

Salvaging a genetics project

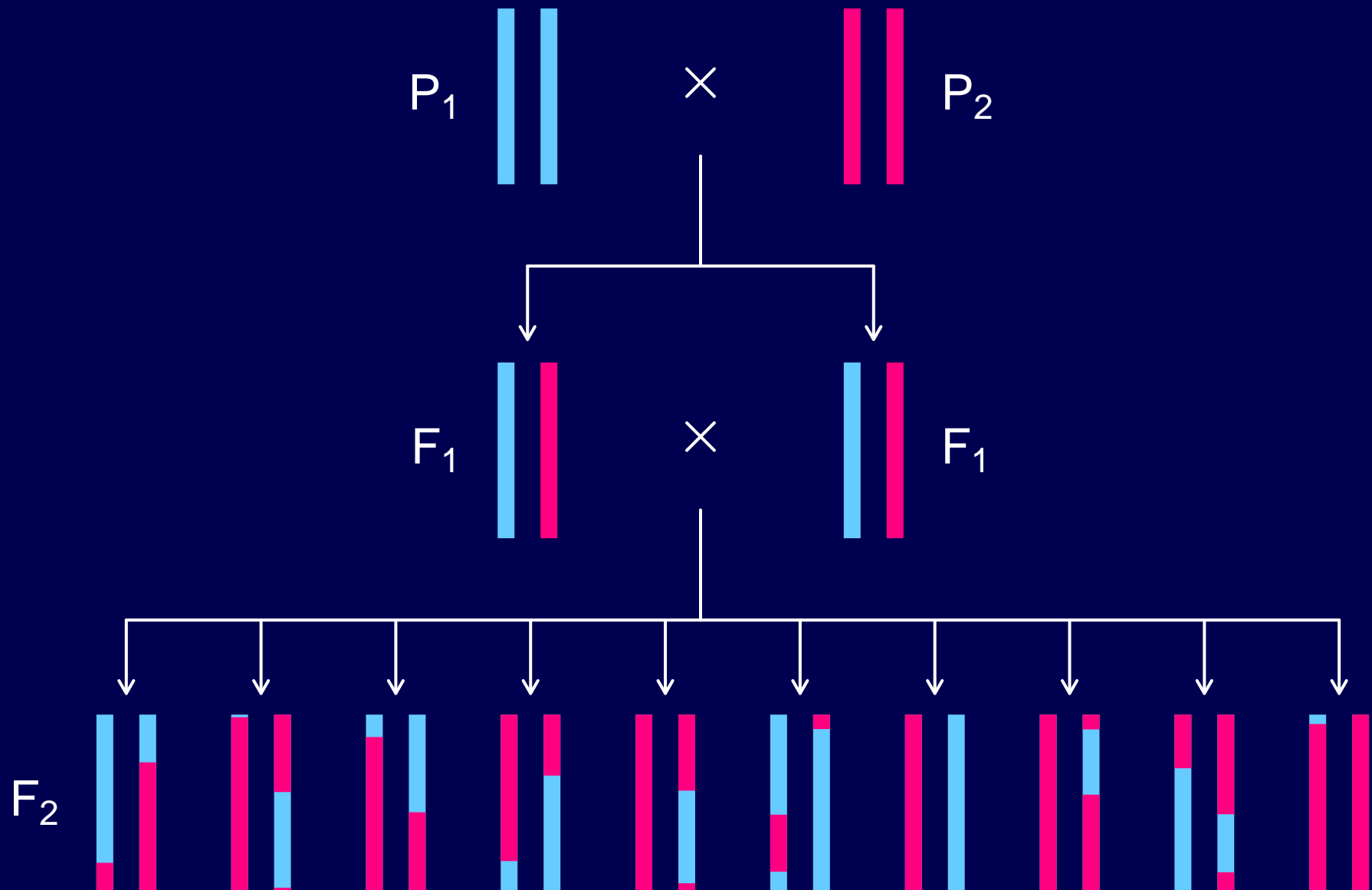
Identifying and correcting sample mix-ups
in high-dimensional data

Karl W Broman

Department of Biostatistics & Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman

Intercross



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

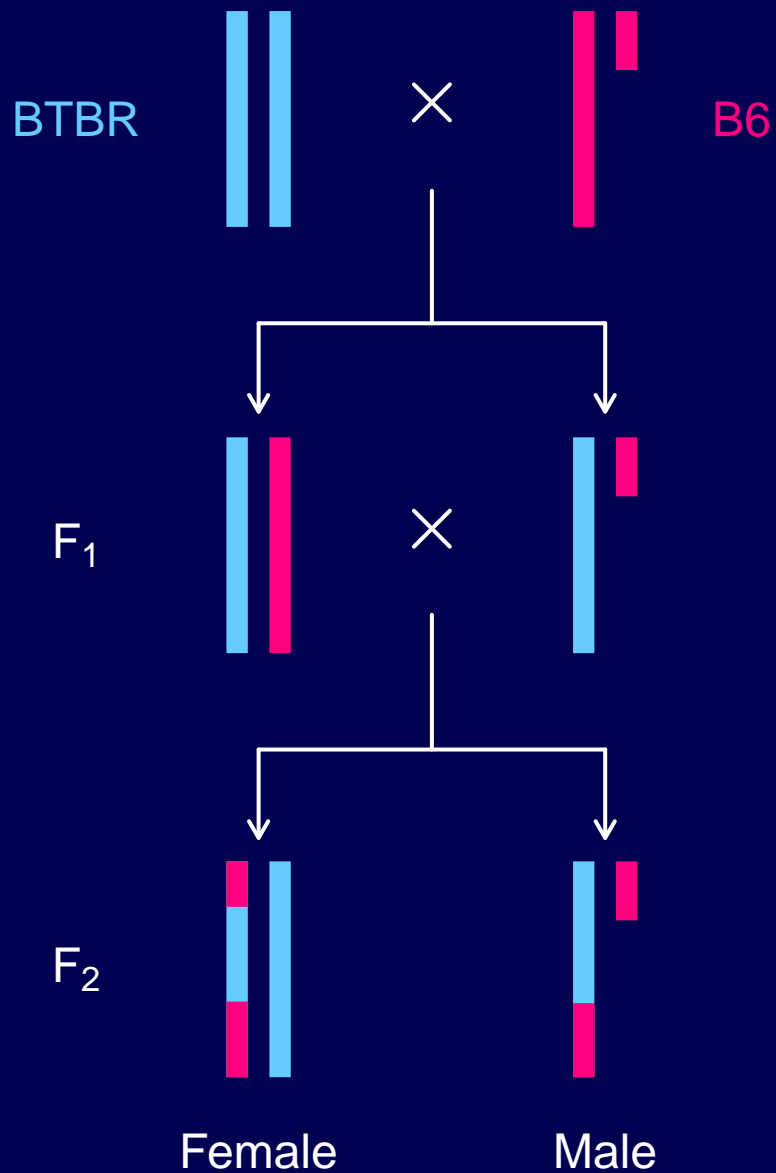
Numerous clinical phenotypes

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

Considerable brain power

Alan Attie, Karl Broman, Aimee Teo Broman, Danielle Greenawalt, Mark Keller, Christina Kendzierski, Amit Kulkarni, Eric Schadt, Brian Yandell, . . .

Sex and the X chr



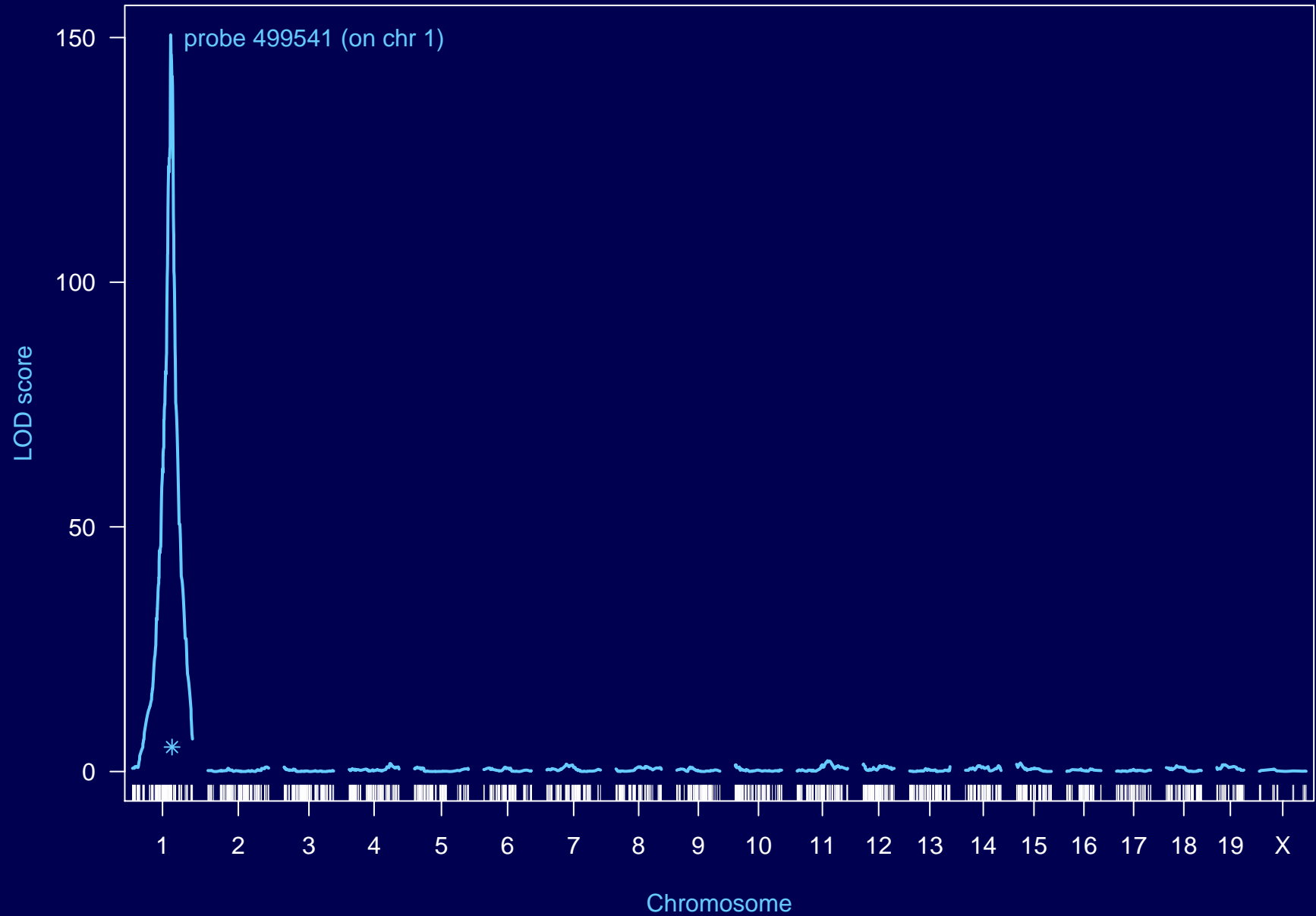
F₂ females: R/R or B/R

F₂ males: hemizygous B or R

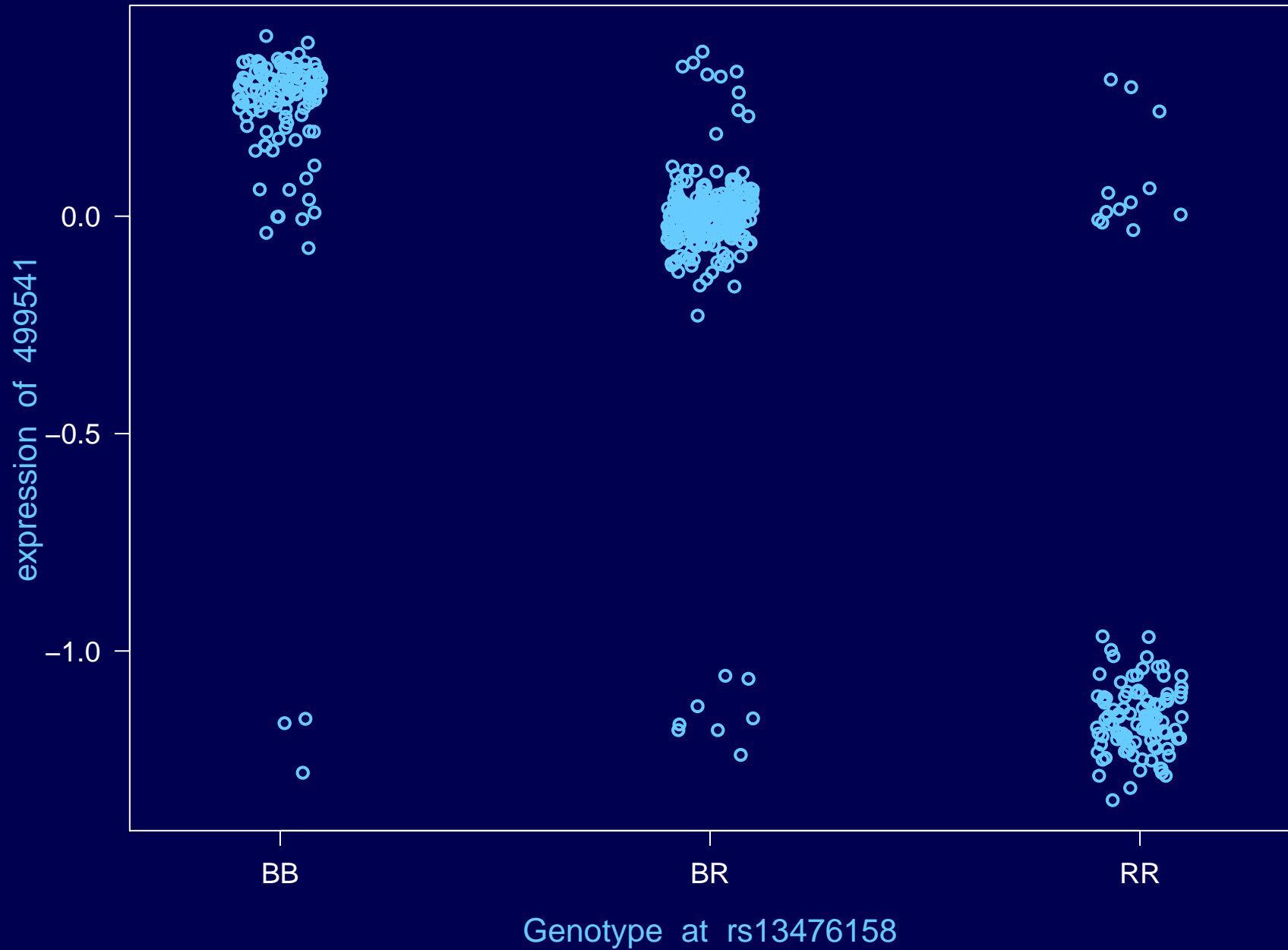
Genotype mix-ups



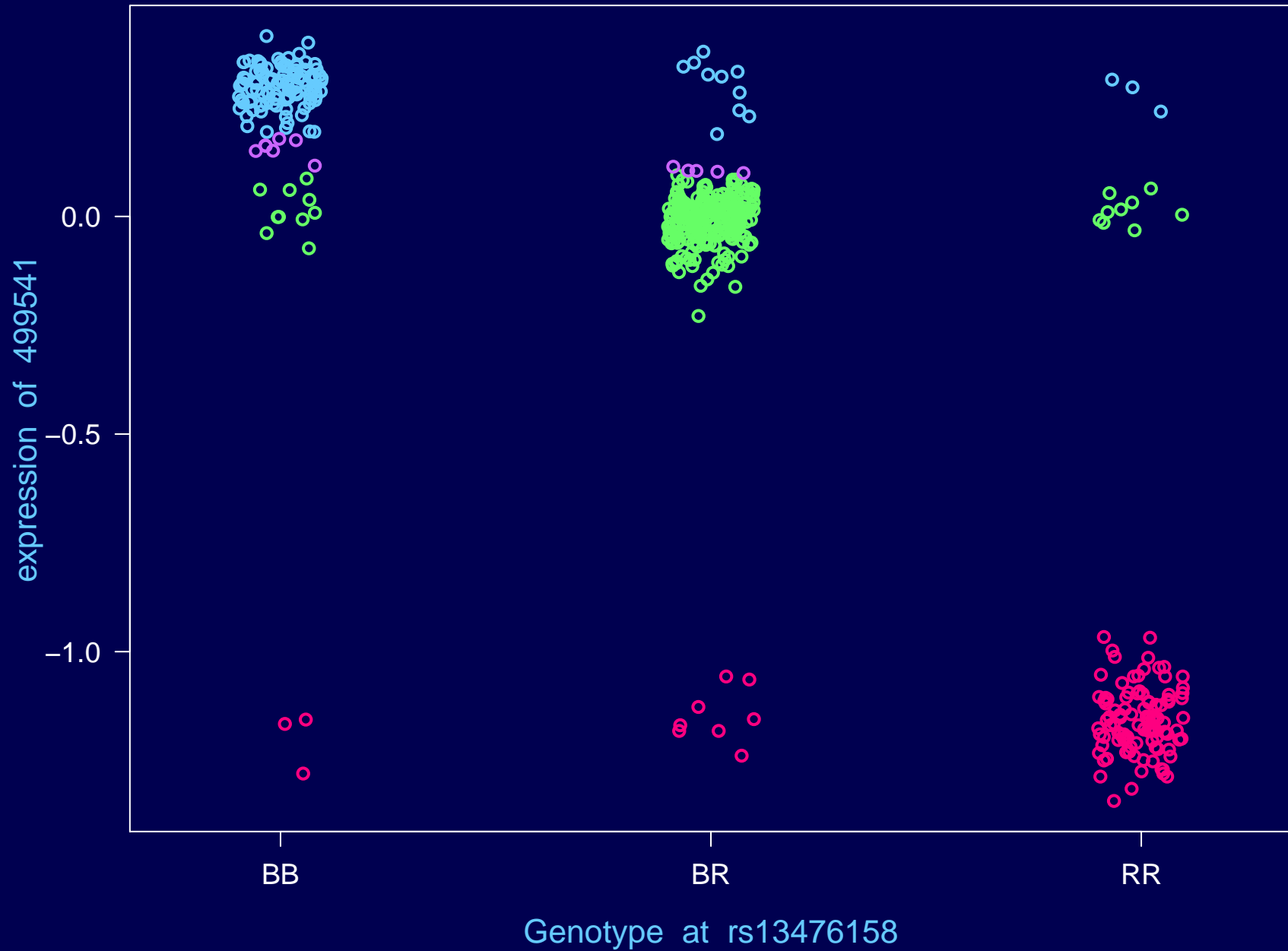
Strong eQTL



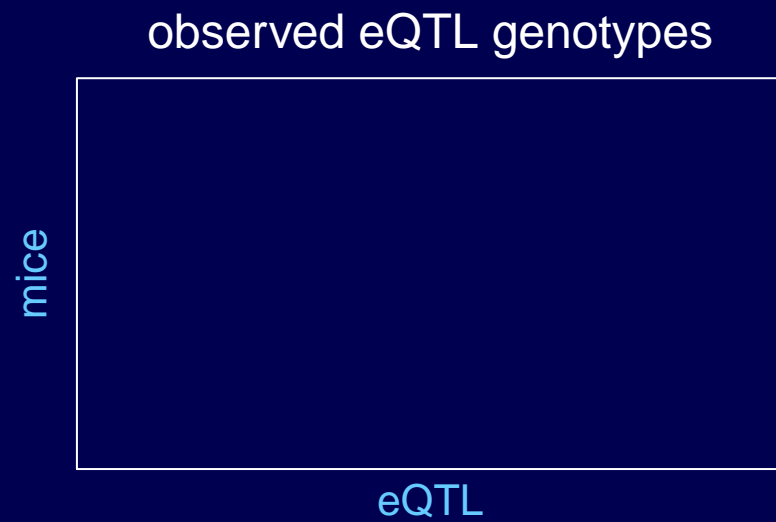
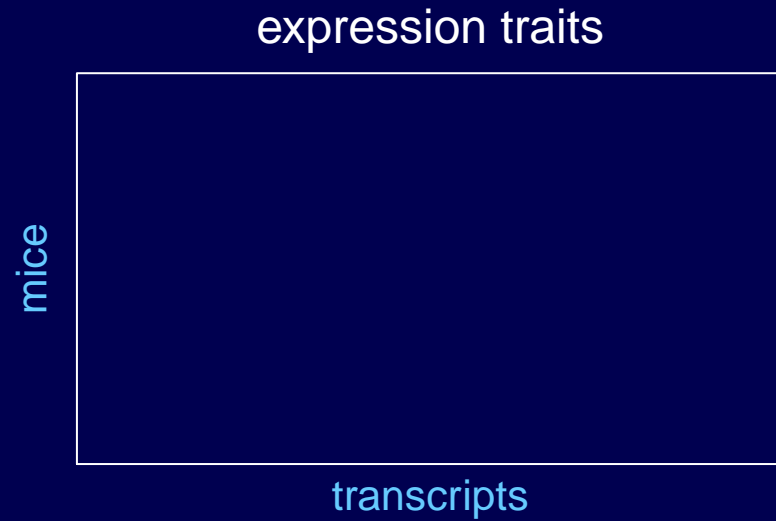
E vs G



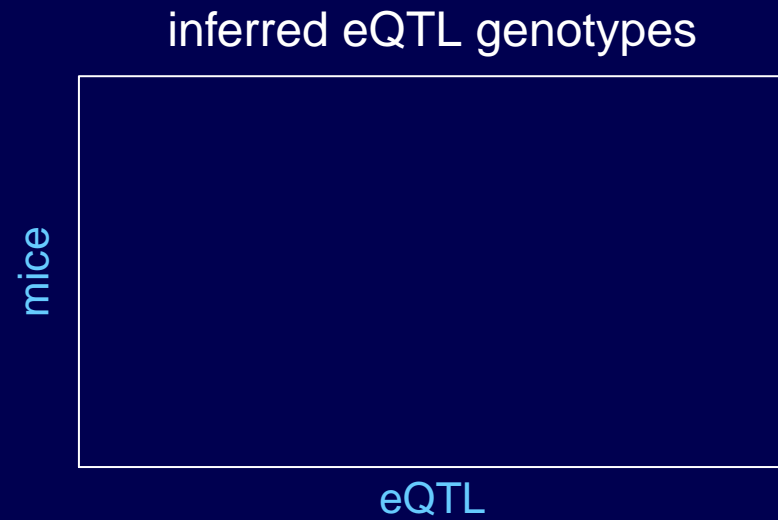
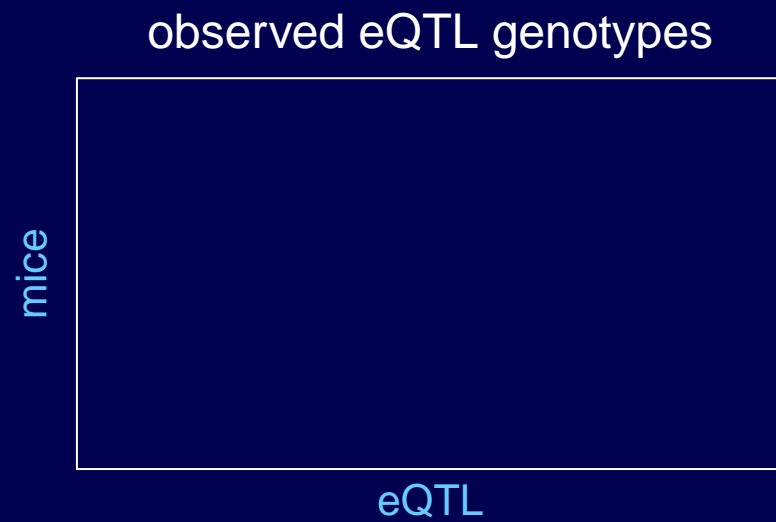
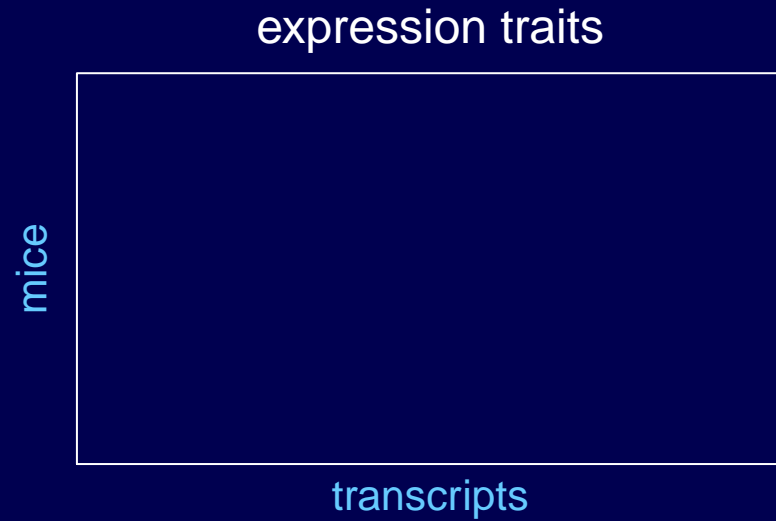
kNN classifier



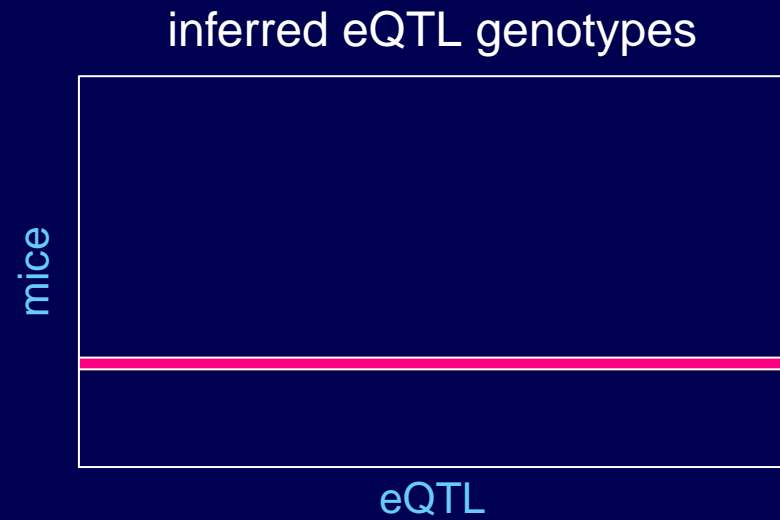
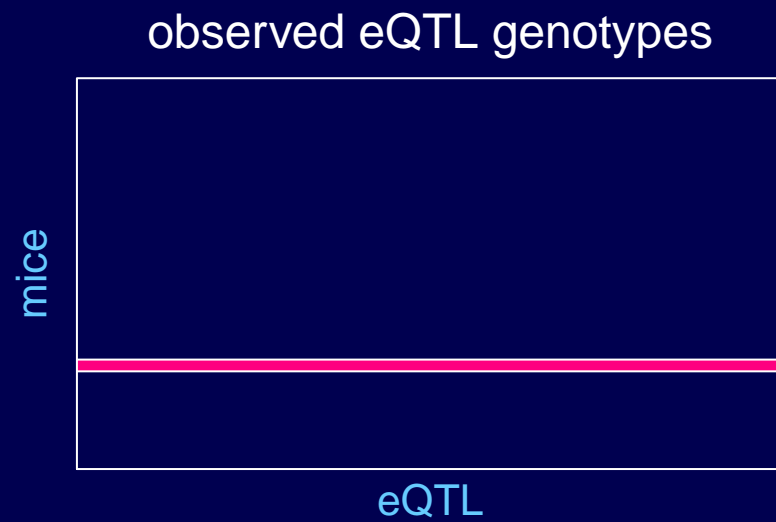
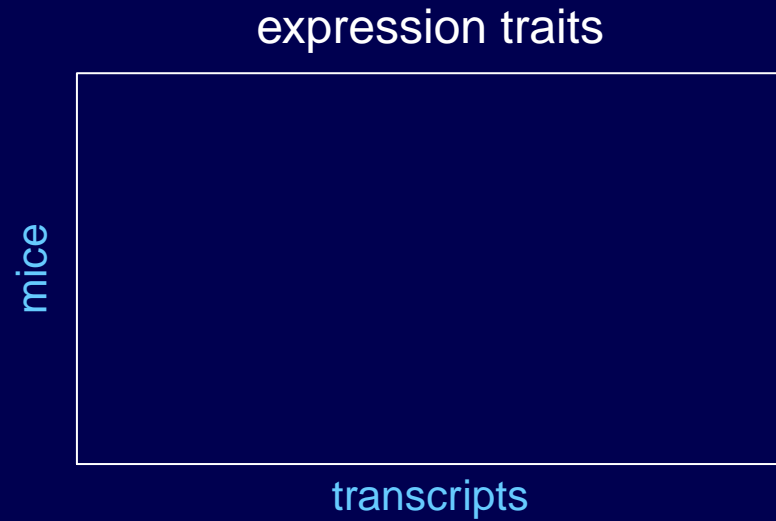
Basic scheme



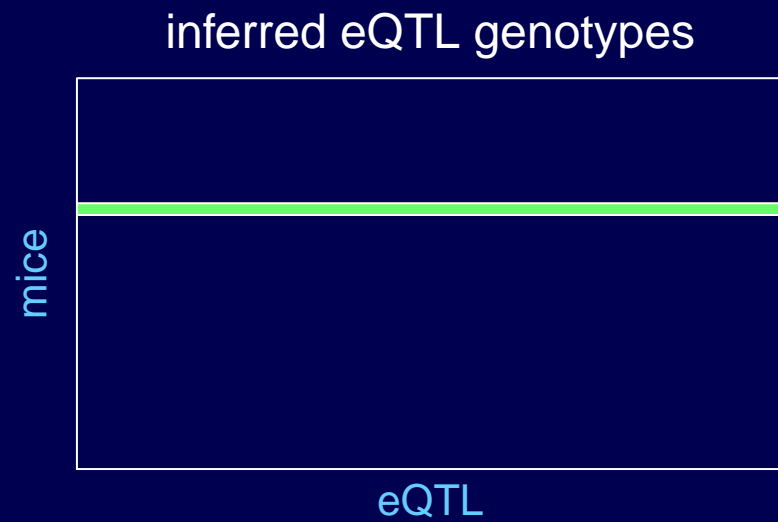
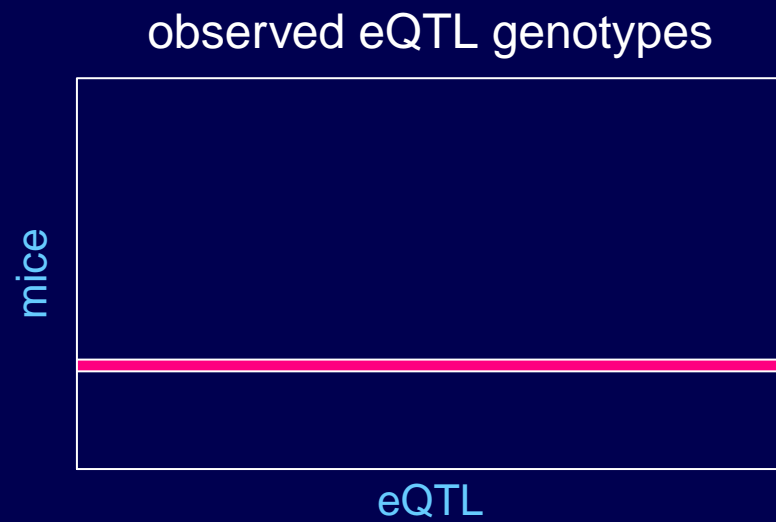
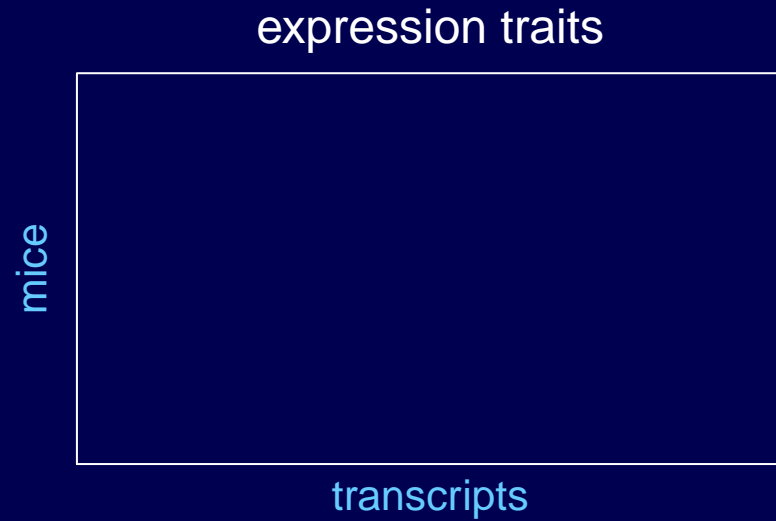
Basic scheme



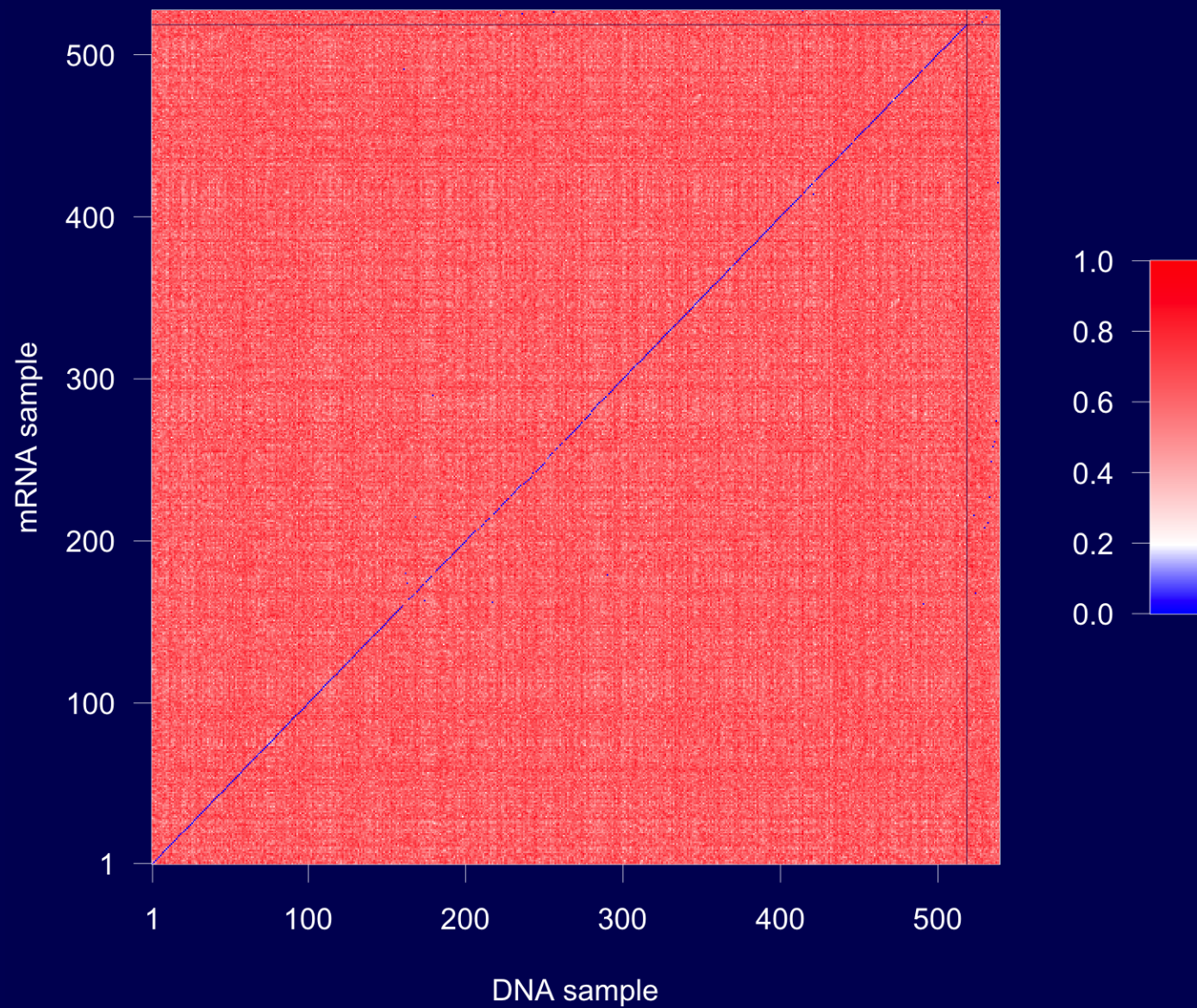
Basic scheme



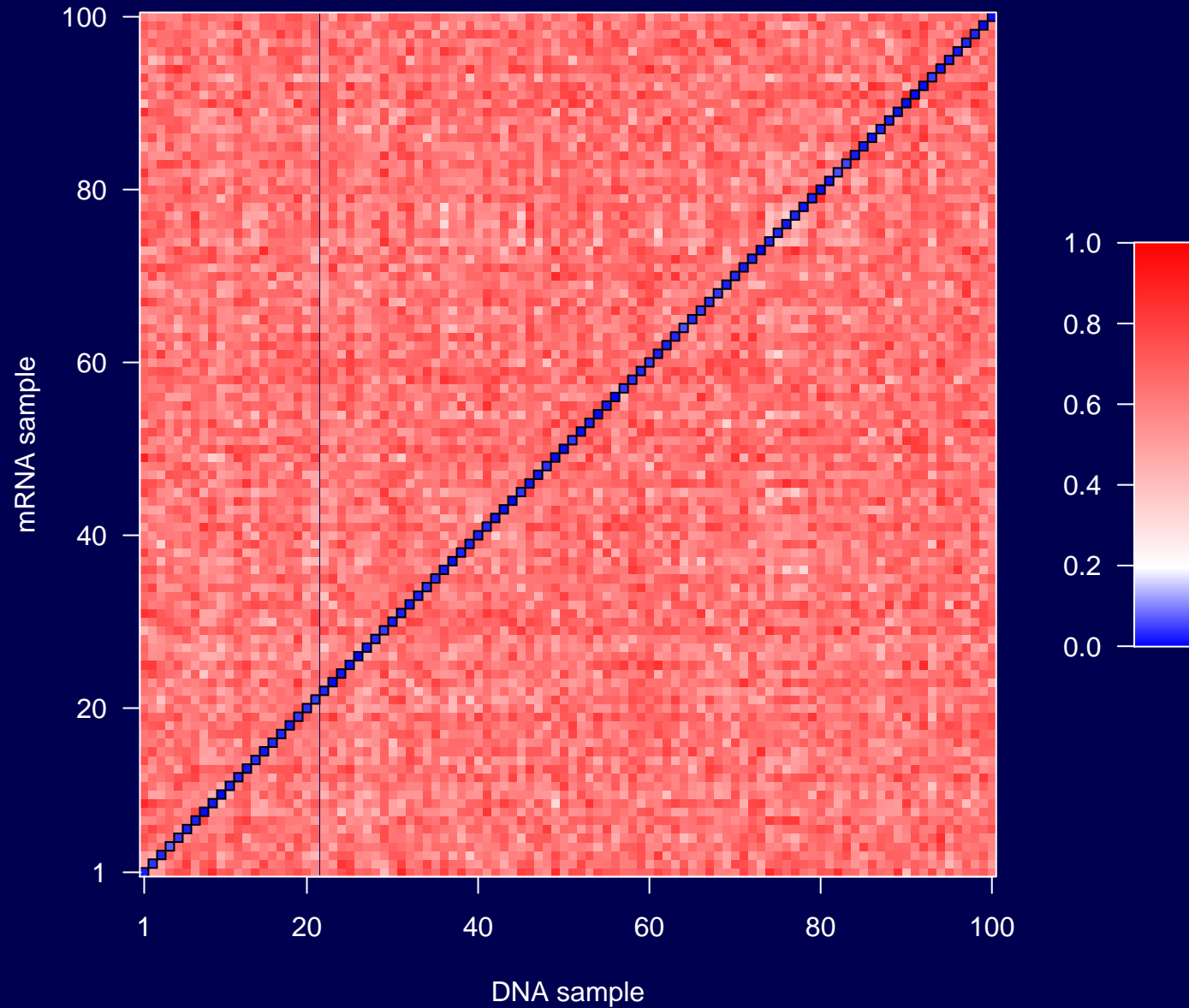
Basic scheme



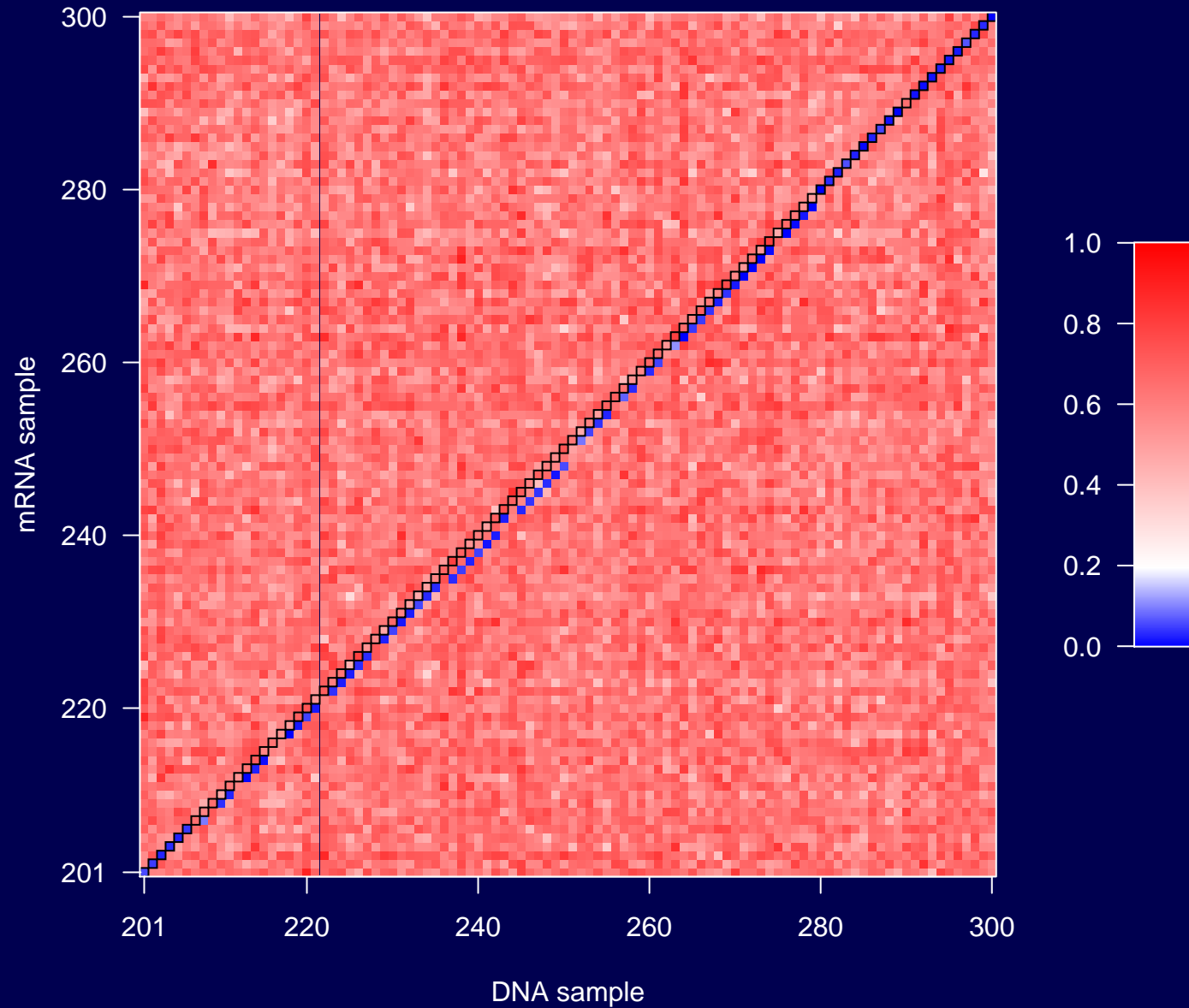
Prop'n mismatches



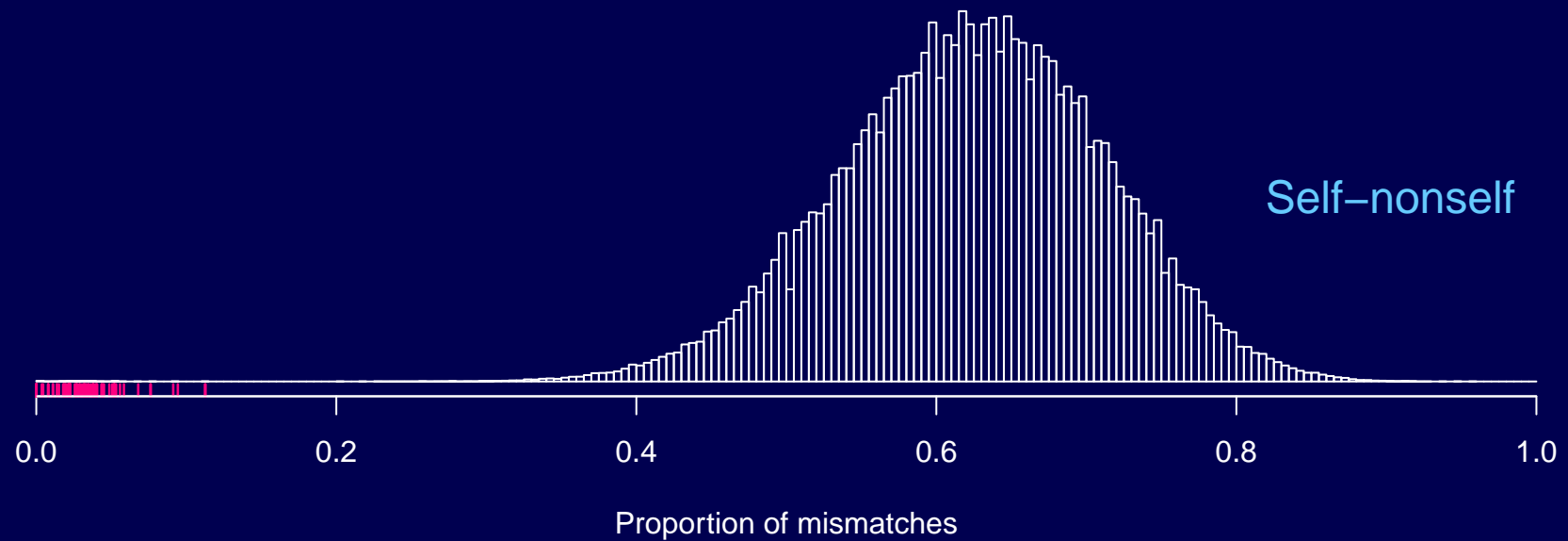
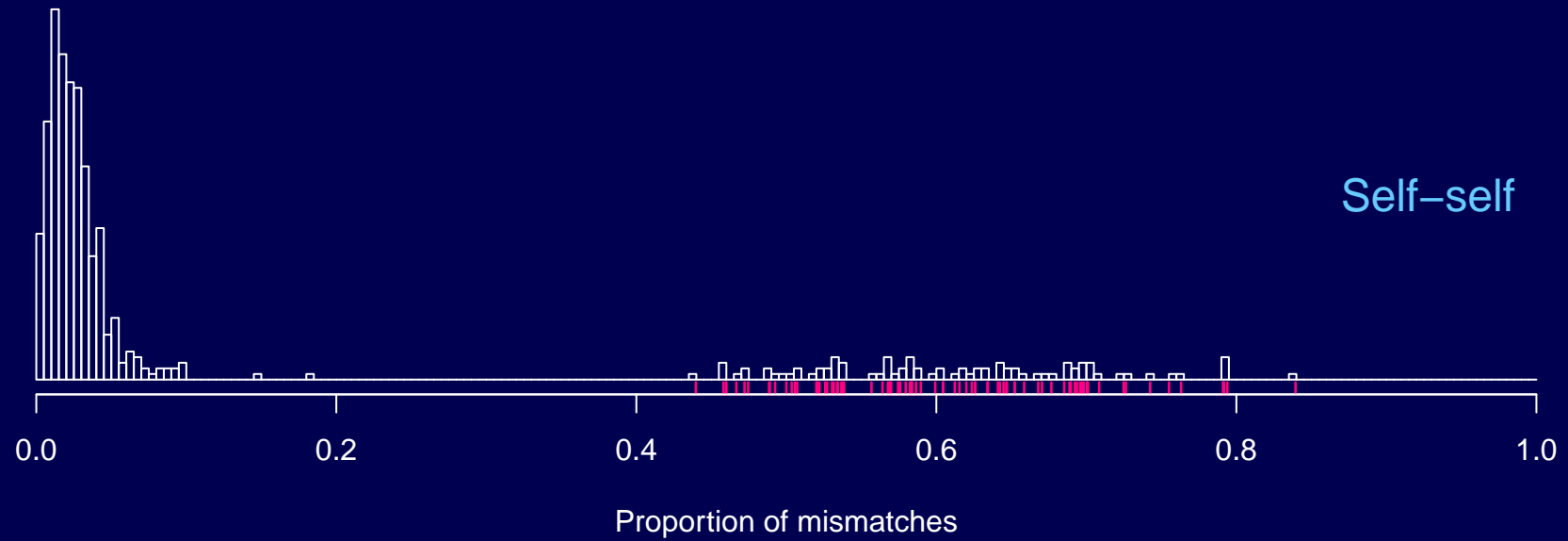
Prop'n mismatches



Prop'n mismatches



Prop'n mismatches



Decisions

Self vs best

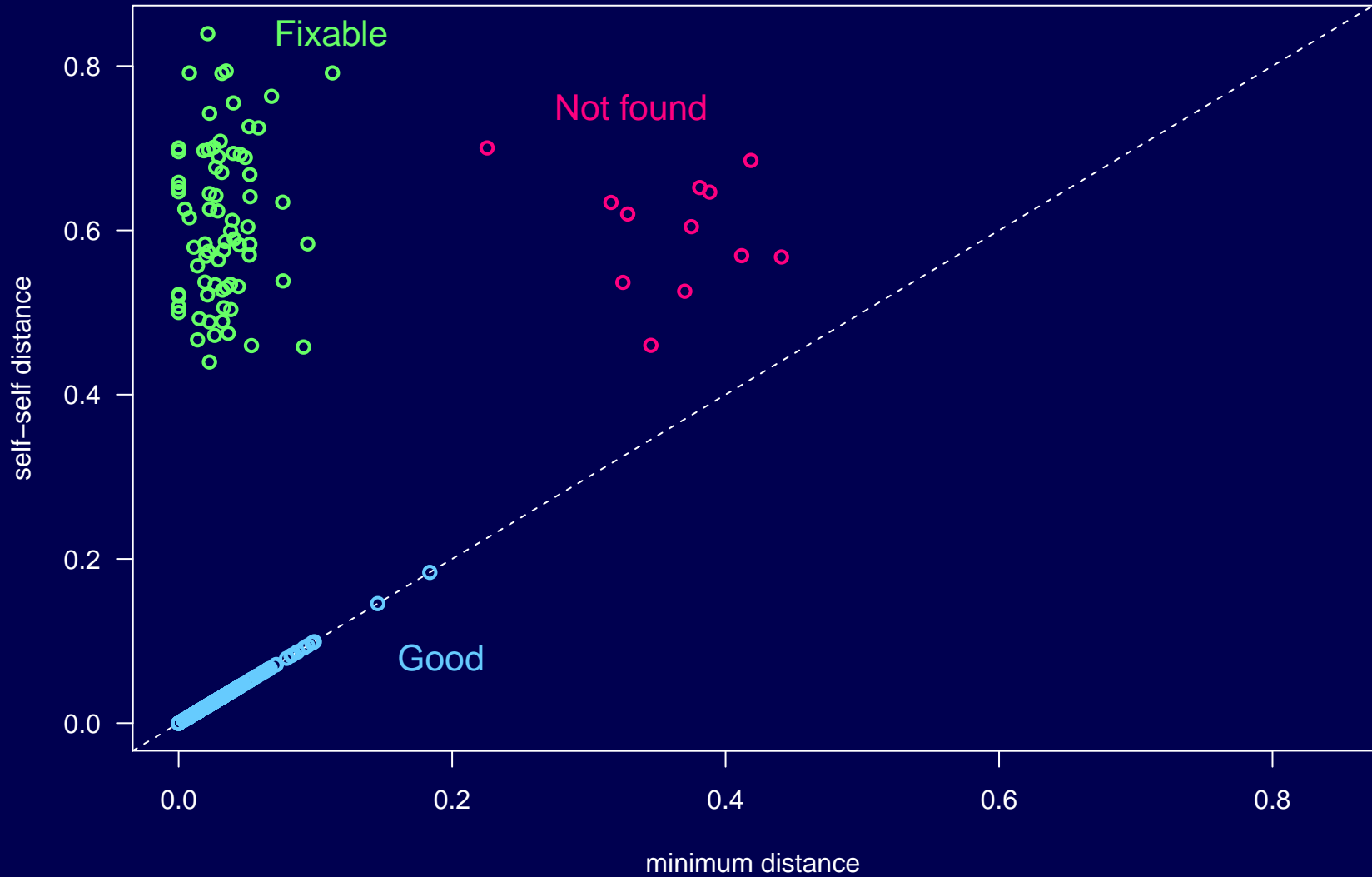
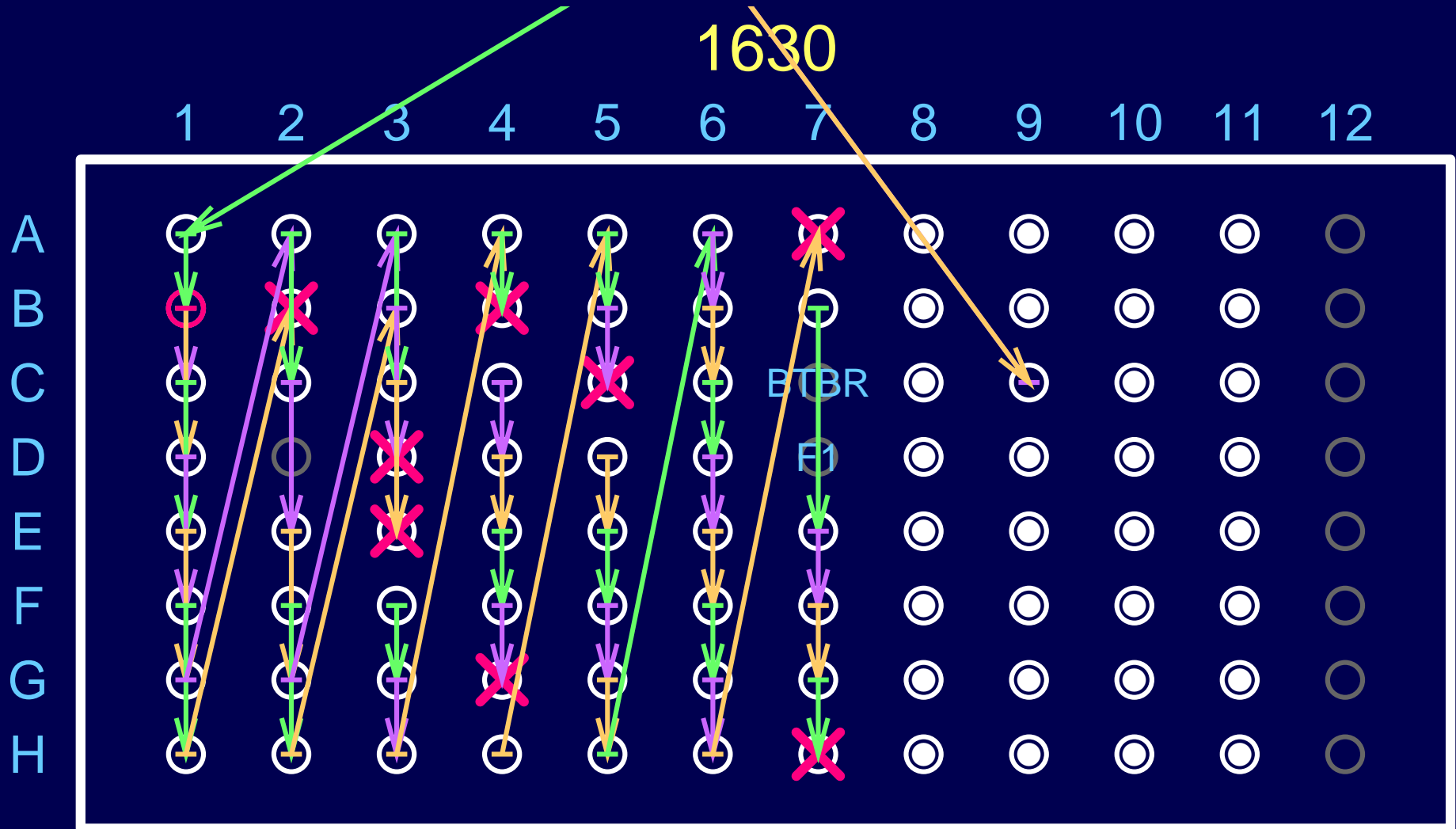
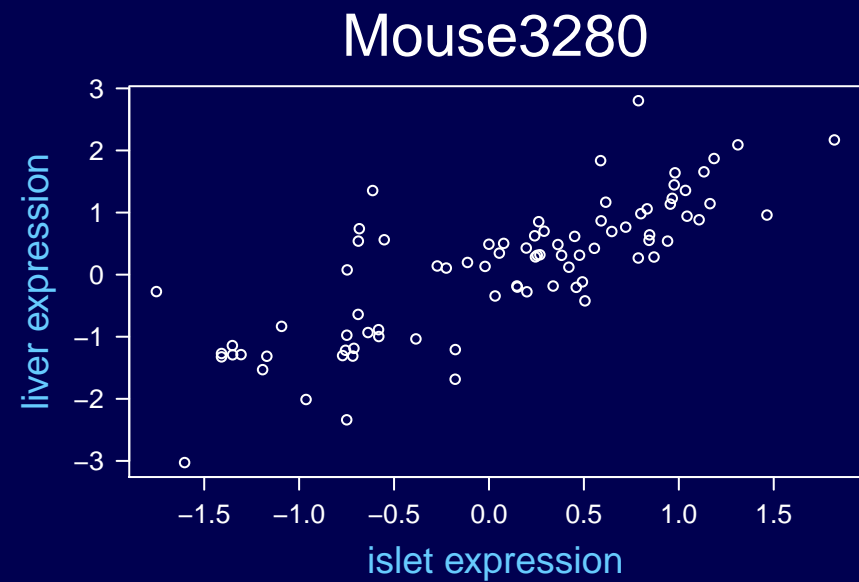


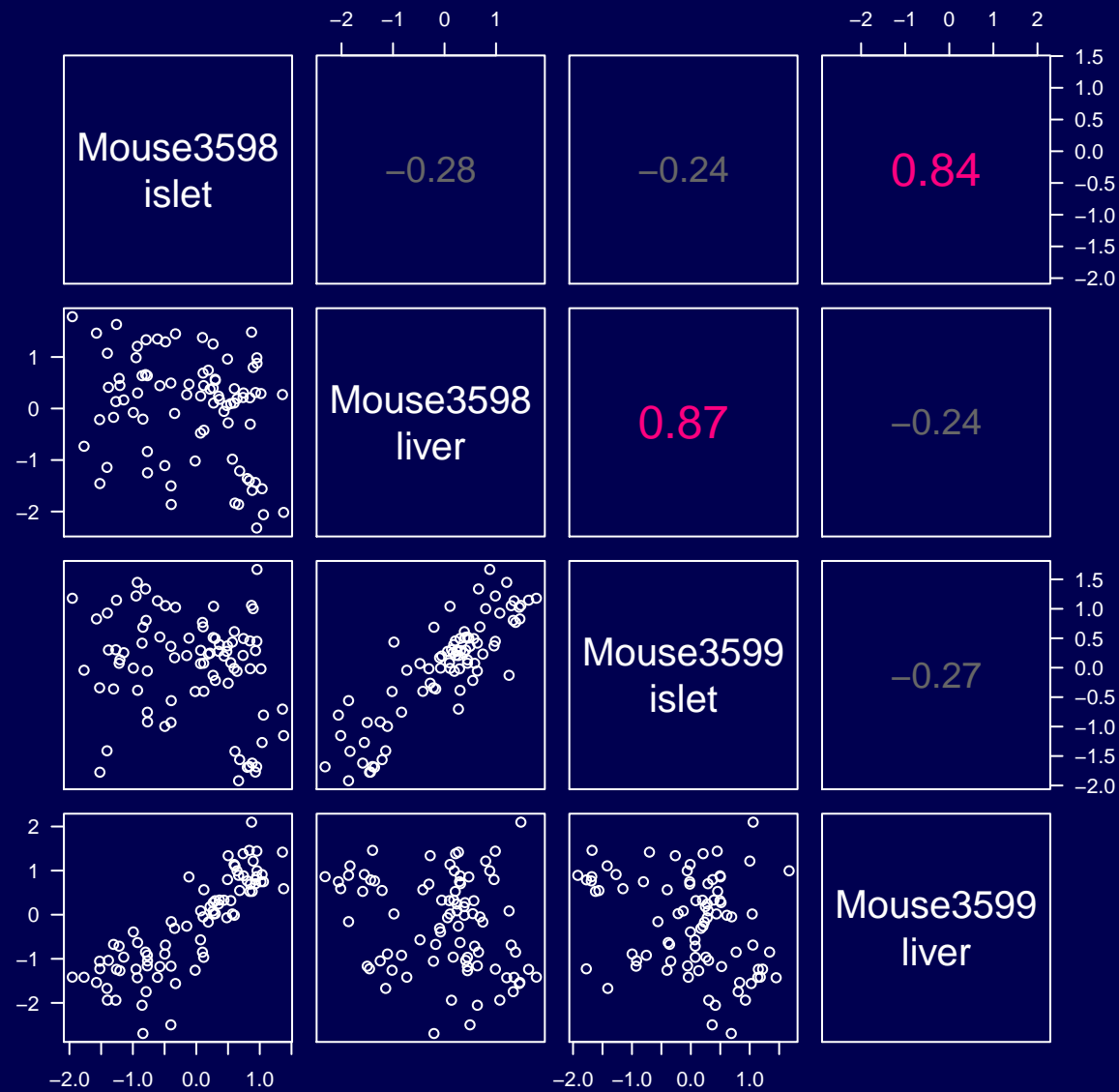
Plate 1630



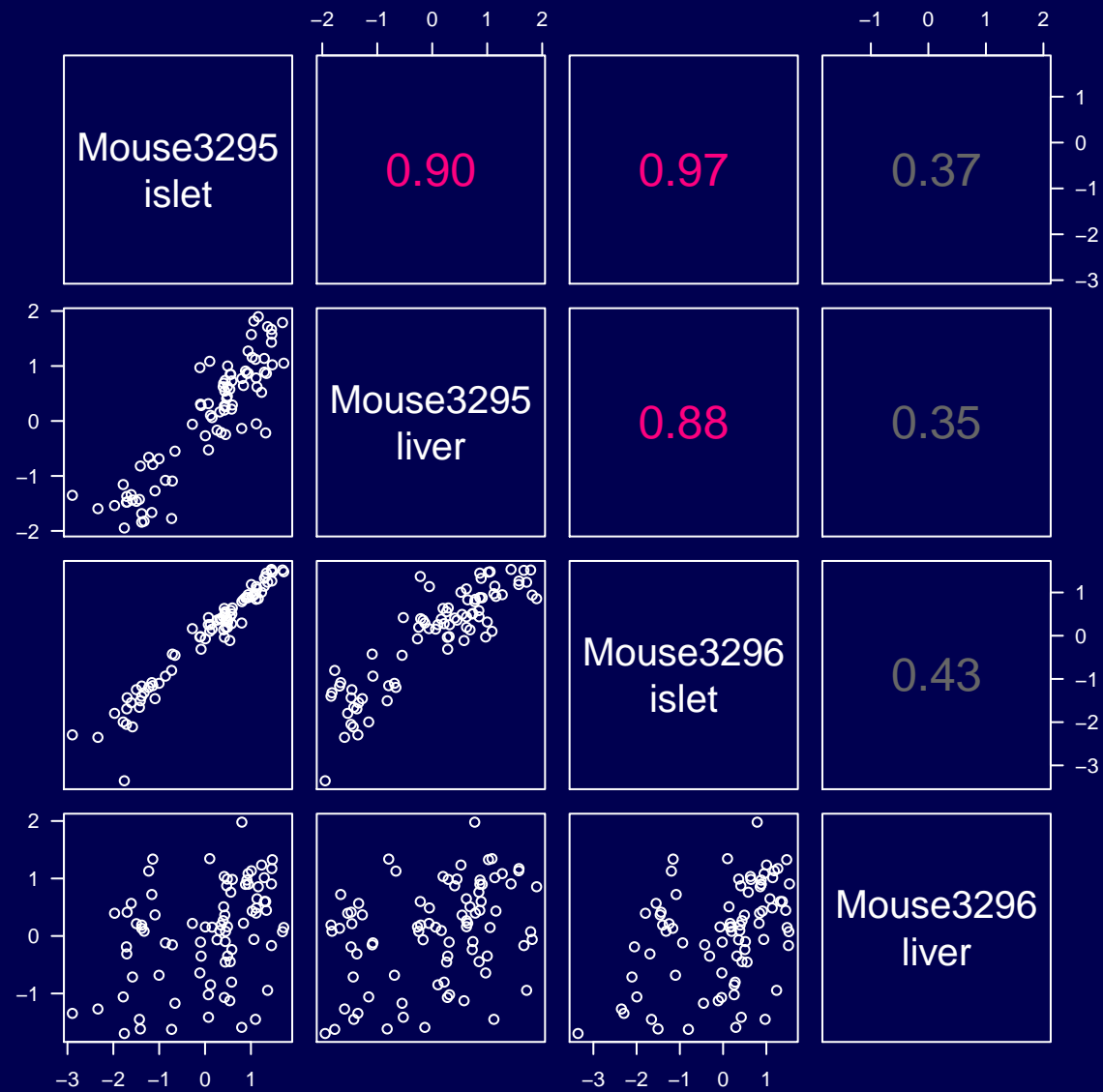
E vs E



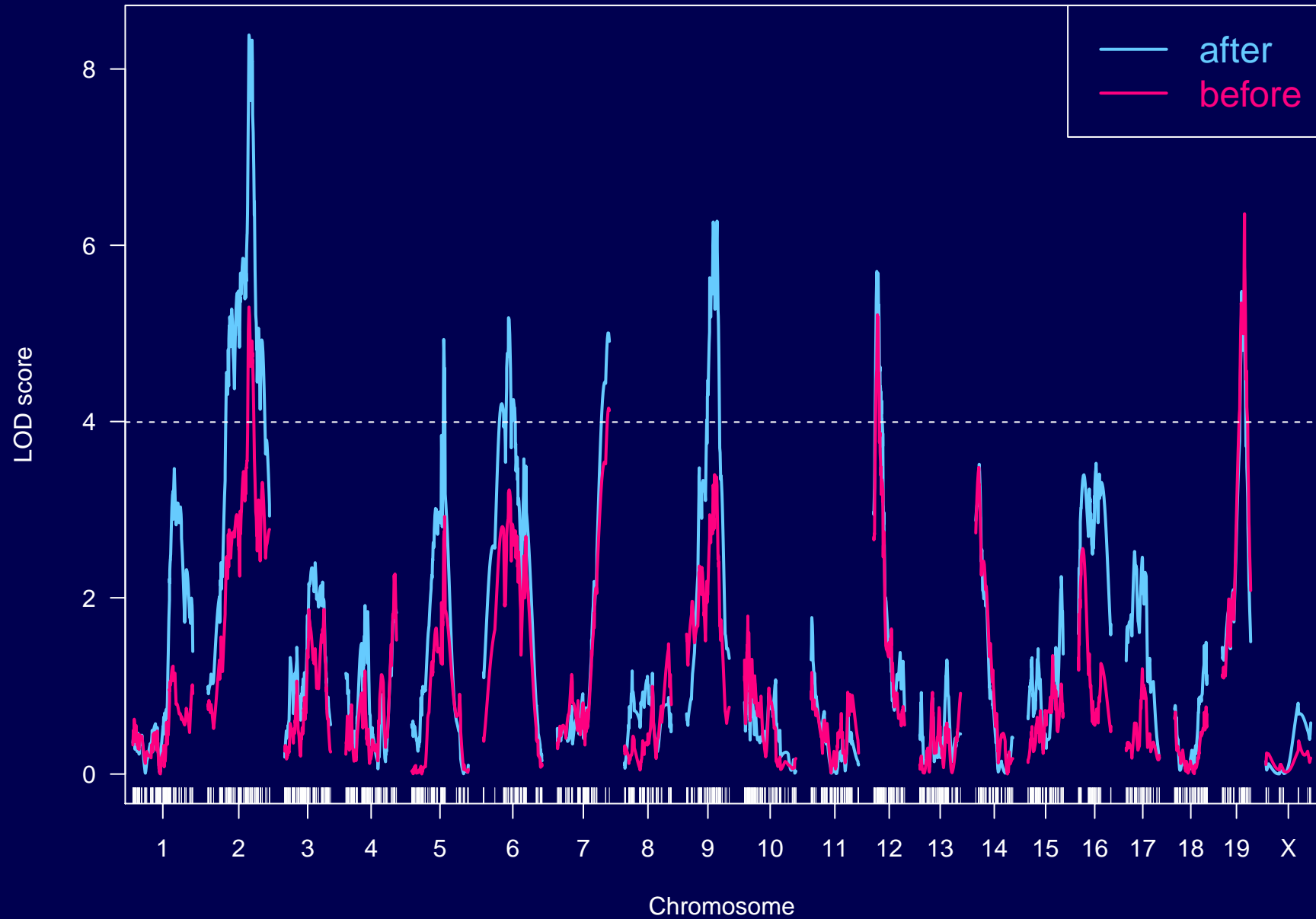
E vs E



E vs E



Insulin QTL



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
- Very similar to **MixupMapper** (Westra et al., *Bioinformatics* 27:2104–2111, 2011)
 - Multiple tissues
 - Direct tissue-tissue comparisons
 - Predict genotype rather than expression phenotype

Acknowledgments

Alan Attie
Mark Keller

Biochemistry, UW–Madison

Brian Yandell

Statistics and Horticulture, UW–Madison

Christina Kendzierski
Aimee Teo Broman

Biostatistics & Medical Informatics, UW–Madison

Eric Schadt

Pacific Biosciences of California

Danielle Greenawalt
Amit Kulkarni

Merck & Co., Inc.

Śaunak Sen

UCSF

NIH: R01 GM074244, R01 DK066369