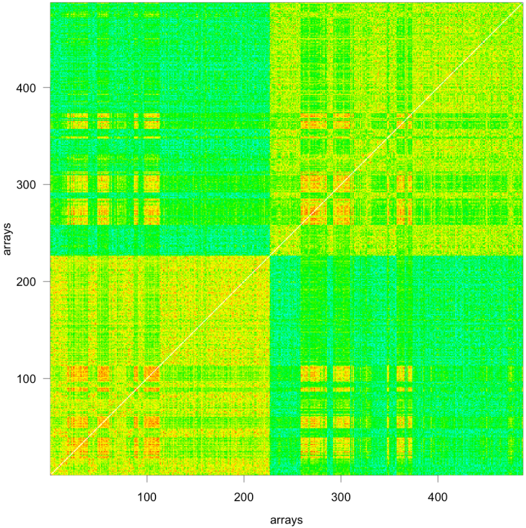
liver



3136 → 3141

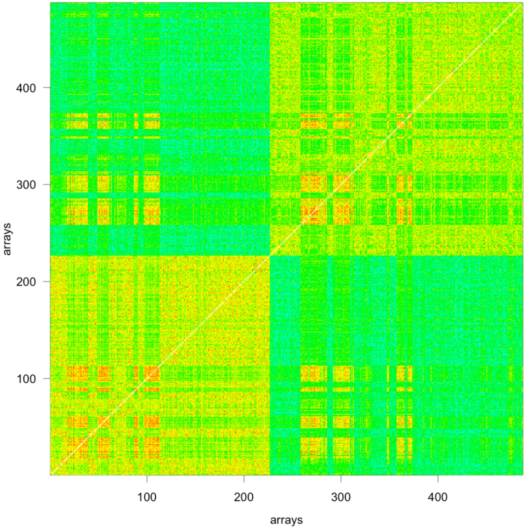
3142 ⟷ 3143

Another example



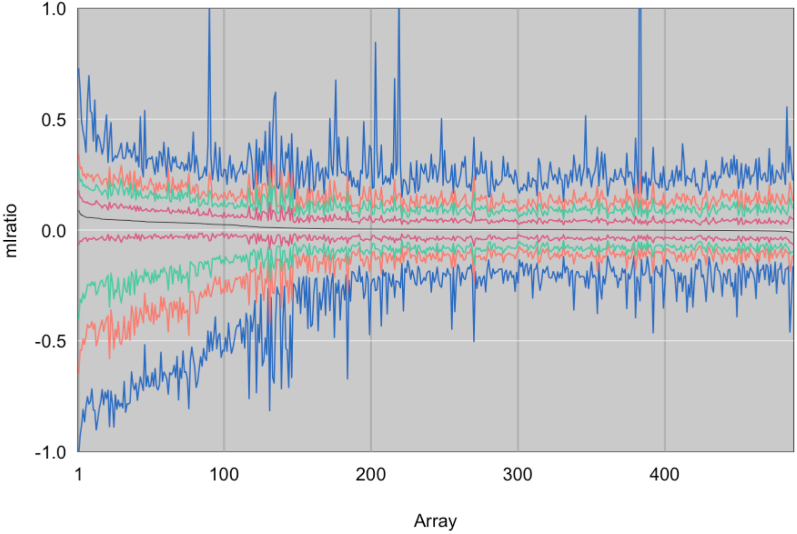
kbroman.org/blog/2012/04/25/microarrays-suck

What the heck?



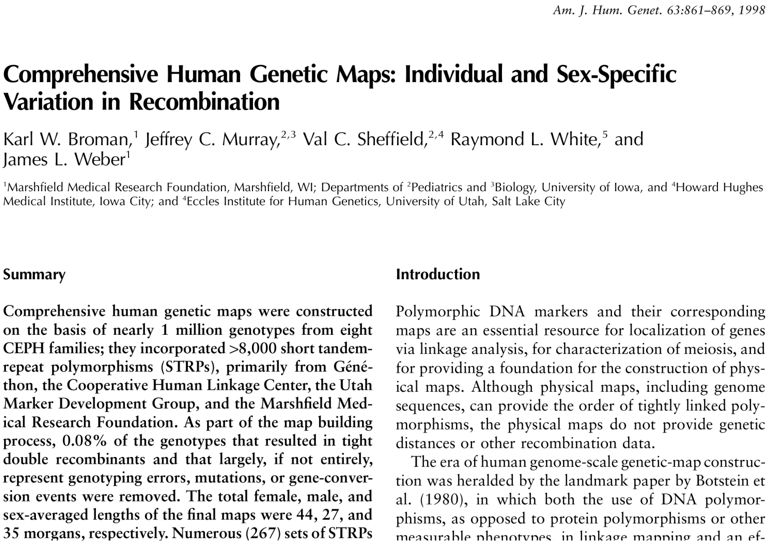
kbroman.org/blog/2012/04/25/microarrays-suck

Dense box plots



Follow up artifacts

They might be the most interesting results



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

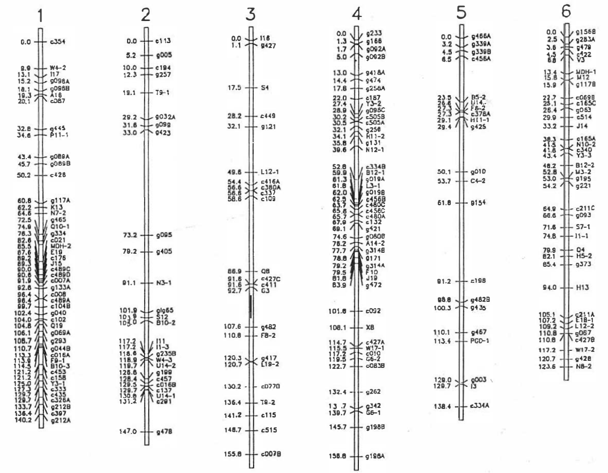
Comprehensive human genetic maps were constructed on the basis of nearly 1 million genotypes from eight CEPH families; they incorporated >8,000 short tandem-repeat polymorphisms (STRPs), primarily from Généthon, the Cooperative Human Linkage Center, the Utah Marker Development Group, and the Marshfield 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.

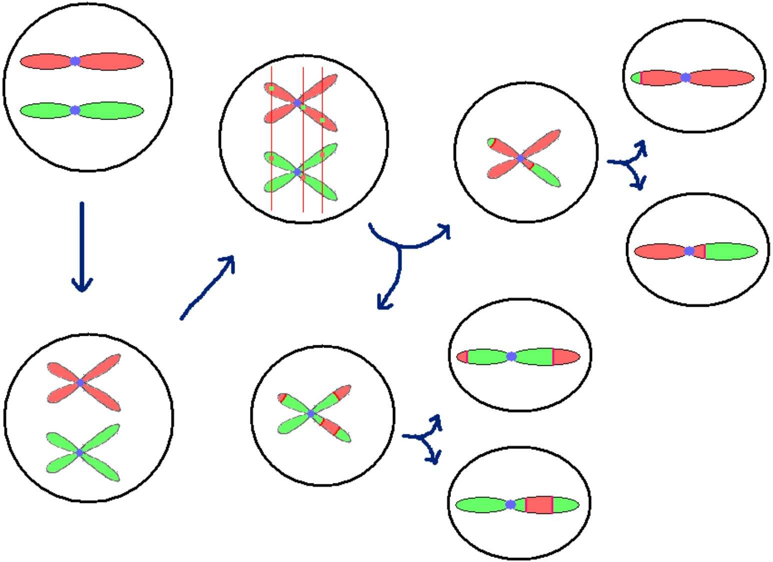
The era of human genome-scale genetic-map construction was heralded by the landmark paper by Botstein et al. (1980), in which both the use of DNA polymorphisms, as opposed to protein polymorphisms or other measurable phenotypes, in linkage mapping and an ef-

Eucalypt genetic map



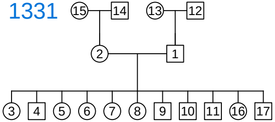
Byrne et al., Theor Appl Genet 91:869–875, 1995

Meiosis

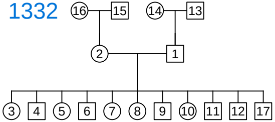


CEPH pedigrees

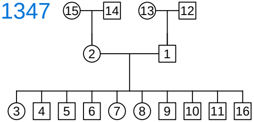
1331



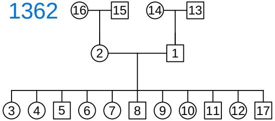
1332



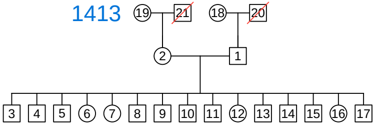
1347



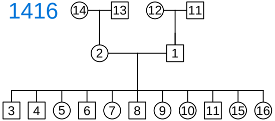
1362



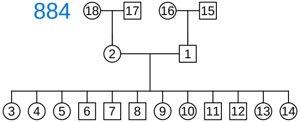
1413



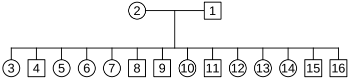
1416



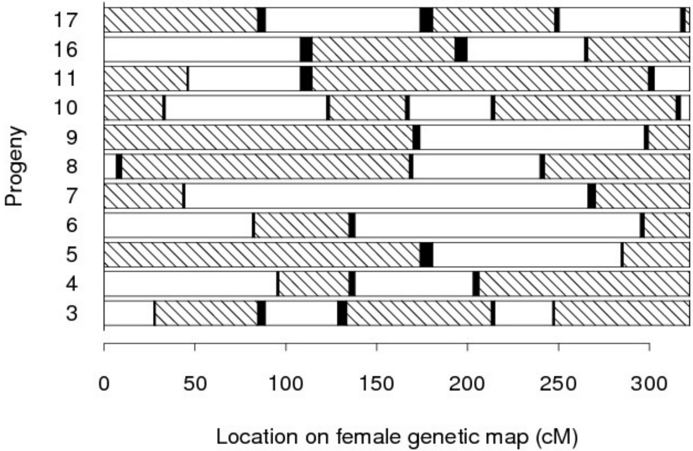
884



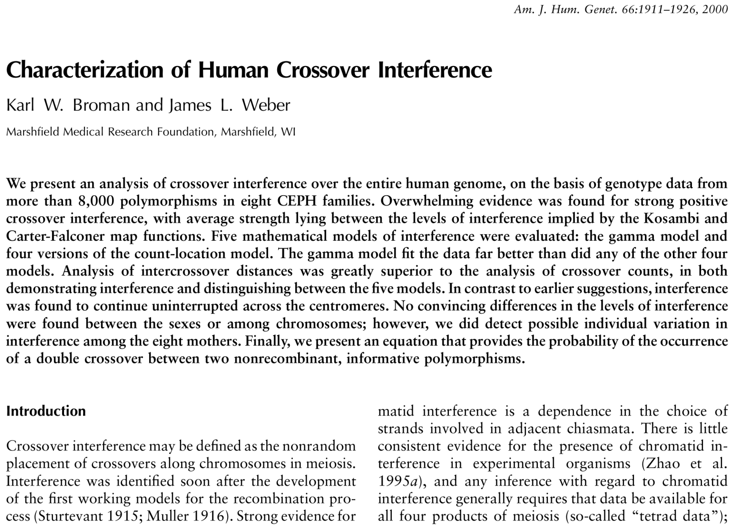
102



Crossover locations



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



Am. J. Hum. Genet. 66:1911–1926, 2000

Characterization of Human Crossover Interference

Karl W. Broman and James L. Weber

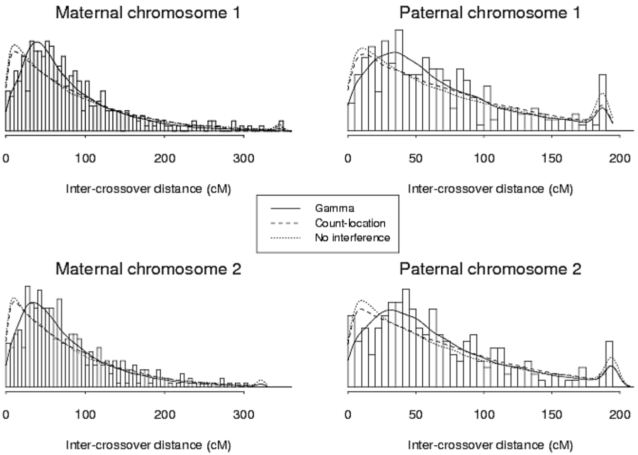
Marshfield Medical Research Foundation, Marshfield, WI

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

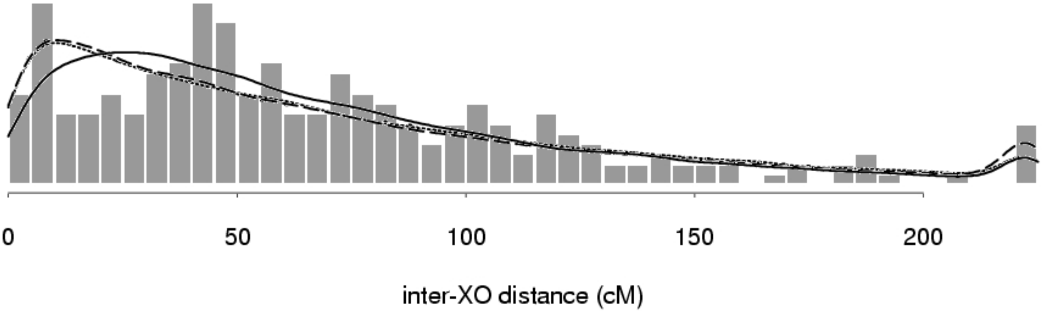
Crossover interference may be defined as the nonrandom placement of crossovers along chromosomes in meiosis. Interference was identified soon after the development of the first working models for the recombination process (Sturtevant 1915; Muller 1916). Strong evidence for matid interference is a dependence 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");

Crossover interference

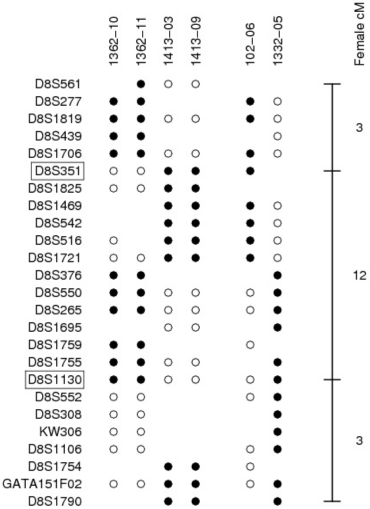


Broman and Weber, Am J Hum Genet 66:1911–1926, 2000

Maternal chr 8



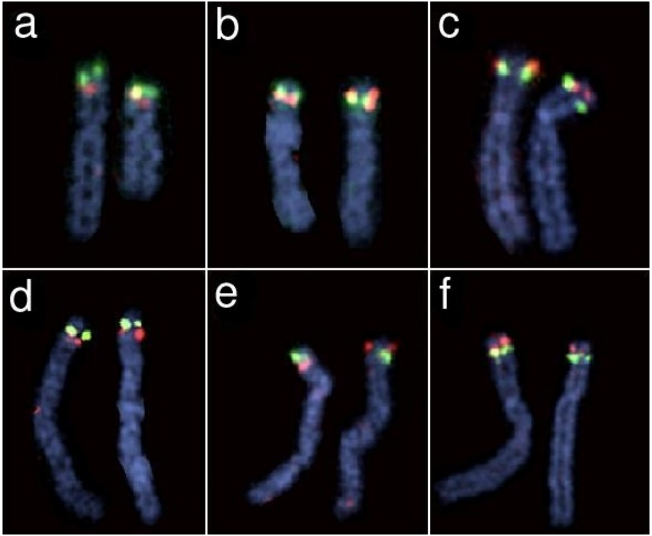
Apparent triple XOs



	1362–10	1362–11	1413–03	1413–09	102–06	1332–05
D8S561		●	○	○		
D8S277	●	●			●	○
D8S1819	●	●	○	○	●	○
D8S439	●	●				○
D8S1706	●	●	○	○	●	○
D8S351	○	○	●	●	●	
D8S1825	○	○	●	●		
D8S1469			●	●	●	○
D8S542			●	●	●	○
D8S516	○		●	●	●	○
D8S1721	○	○	●	●	●	○
D8S376	●	●				●
D8S550	●	●	○	○	○	●
D8S265	●	●	○	○	○	●
D8S1695			○	○		●
D8S1759	●	●			○	
D8S1755	●	●	○	○		●
D8S1130	●	●	○	○	○	●
D8S552	○	○			○	●
D8S308	○	○				●
KW306	○	○				●
D8S1106	○	○			○	●
D8S1754			●	●	○	
GATA151F02	○	○	●	●	○	●
D8S1790			●	●		●

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

Chr 8p inversion



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

Capturing EDA

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

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?”

Basic principles

Step 1: slow down and document.

Step 2: have sympathy for your future self.

Step 3: have a system.

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`

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

– Tukey