

Salvaging a genetics project

Identifying and correcting sample mix-ups in high-dimensional data

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org

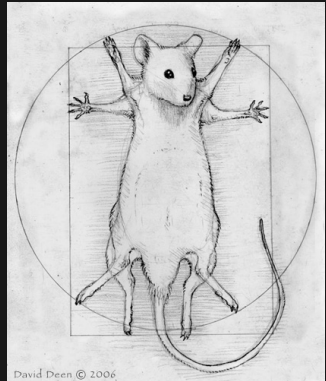
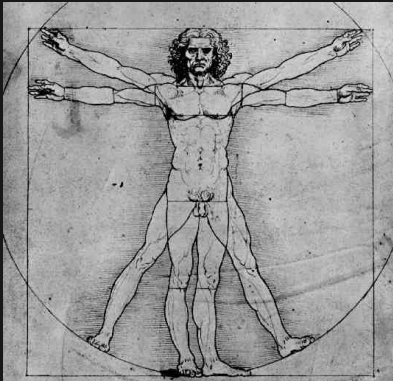
github.com/kbroman

@kwbroman

bit.ly/berkeley2017

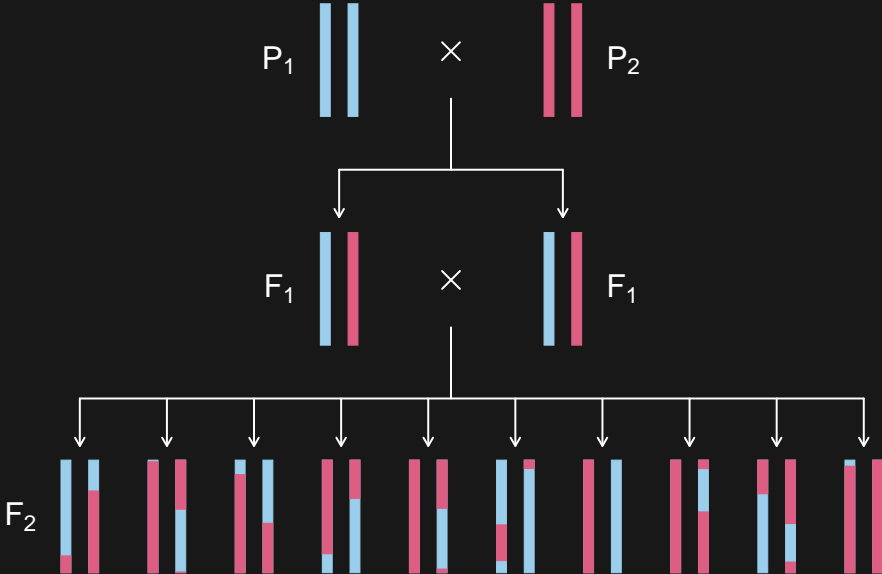




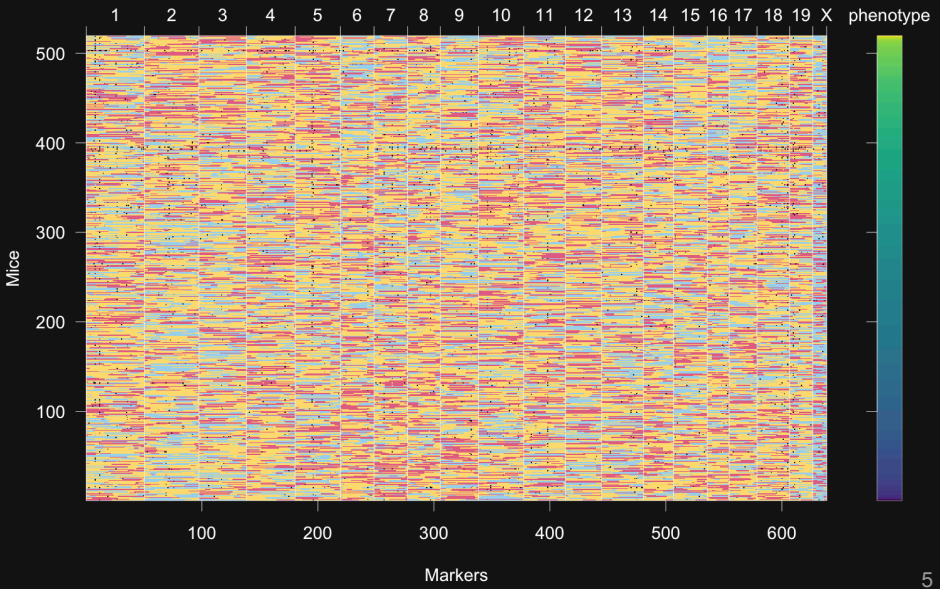


daviddeen.com

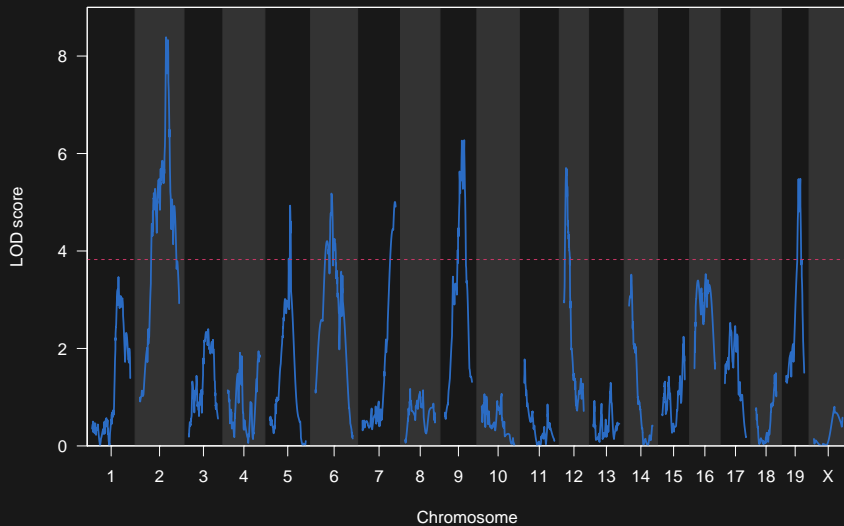
Intercross



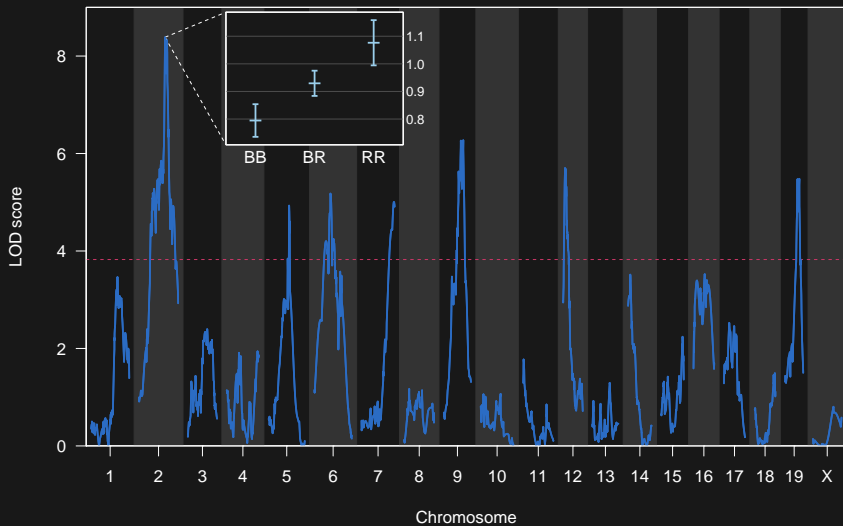
Data



QTL mapping



QTL mapping

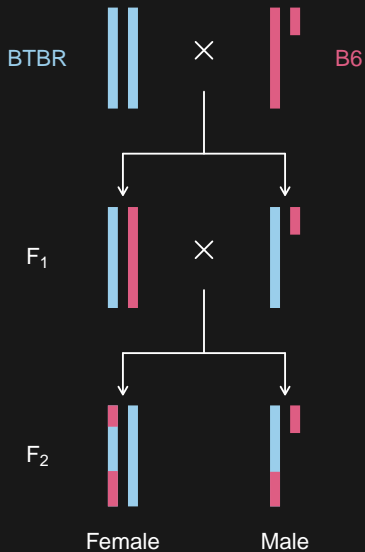


Attie project

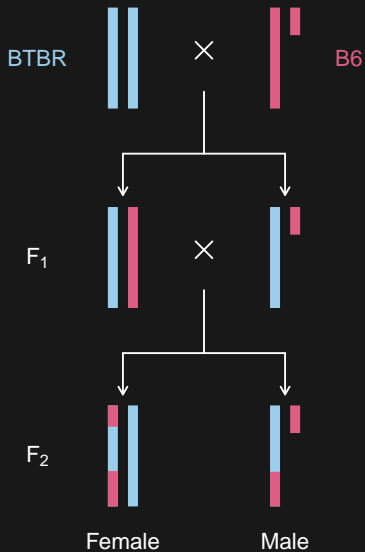
~500 B6 × BTBR intercross mice, all ob/ob

- ▶ 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)

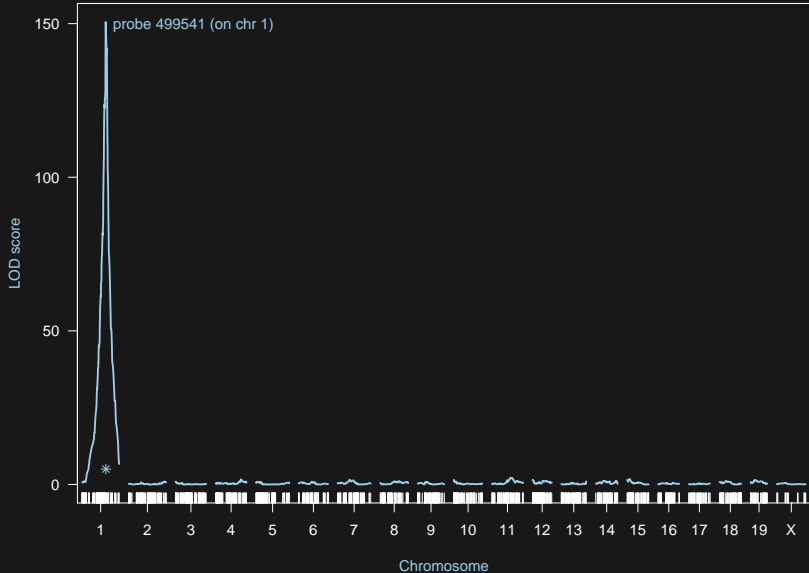
Sex and the X chr



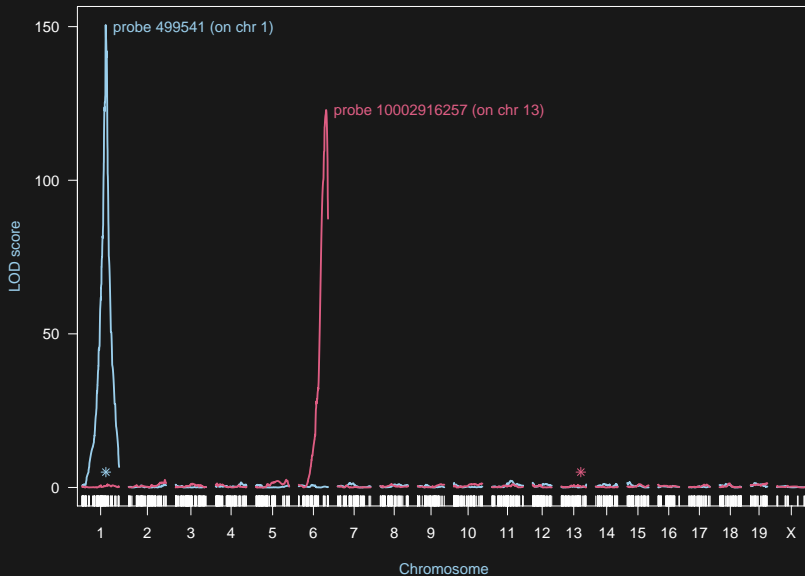
Sex and the X chr



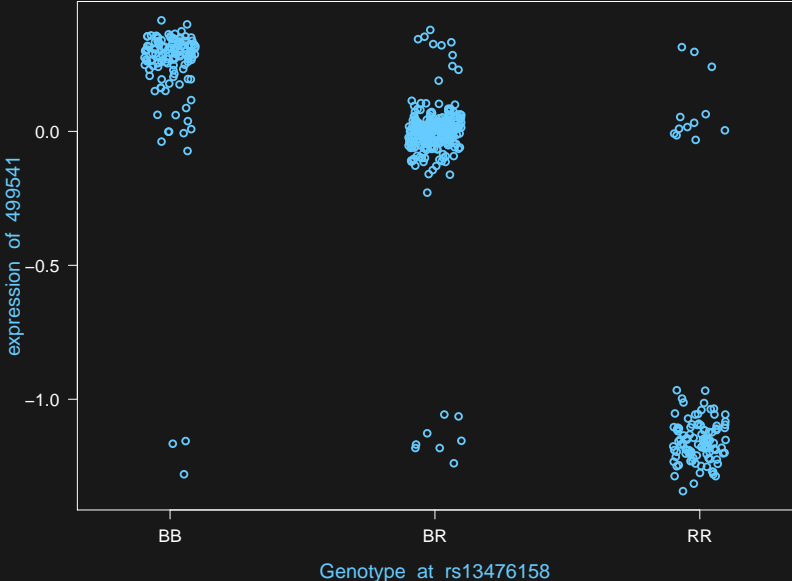
Strong eQTL



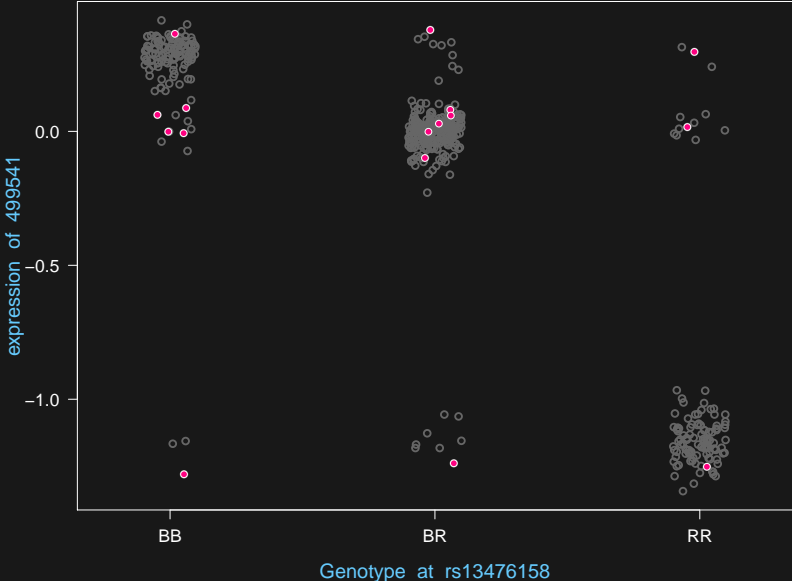
Strong eQTL



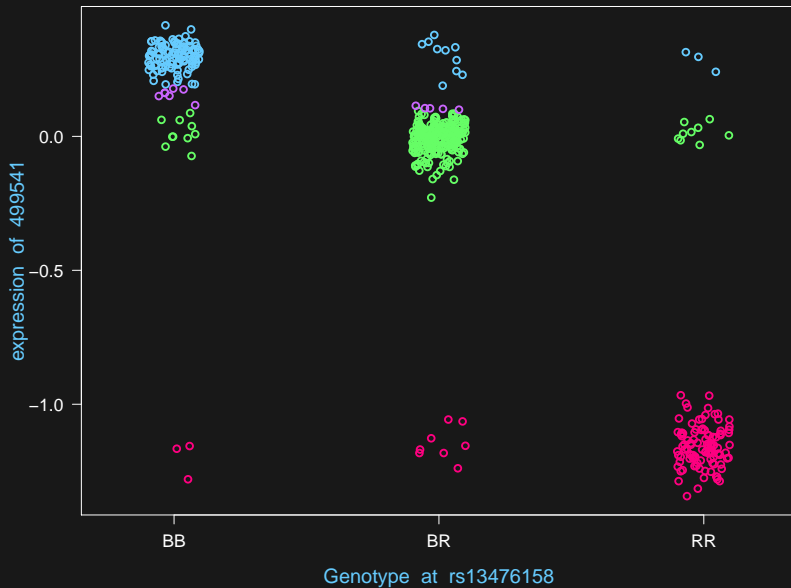
E vs G



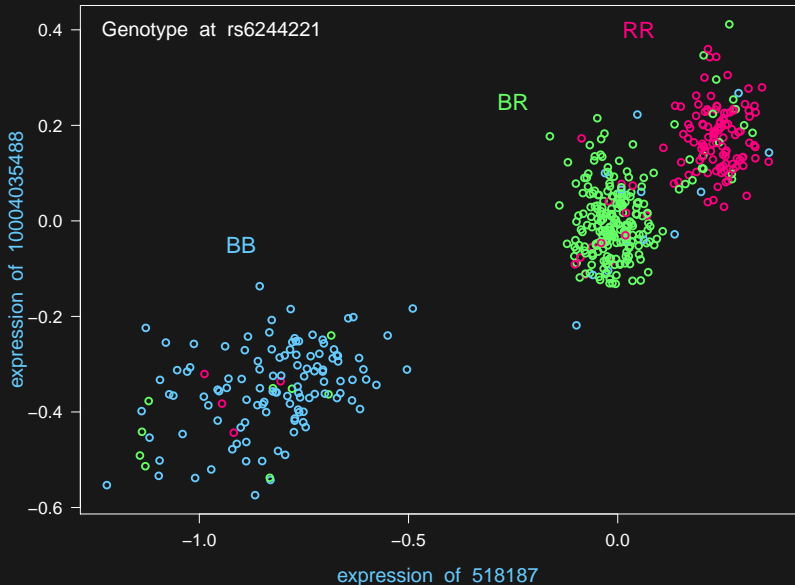
E vs G



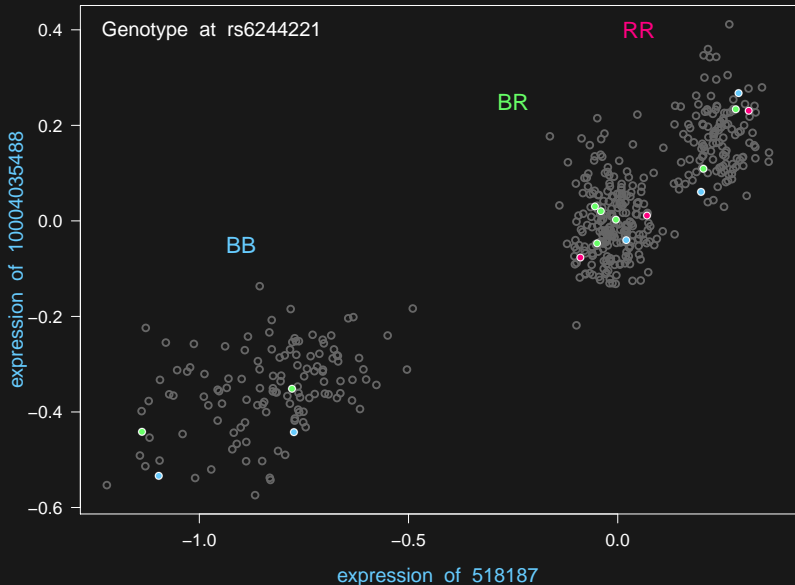
kNN classifier



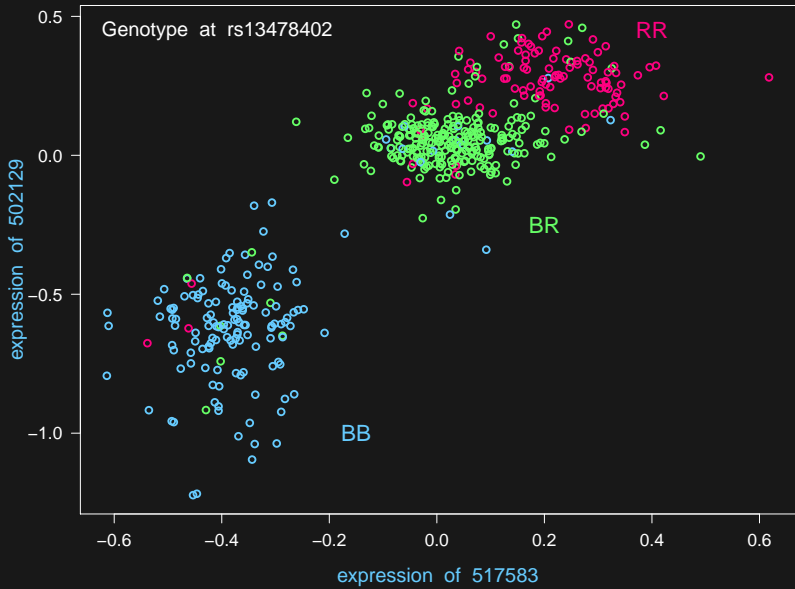
E vs G



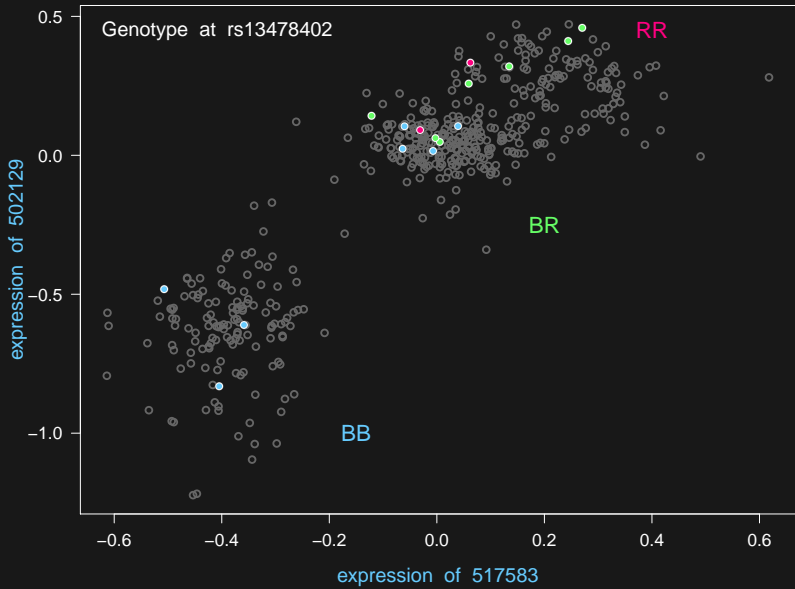
E vs G



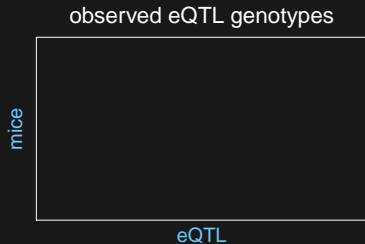
E vs G



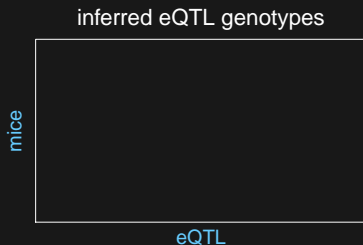
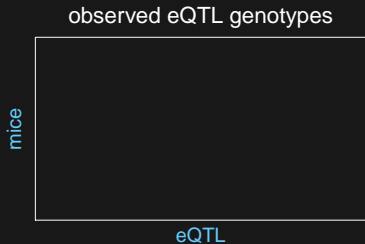
E vs G



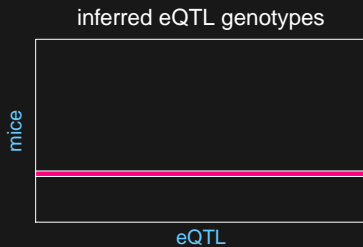
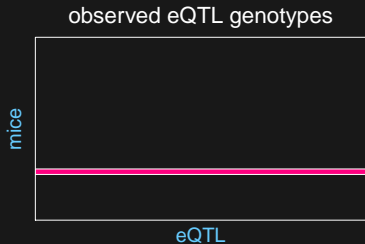
Basic scheme



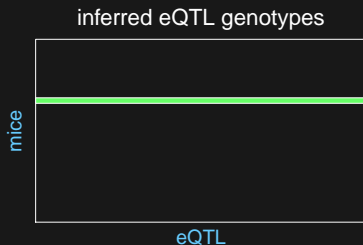
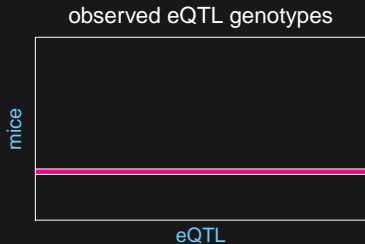
Basic scheme



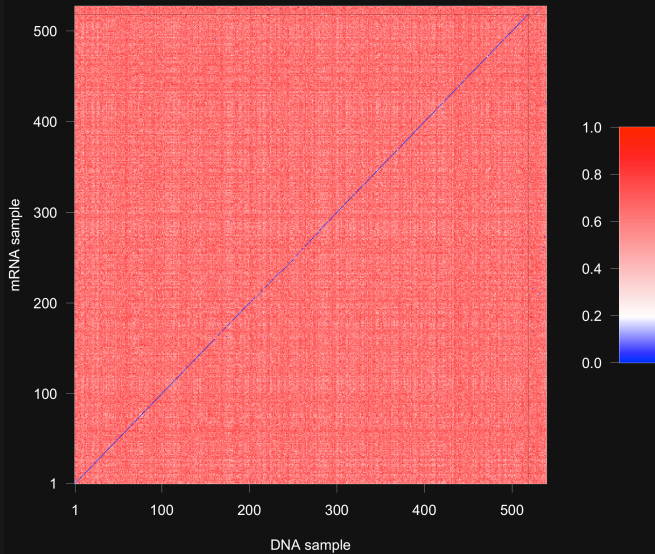
Basic scheme



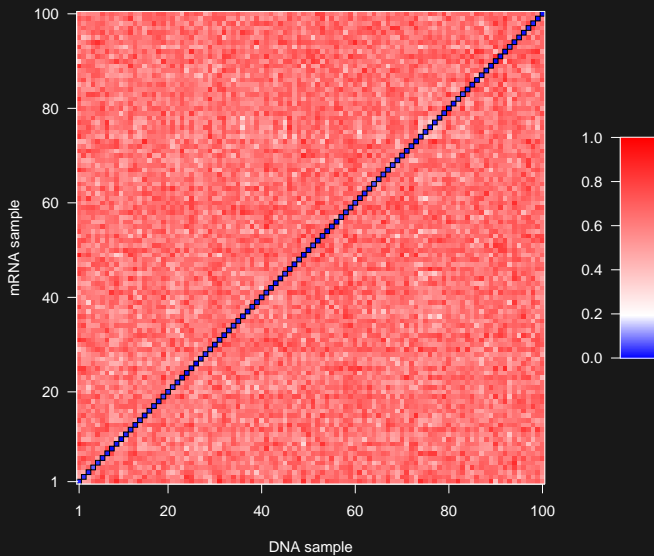
Basic scheme



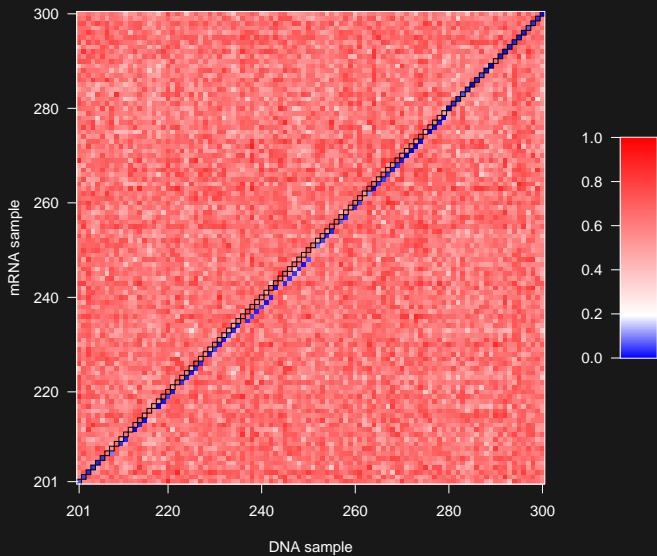
Prop'n mismatches



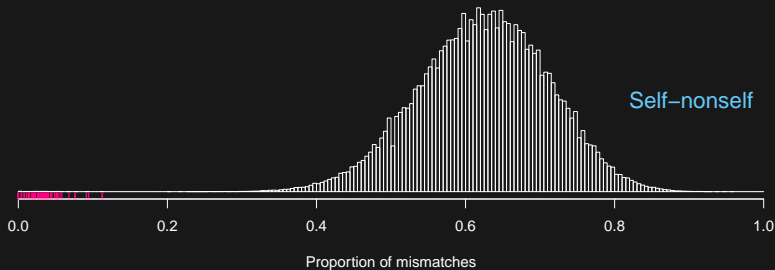
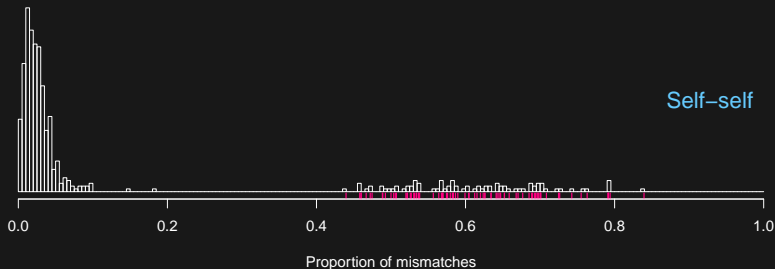
Prop'n mismatches



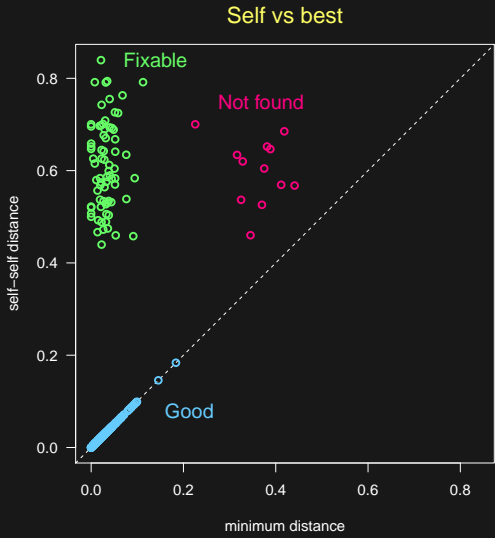
Prop'n mismatches



Prop'n mismatches



Decisions



Genotype mix-ups

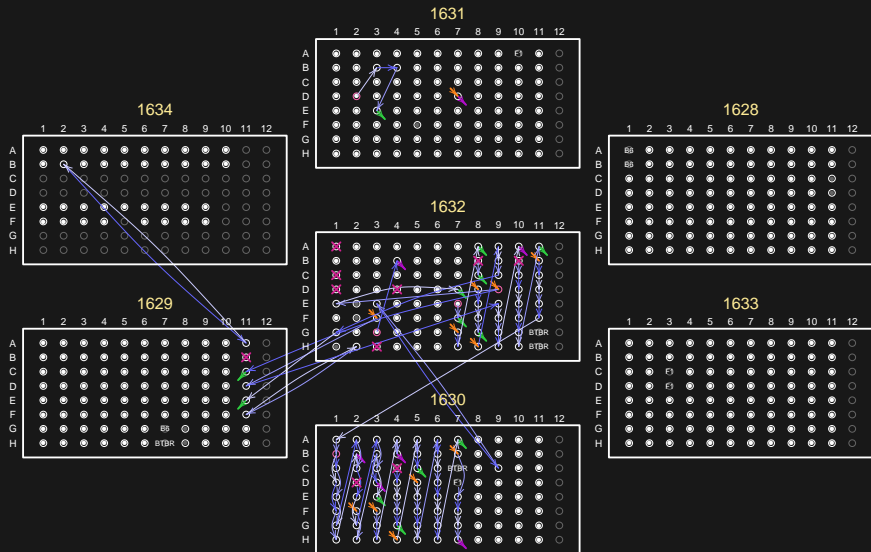
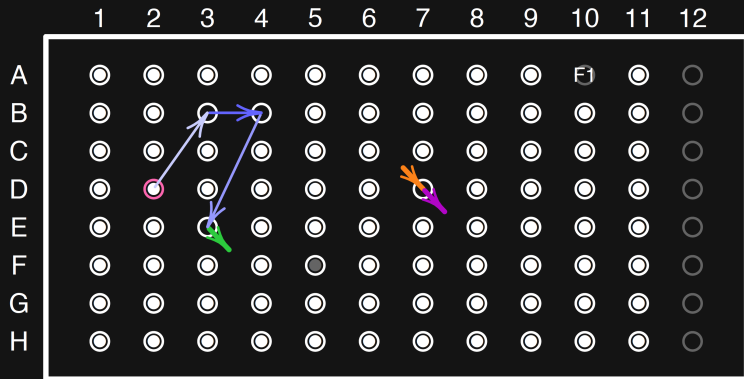


Plate 1631

1631



Plates 1632 and 1630

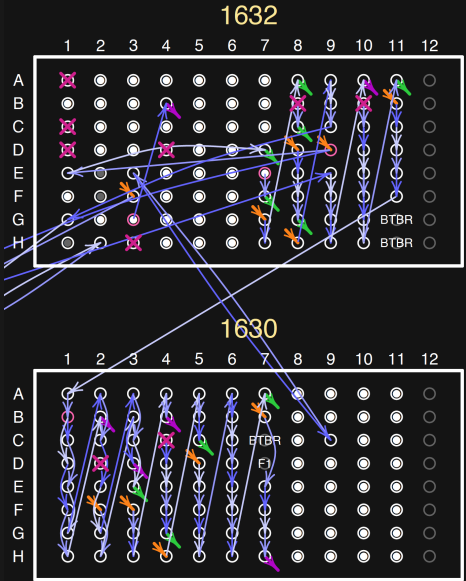
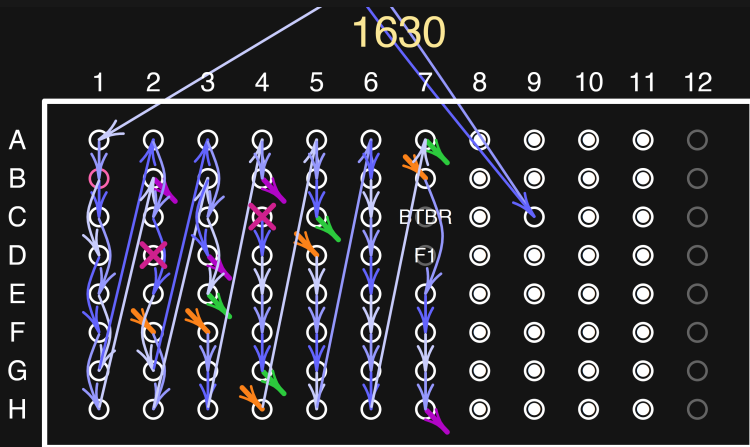


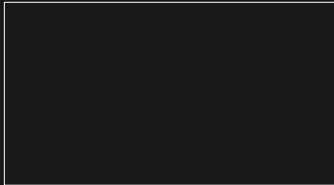
Plate 1630



E vs E

expression in islet

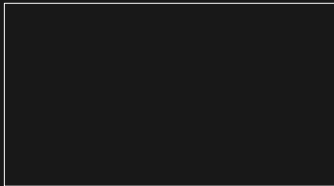
mice



transcripts

expression in liver

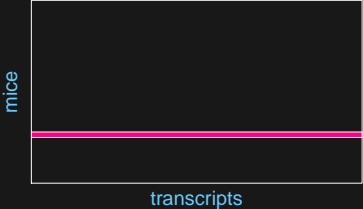
mice



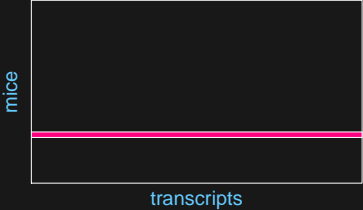
transcripts

E vs E

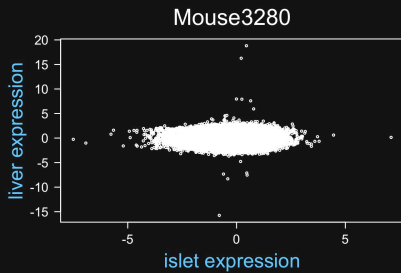
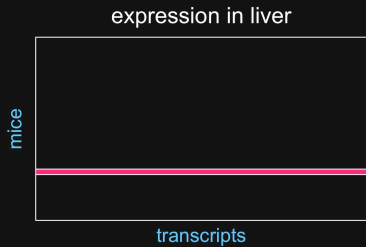
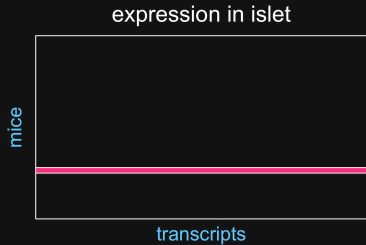
expression in islet



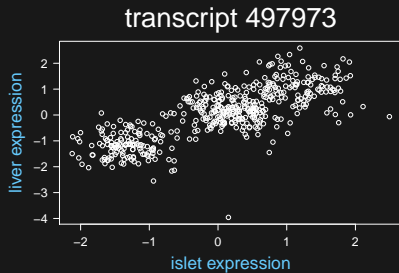
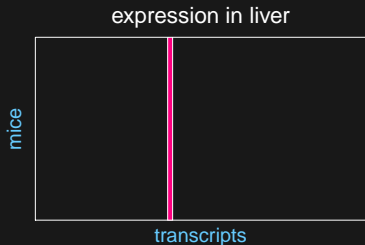
expression in liver



E vs E



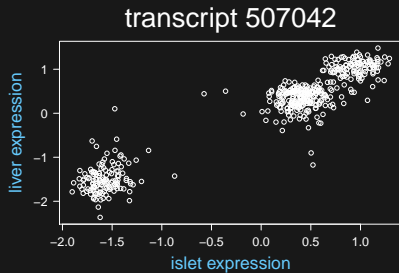
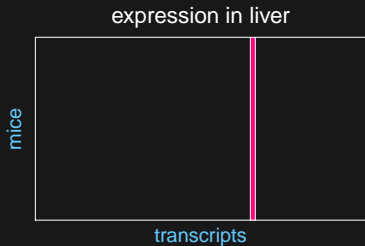
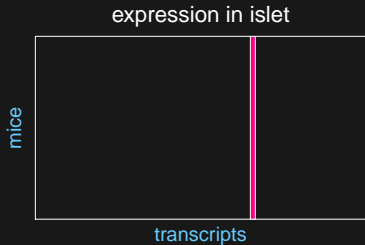
E vs E



E vs E



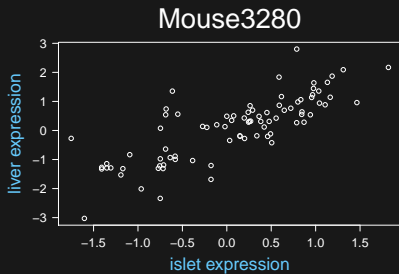
E vs E



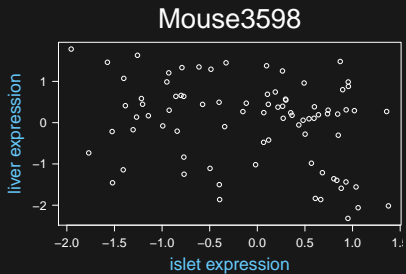
E vs E



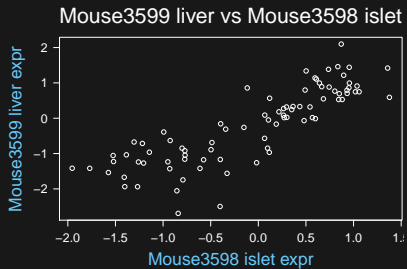
E vs E



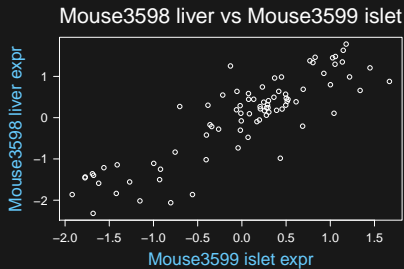
E vs E



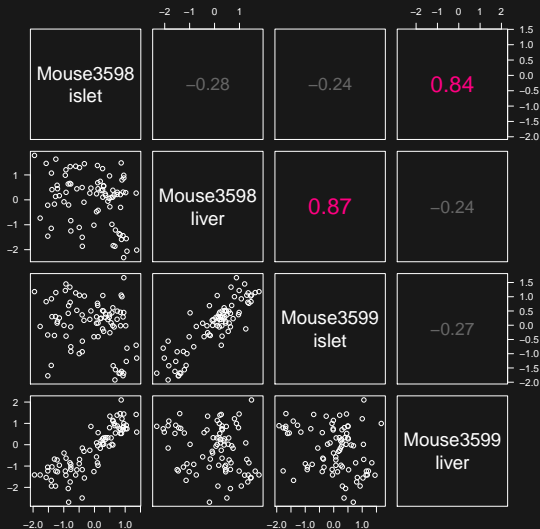
E vs E



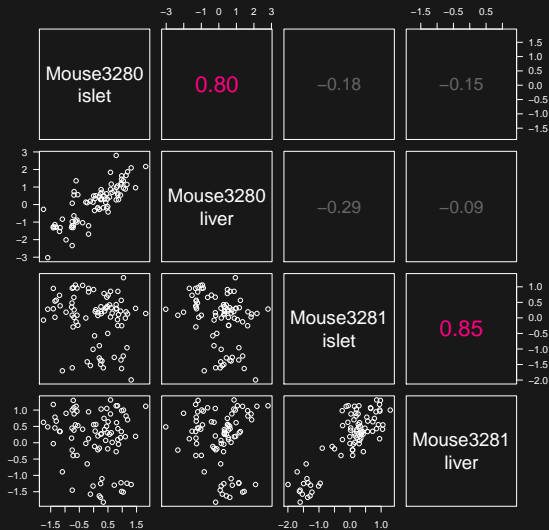
E vs E



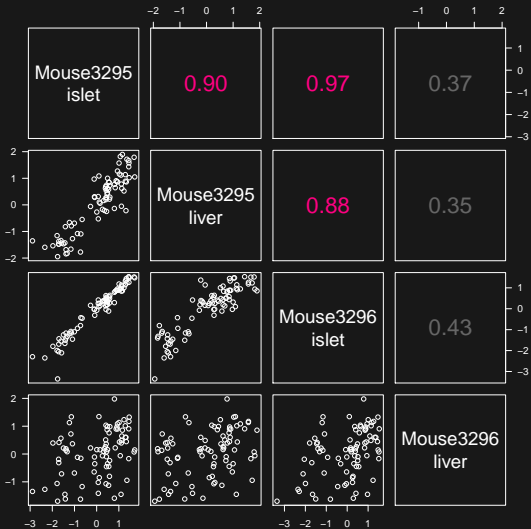
E vs E



E vs E

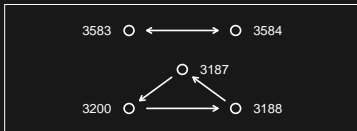


E vs E

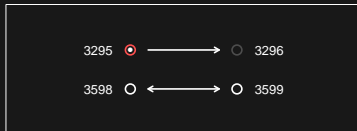


Expression mix-ups

adipose



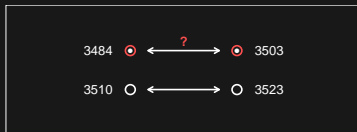
islet



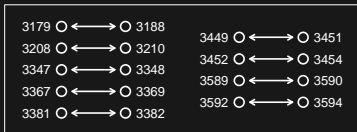
gastroc



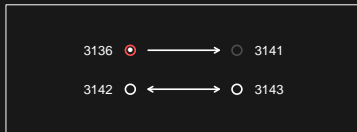
kidney



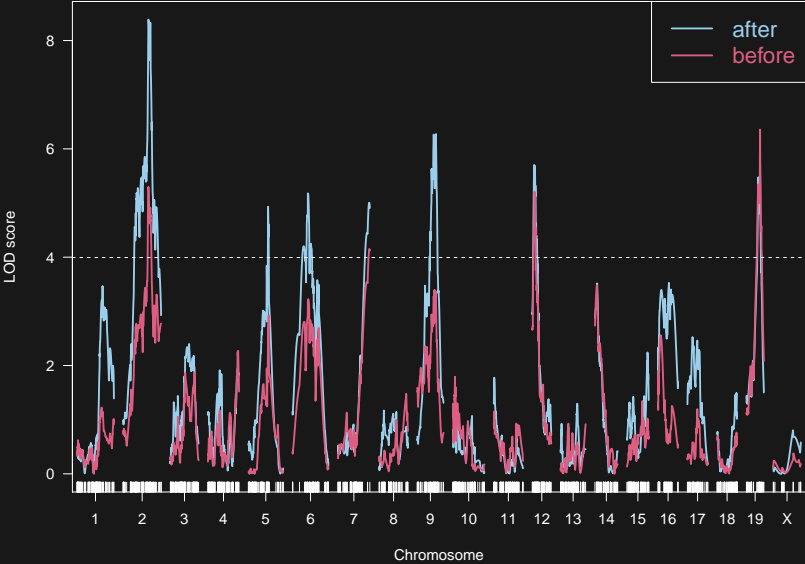
hypo



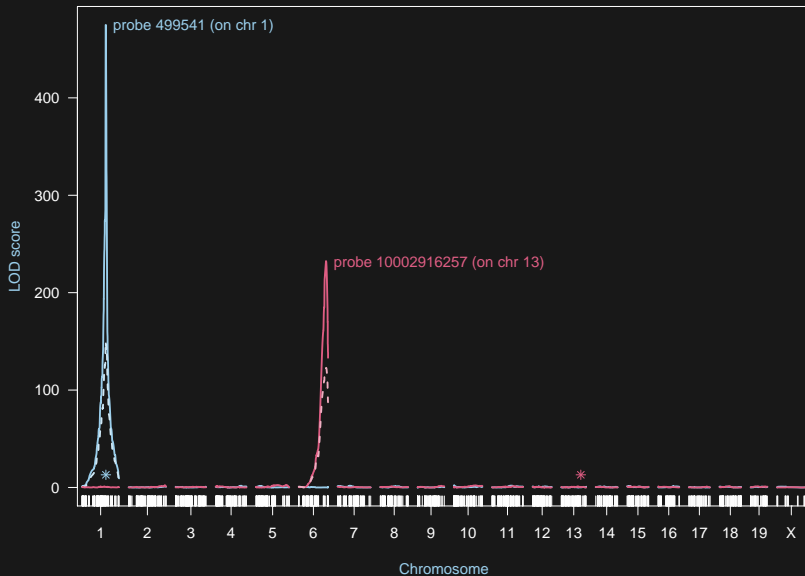
liver



Insulin QTL



Strong eQTL



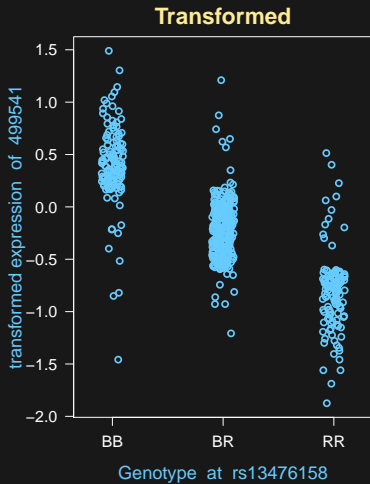
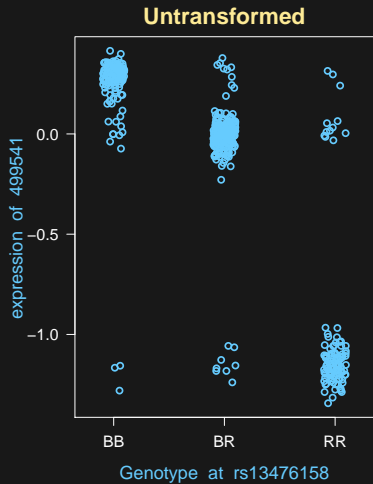
Summary

- ▶ Sample mix-ups happen
- ▶ With eQTL data, we can both identify and **correct** mix-ups
- ▶ There is great value in having expression on multiple tissues
- ▶ The general idea here has wide application for high-throughput data
- ▶ **Broman et al. (2015) G3 5:2177-2186**
doi: 10.1534/g3.115.019778
- ▶ Related work:
 - Westra et al. (2011) Bioinformatics 27:2104–2111
 - Schadt et al. (2012) Nat Genet 44:603–608
 - Ekstrøm and Feenstra (2012) Stat Appl Genet Mol Biol 3:Article 13
 - Lynch et al. (2012) PLoS ONE 7:e41815

Lessons

- ▶ Don't fully trust anyone
 - Including yourself
- ▶ Make lots of plots
 - Don't rely on summary statistics, like LOD scores
 - Look at responses on the original scale
- ▶ Follow up all aberrations
- ▶ Take your time with data cleaning
 - A month, two months, a year?
- ▶ If you have big rectangles whose rows correspond, check that they **actually** correspond

E vs G



Lessons

- ▶ Don't fully trust anyone
 - Including yourself
- ▶ Make lots of plots
 - Don't rely on summary statistics, like LOD scores
 - Look at responses on the original scale
- ▶ Follow up all aberrations
- ▶ Take your time with data cleaning
 - A month, two months, a year?
- ▶ If you have big rectangles whose rows correspond, check that they **actually** correspond

Acknowledgments

Alan Attie
Mark Keller

Biochemistry, UW–Madison

Brian Yandell

Statistics and Horticulture, UW–Madison

Christina Kendziorski
Aimee Teo Broman

Biostat & Medical Info, UW–Madison

Eric Schadt

Mount Sinai

Danielle Greenawalt
Amit Kulkarni

Merck & Co., Inc.

Śaunak Sen

UT-Memphis

NIH: R01 GM074244, R01 DK066369

Slides: bit.ly/berkeley2017



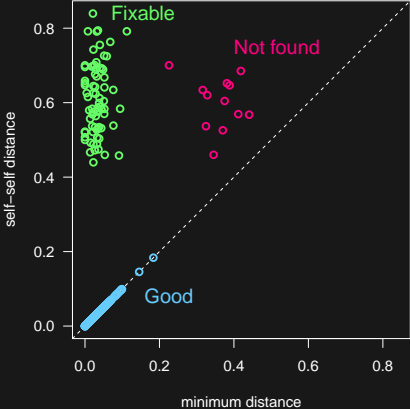
kbroman.org

github.com/kbroman

@kwbroman

Decisions

Self vs best



Next-best vs best

