

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

Karl W Broman

Department of Biostatistics & Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman

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

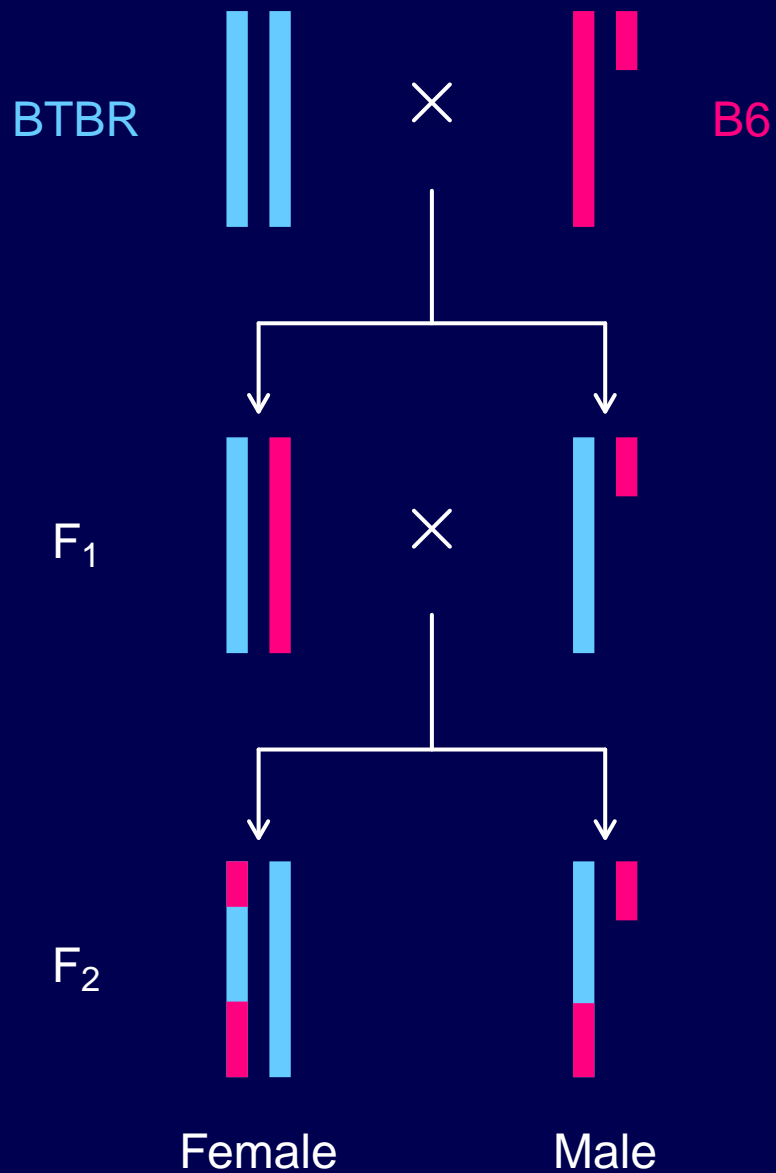
kidney

liver

Numerous clinical phenotypes

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

Sex and the X chr



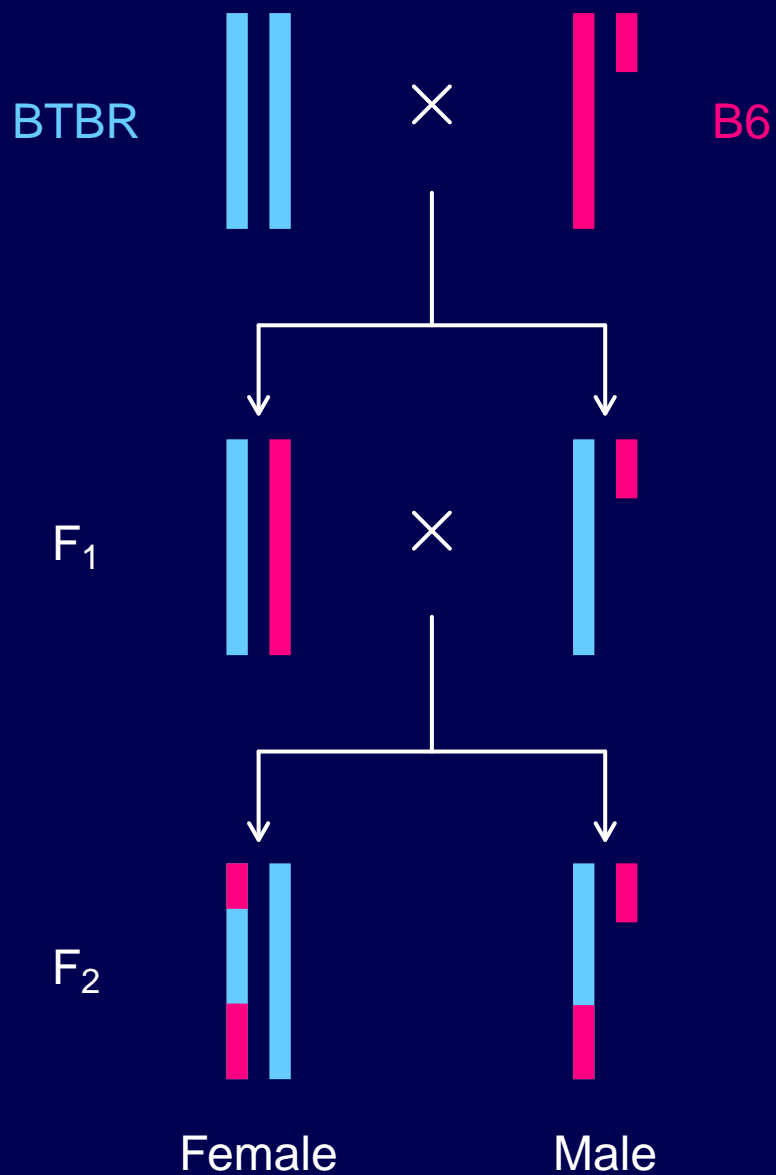
F₂ females: R/R or B/R

F₂ males: hemizygous B or R

Genotype mix-ups



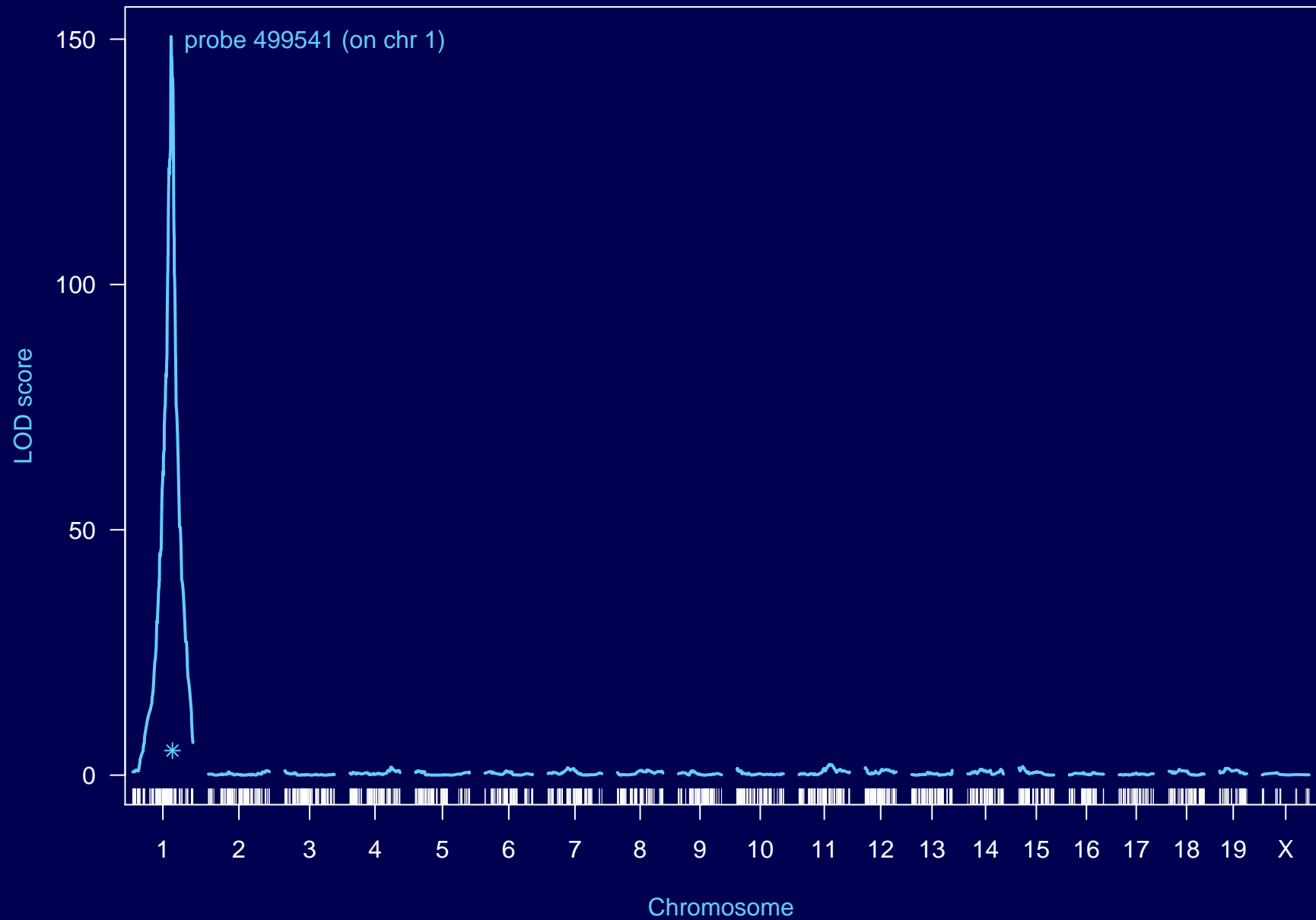
Sex and the X chr



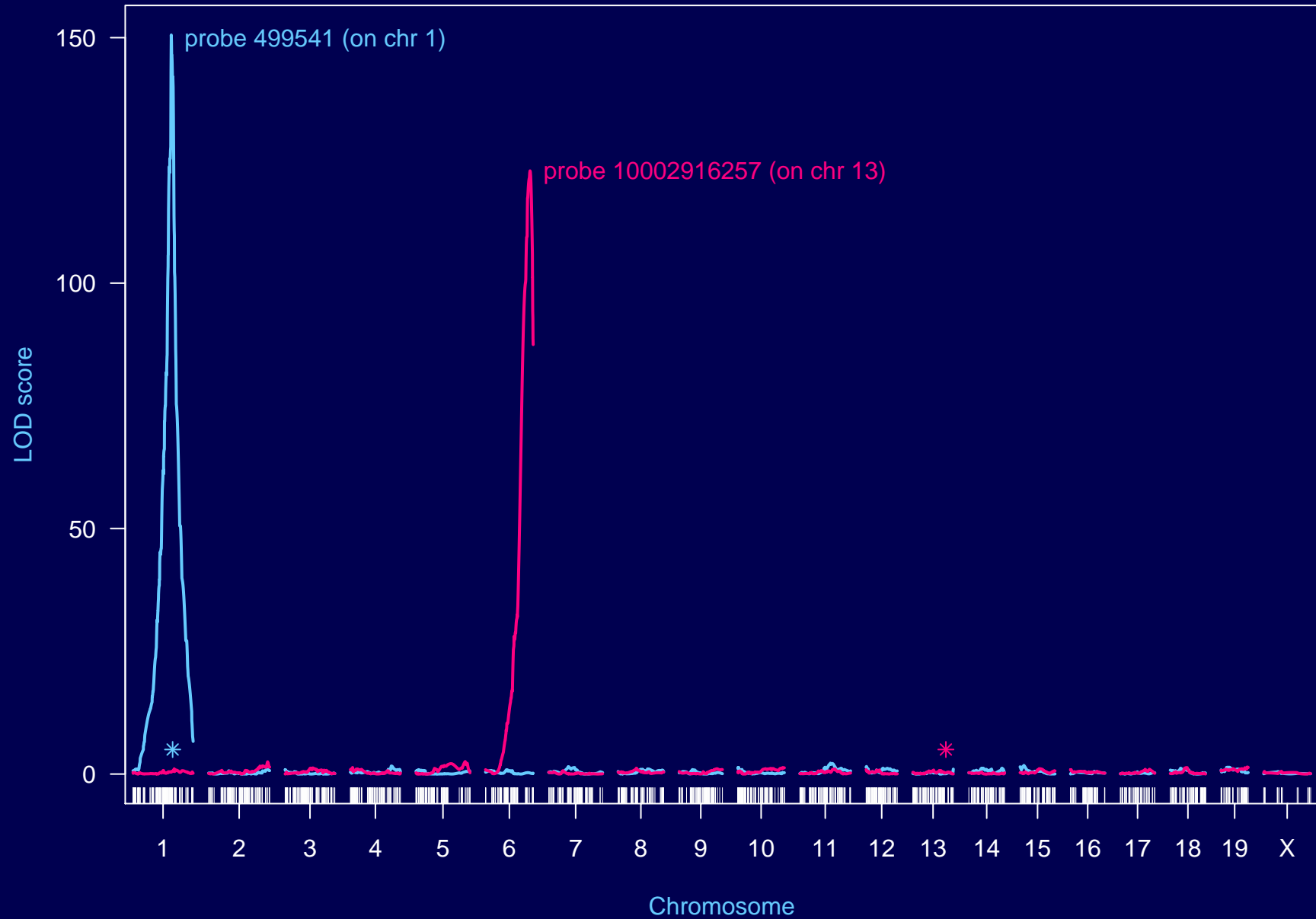
F₂ females: R/R or B/R

F₂ males: hemizygous B or R

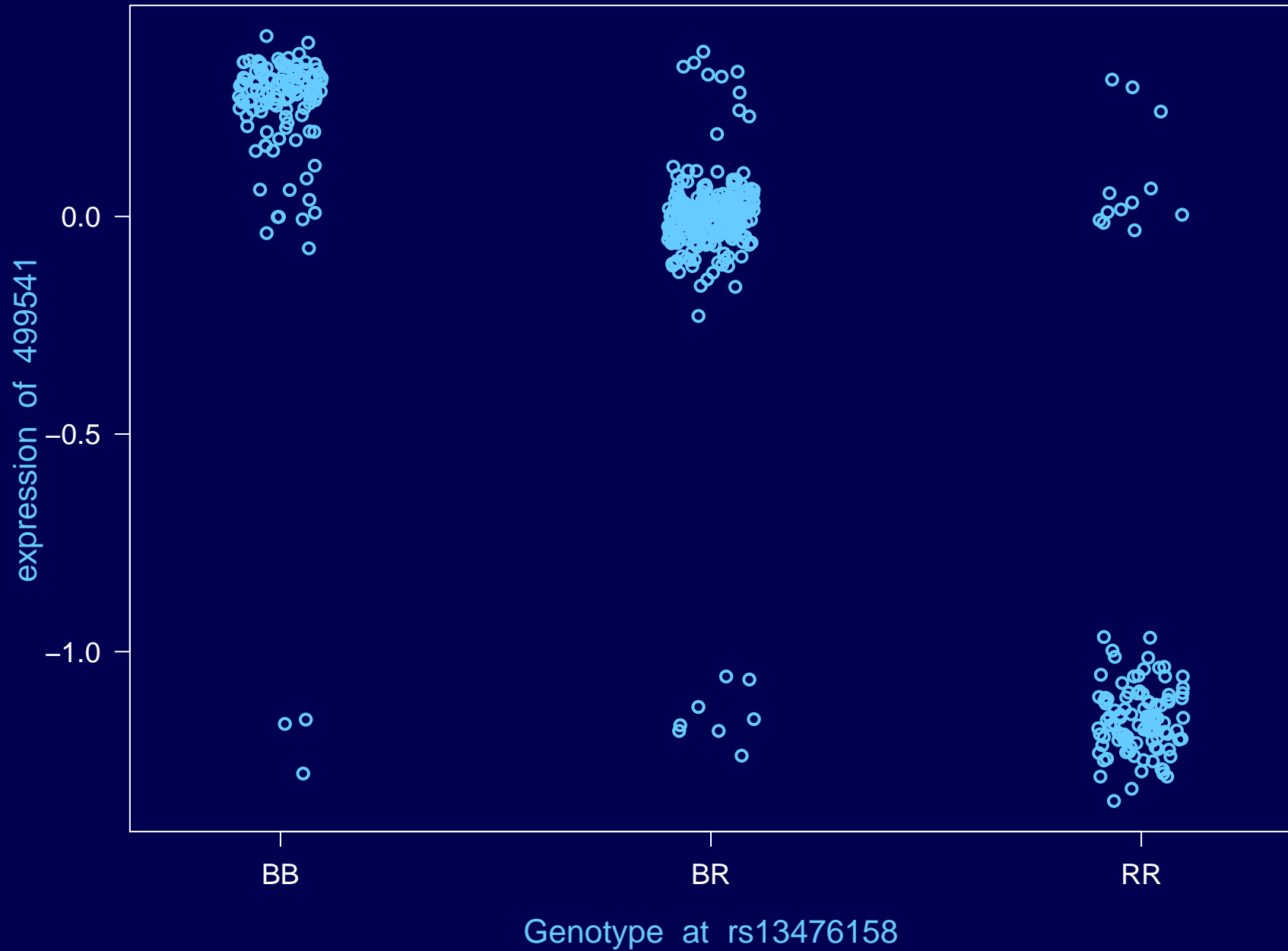
Strong eQTL



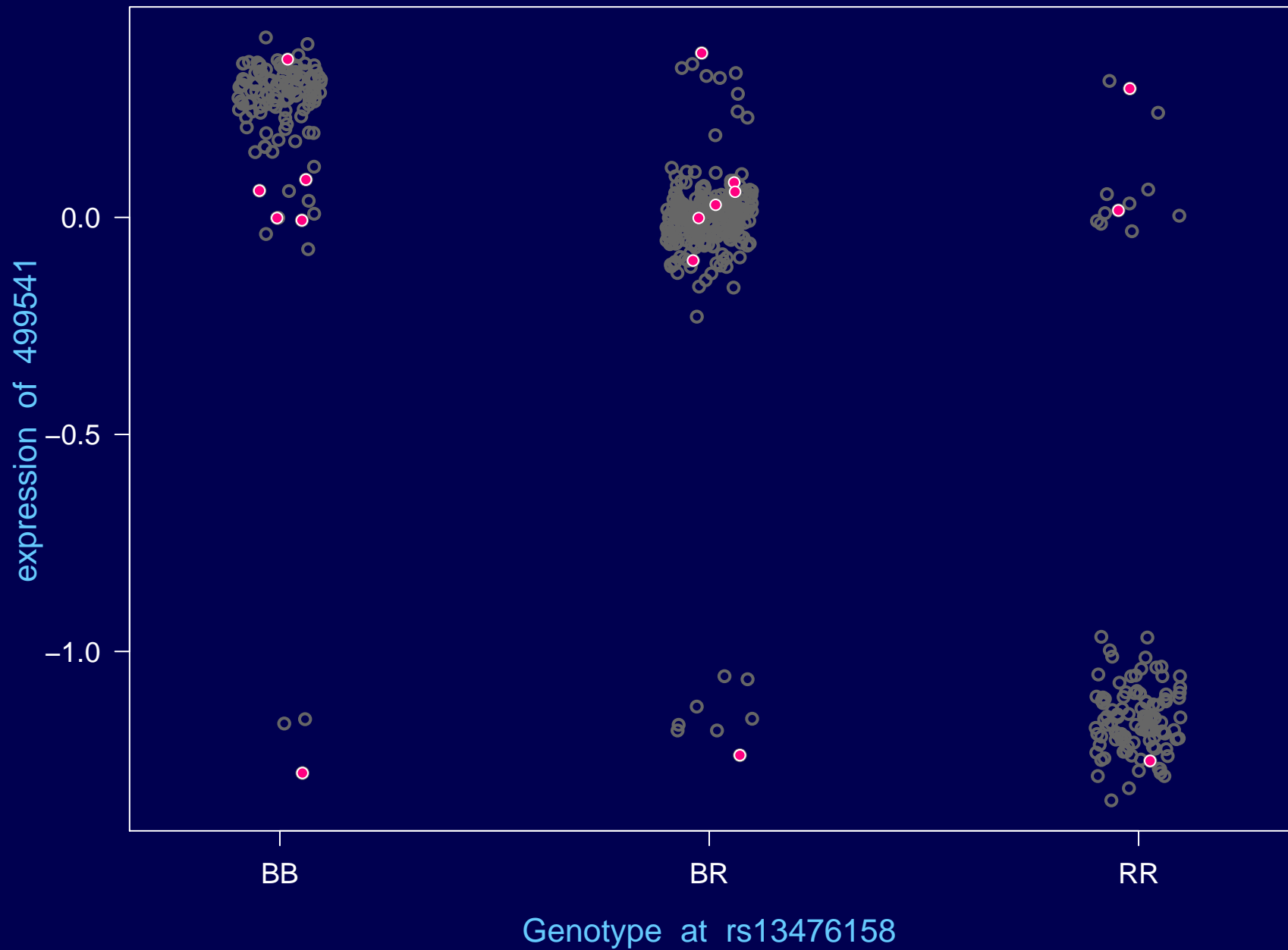
Strong eQTL



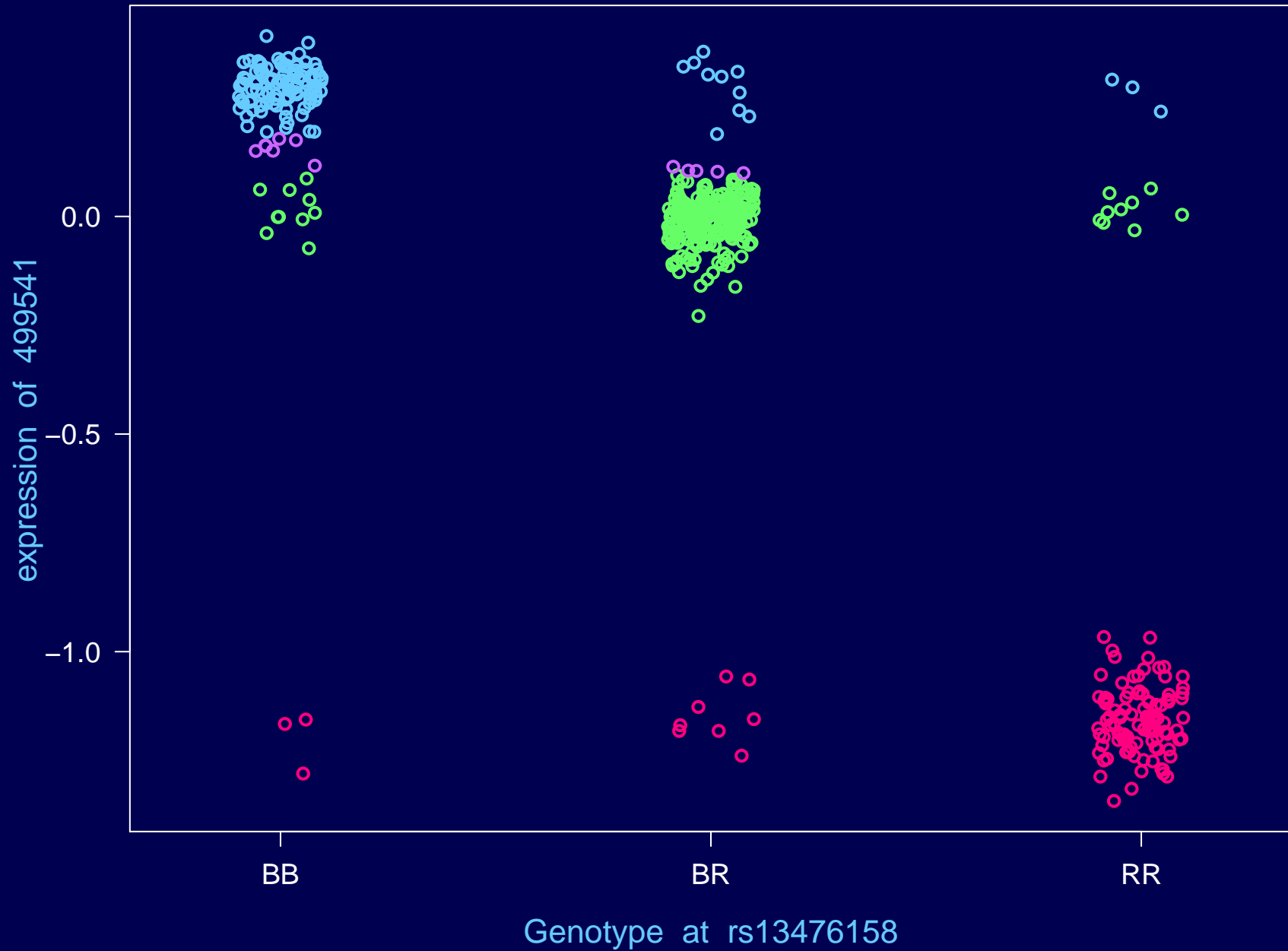
E vs G



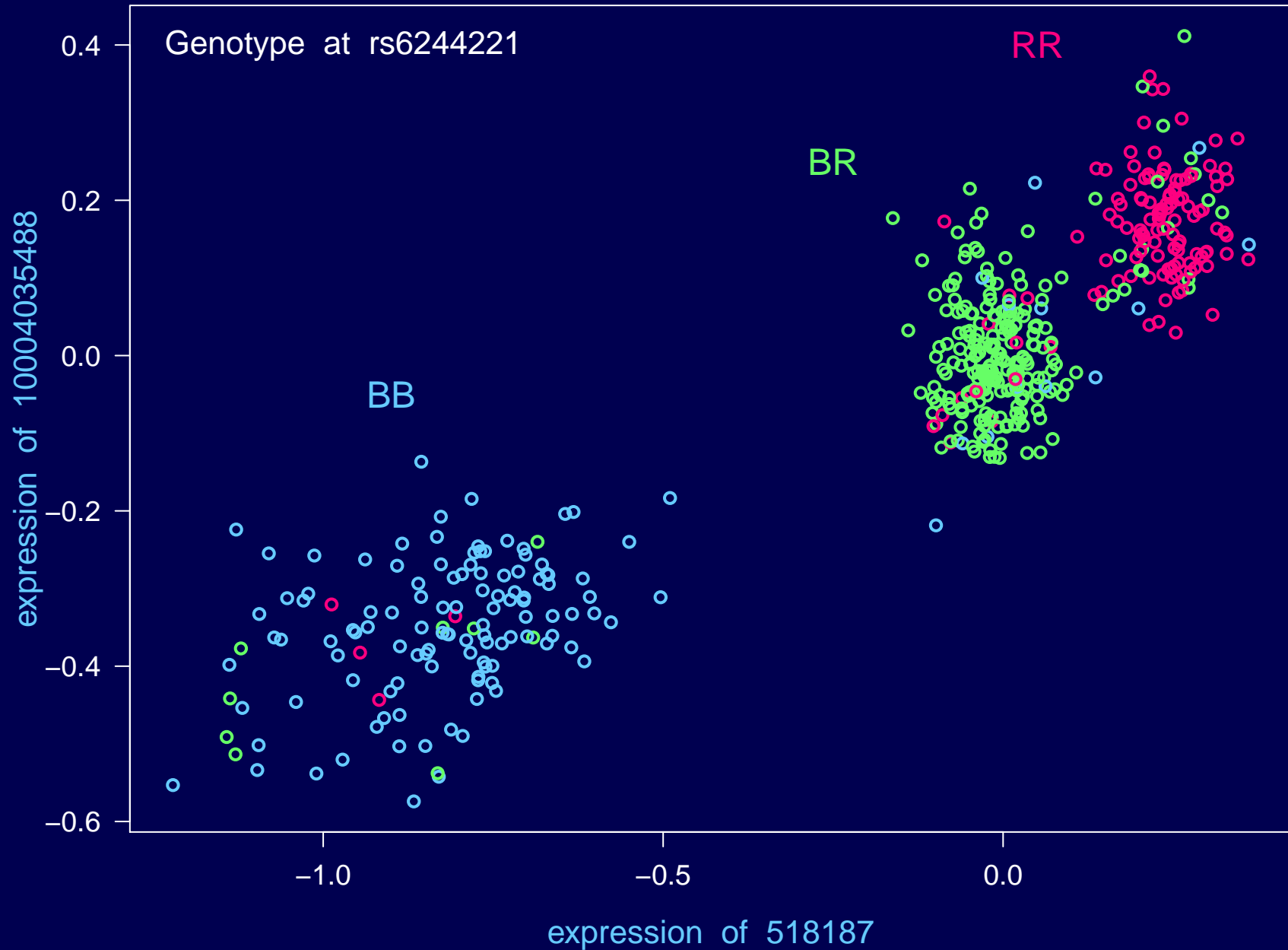
E vs G



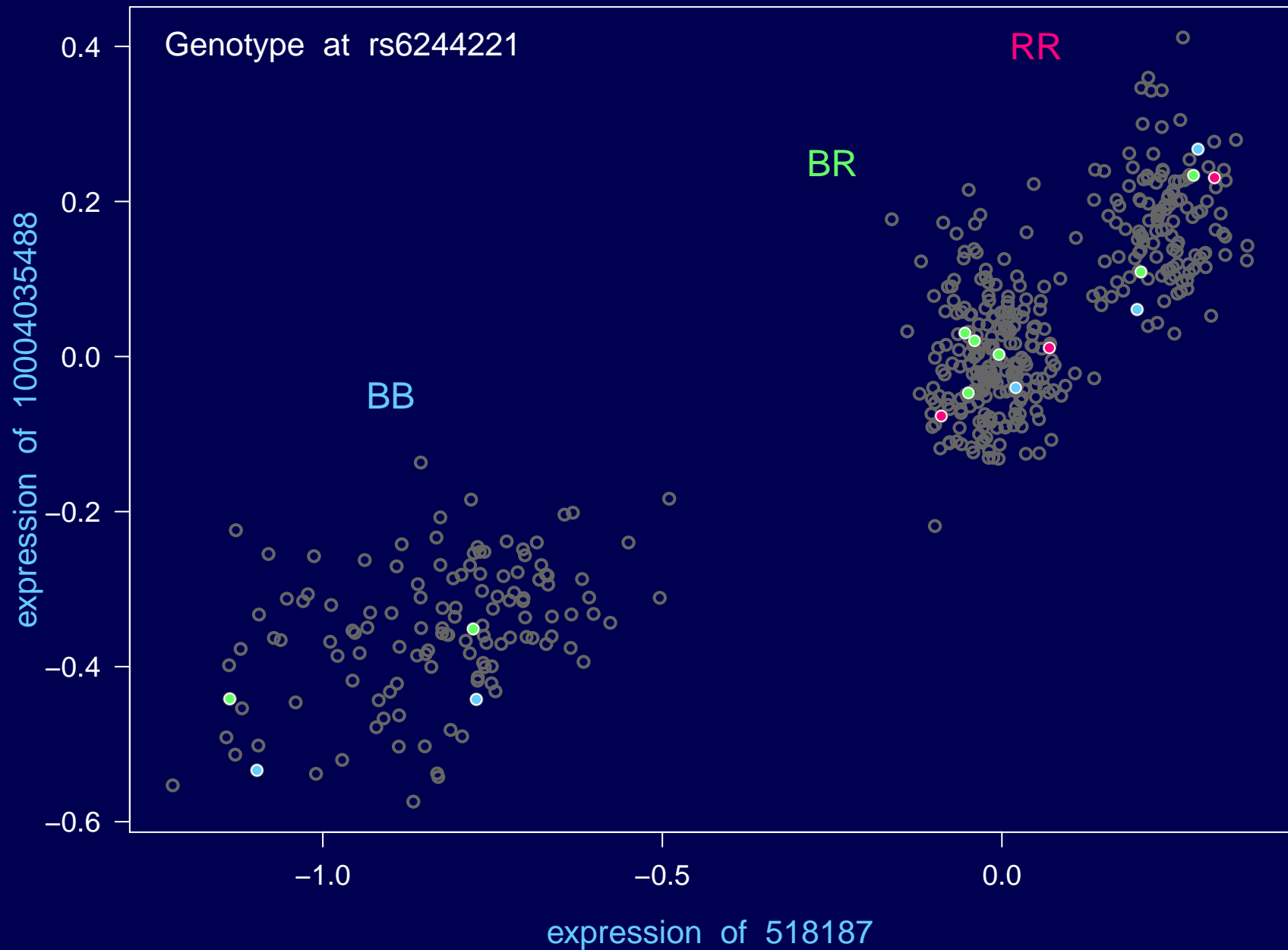
kNN classifier



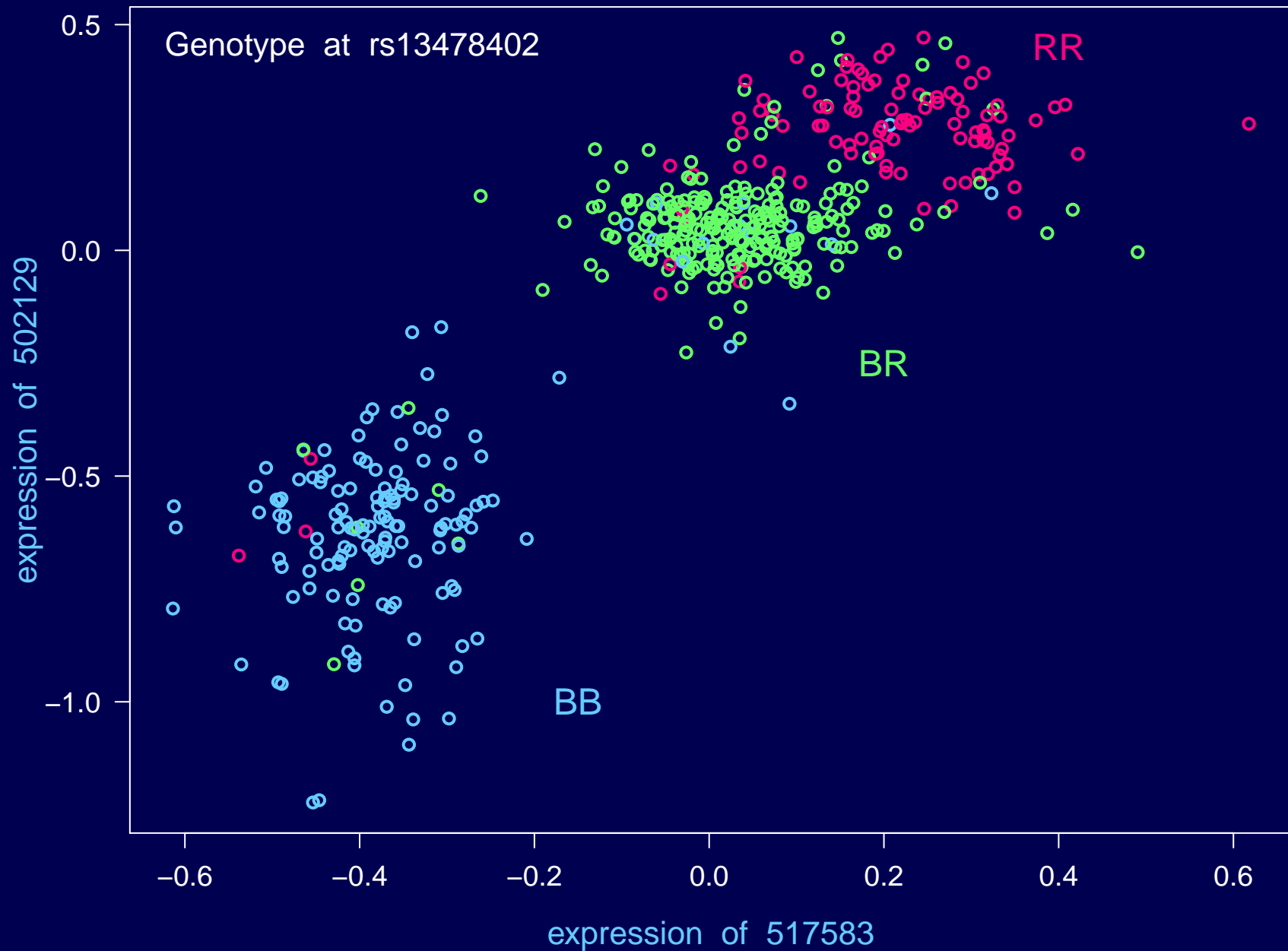
E vs G



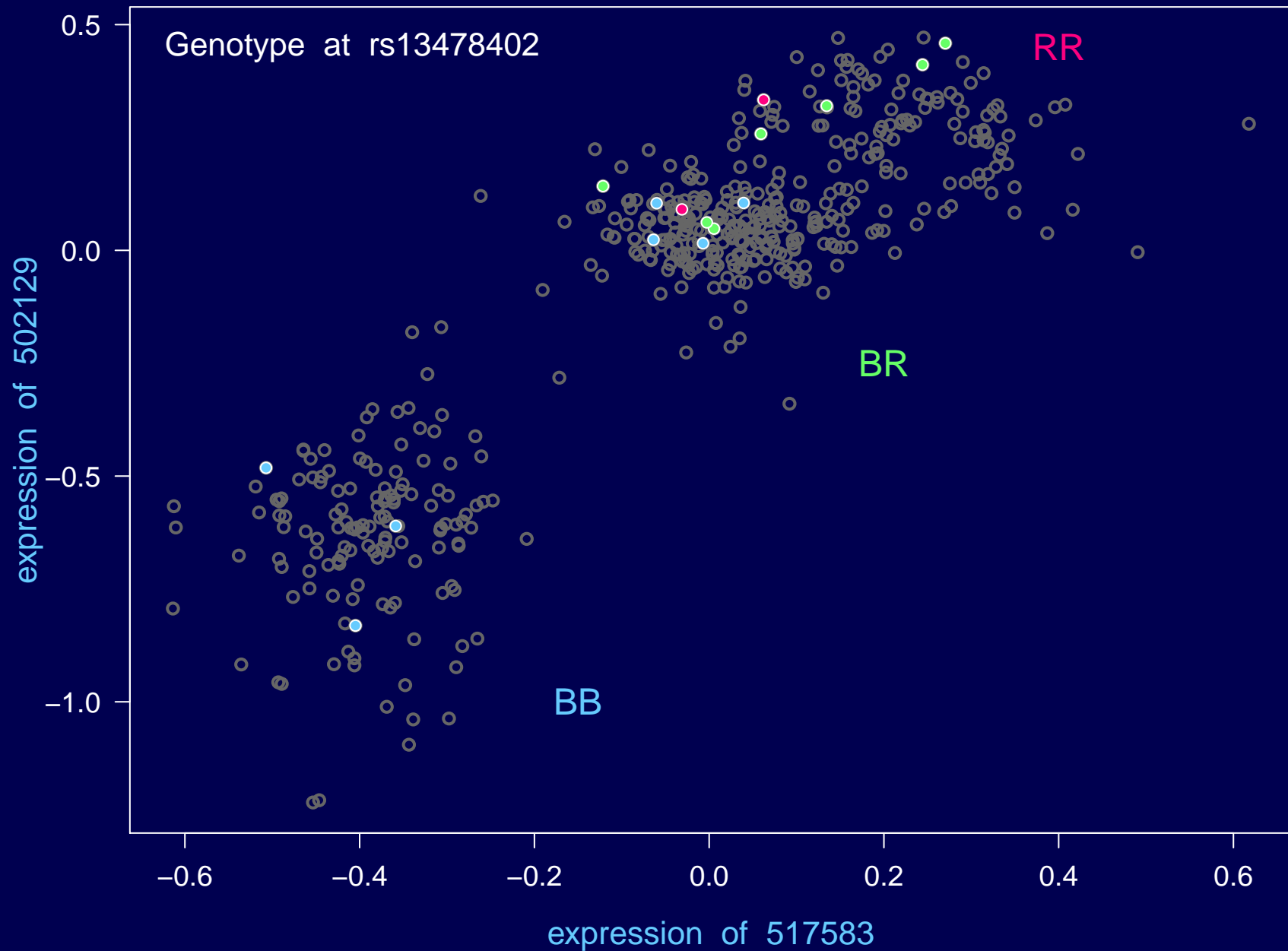
E vs G



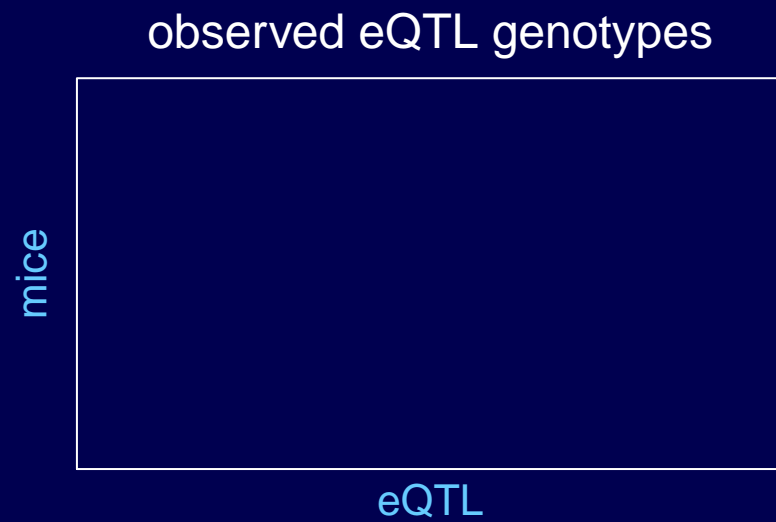
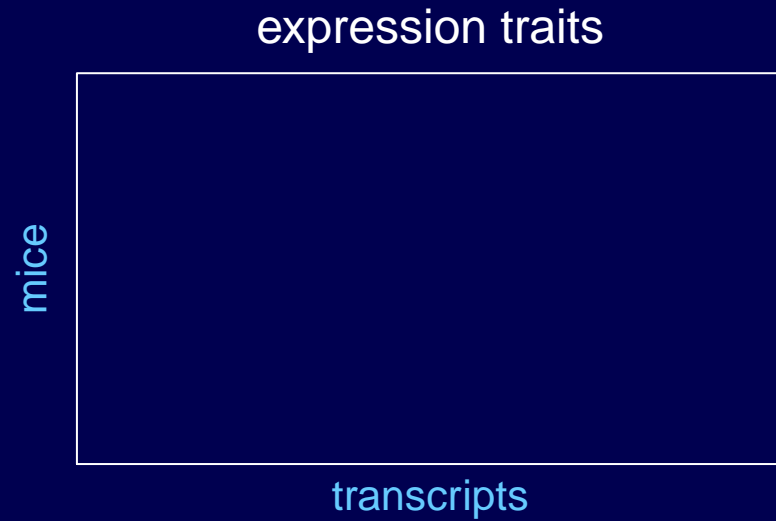
E vs G



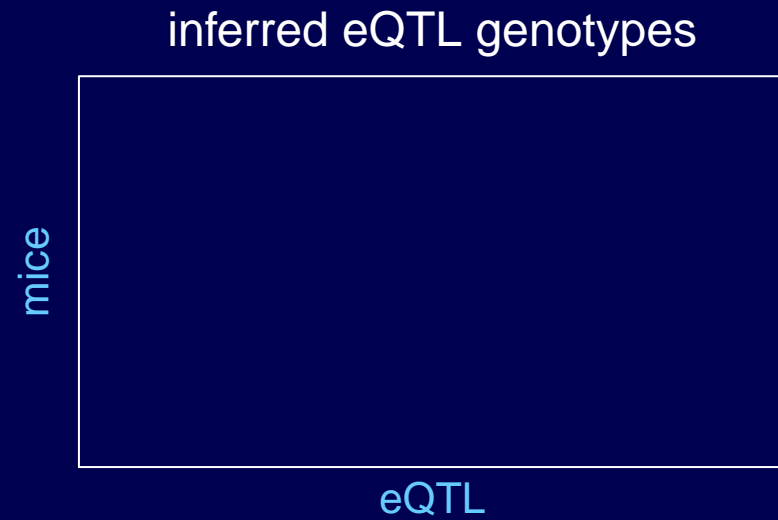
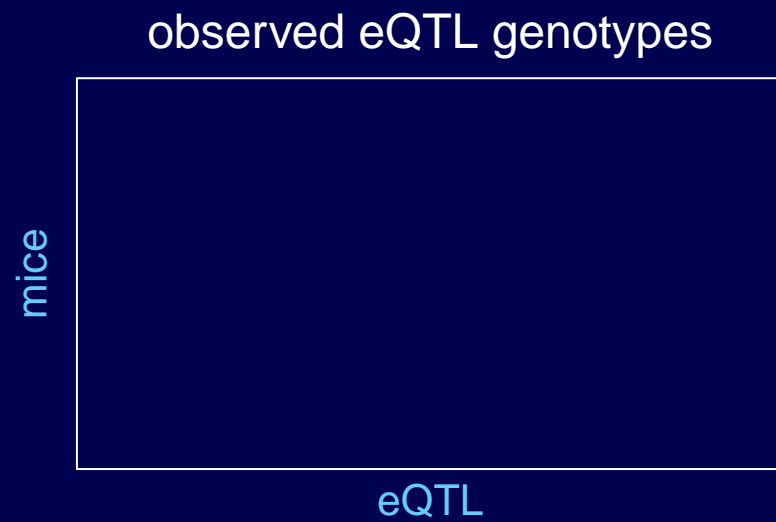
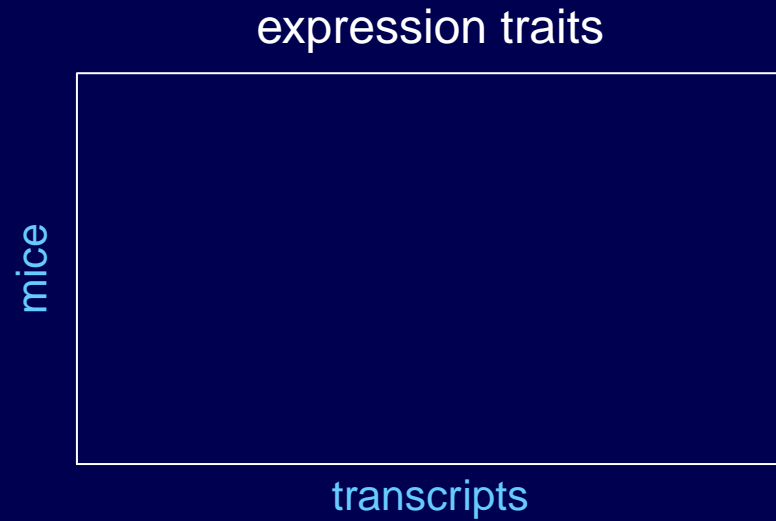
E vs G



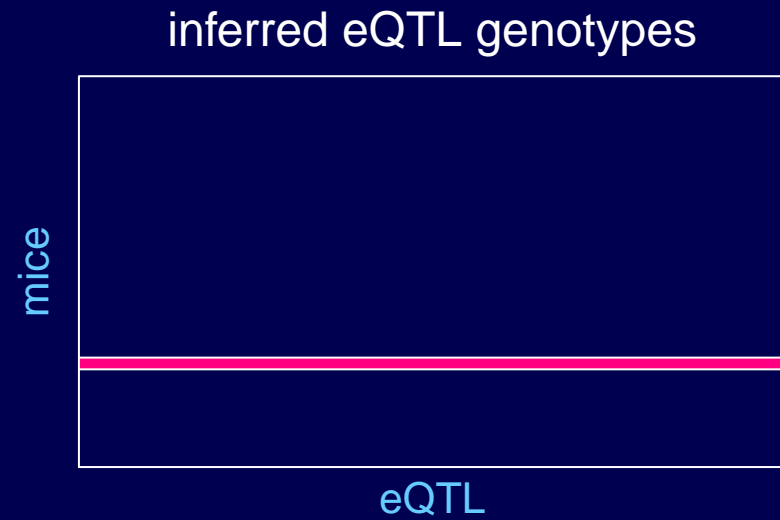
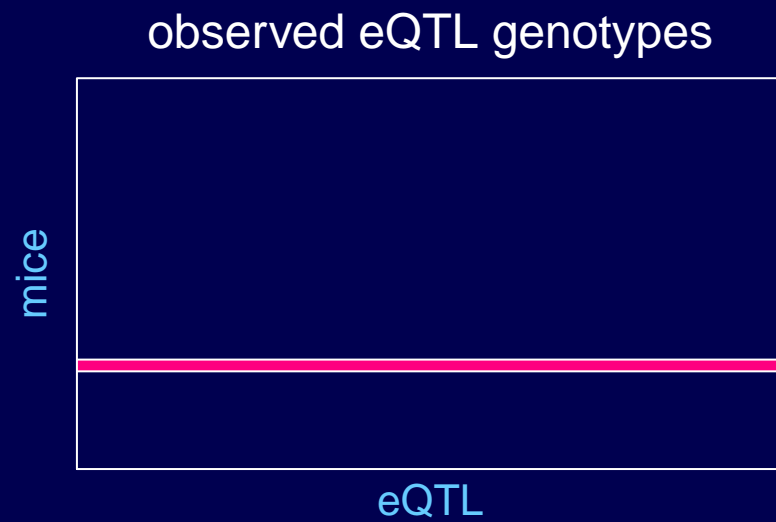
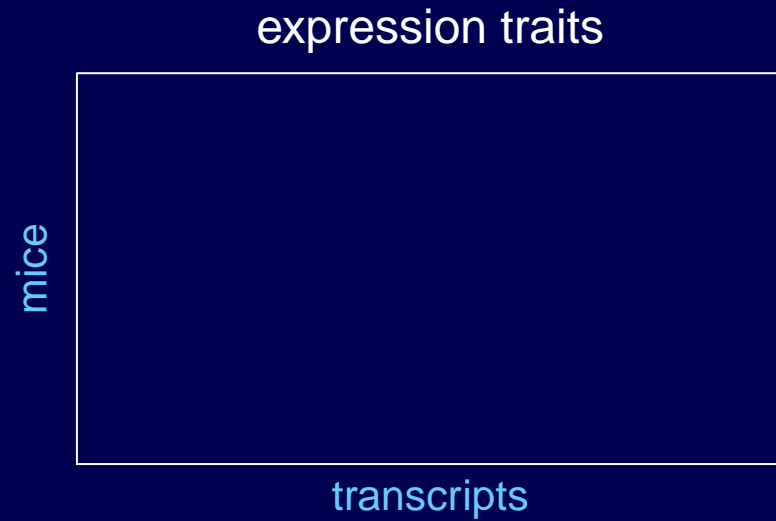
Basic scheme



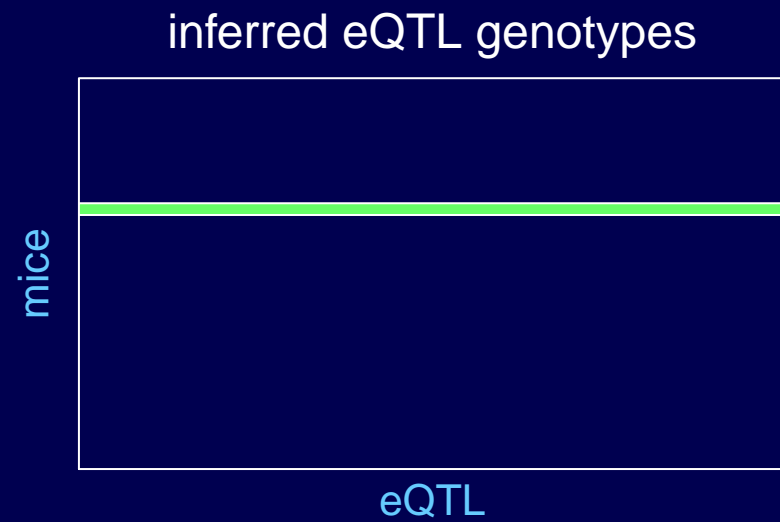
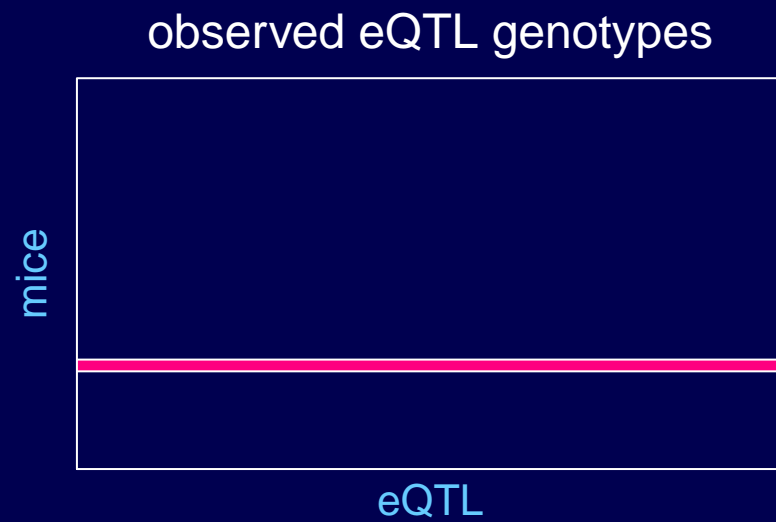
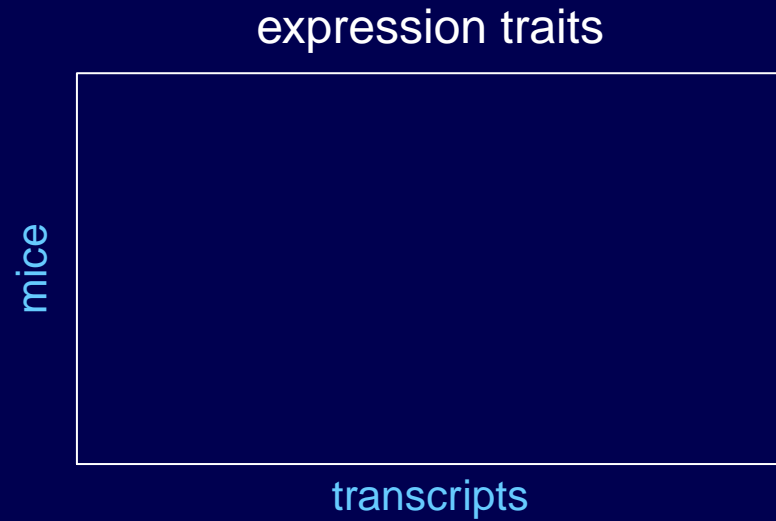
Basic scheme



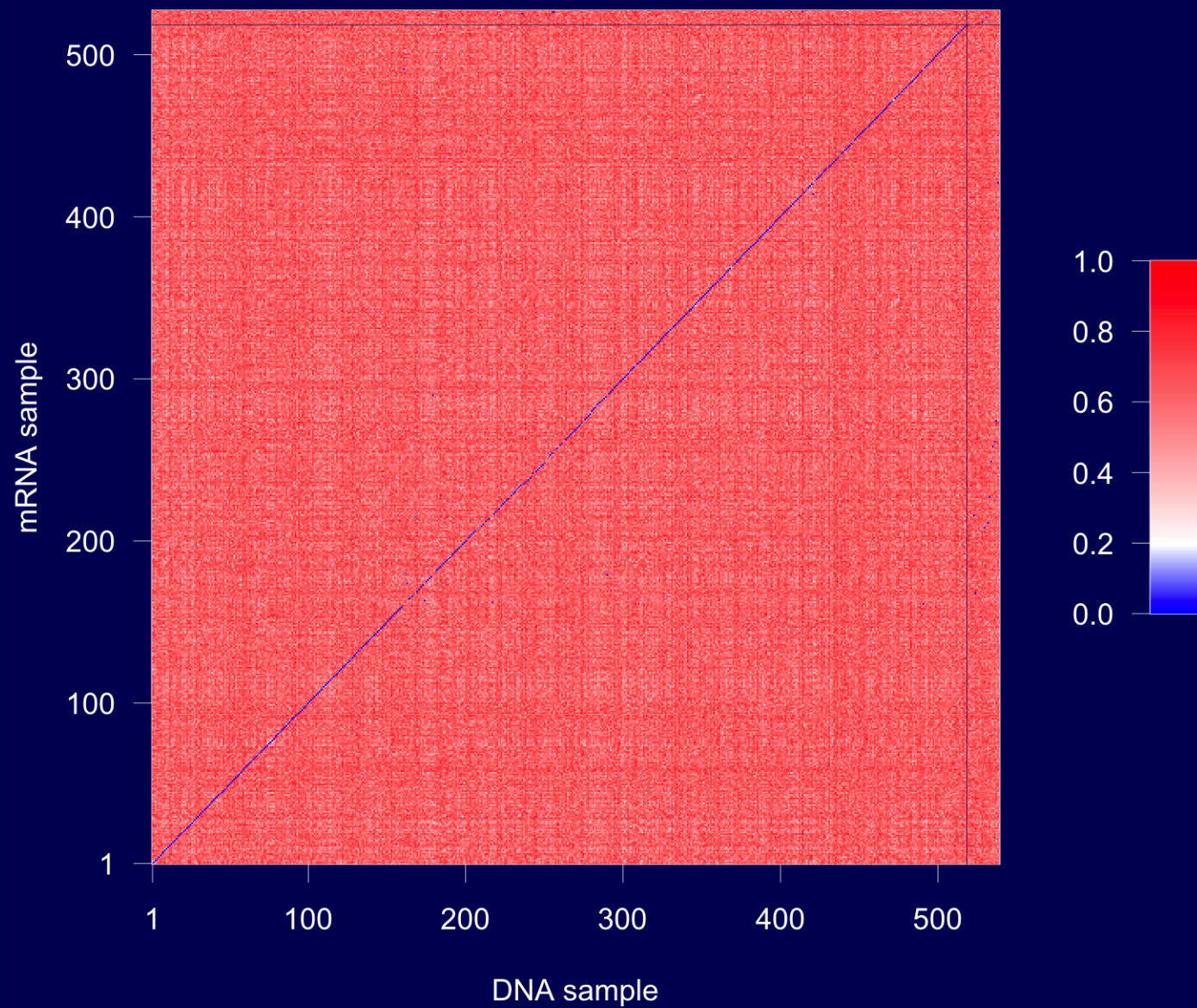
Basic scheme



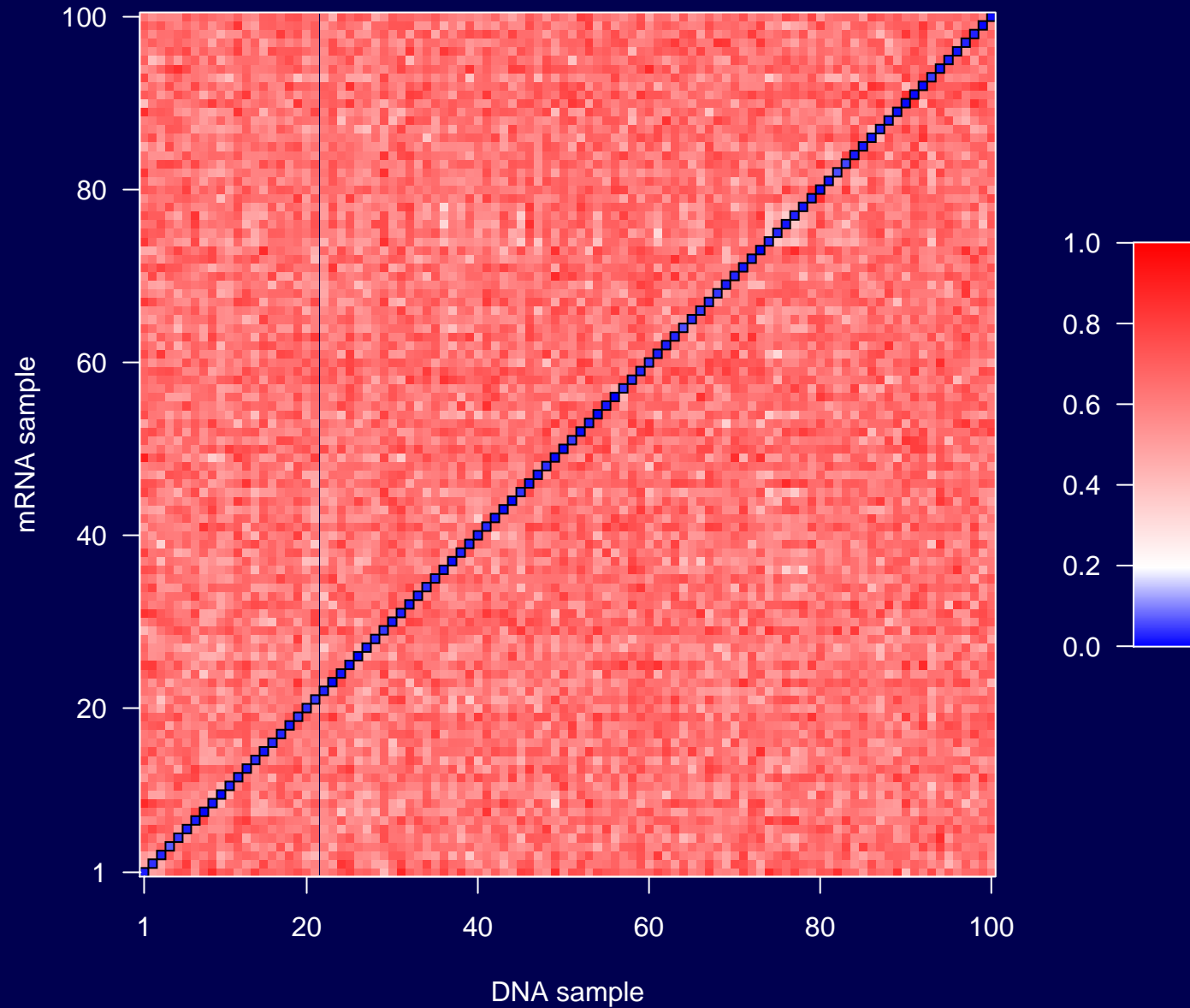
Basic scheme



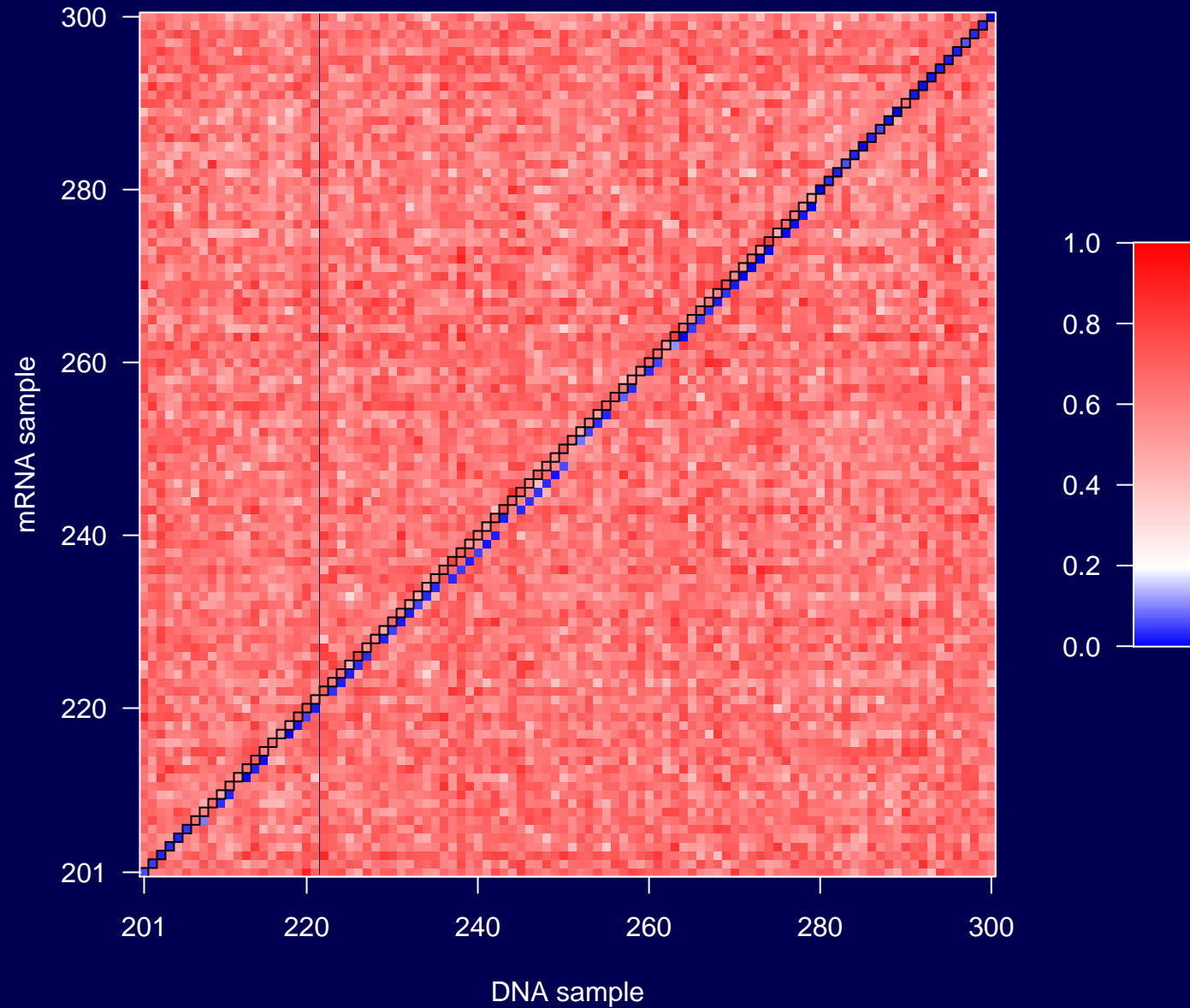
Prop'n mismatches



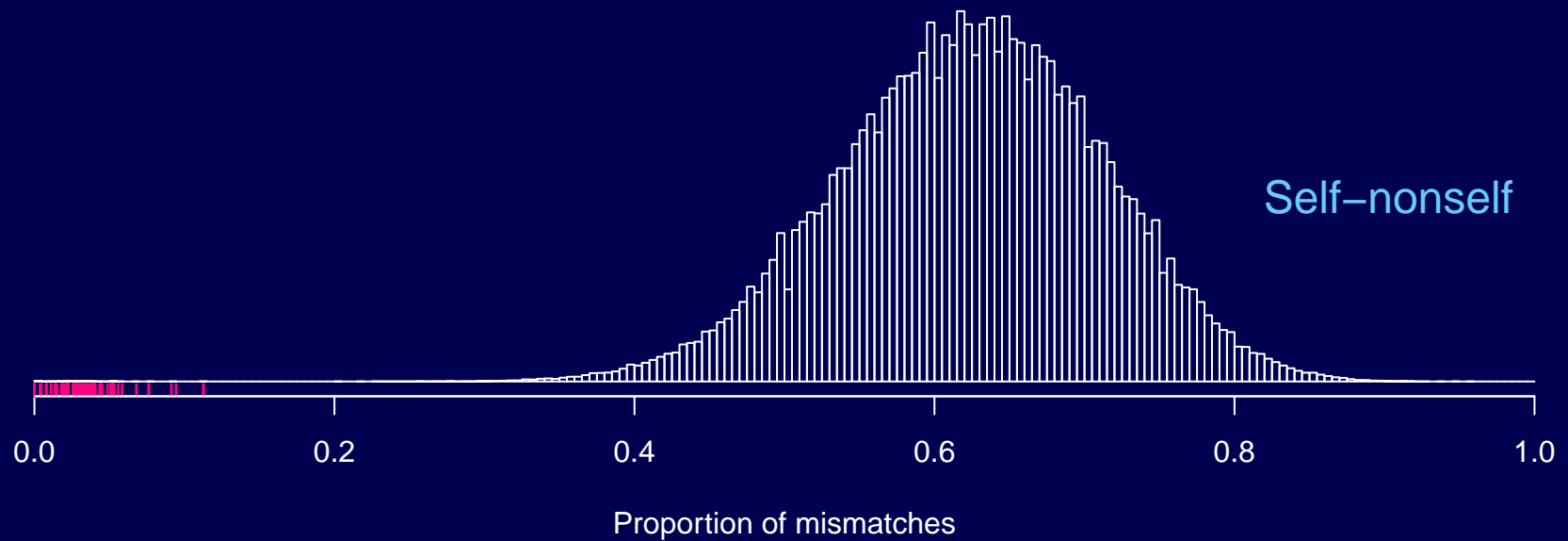
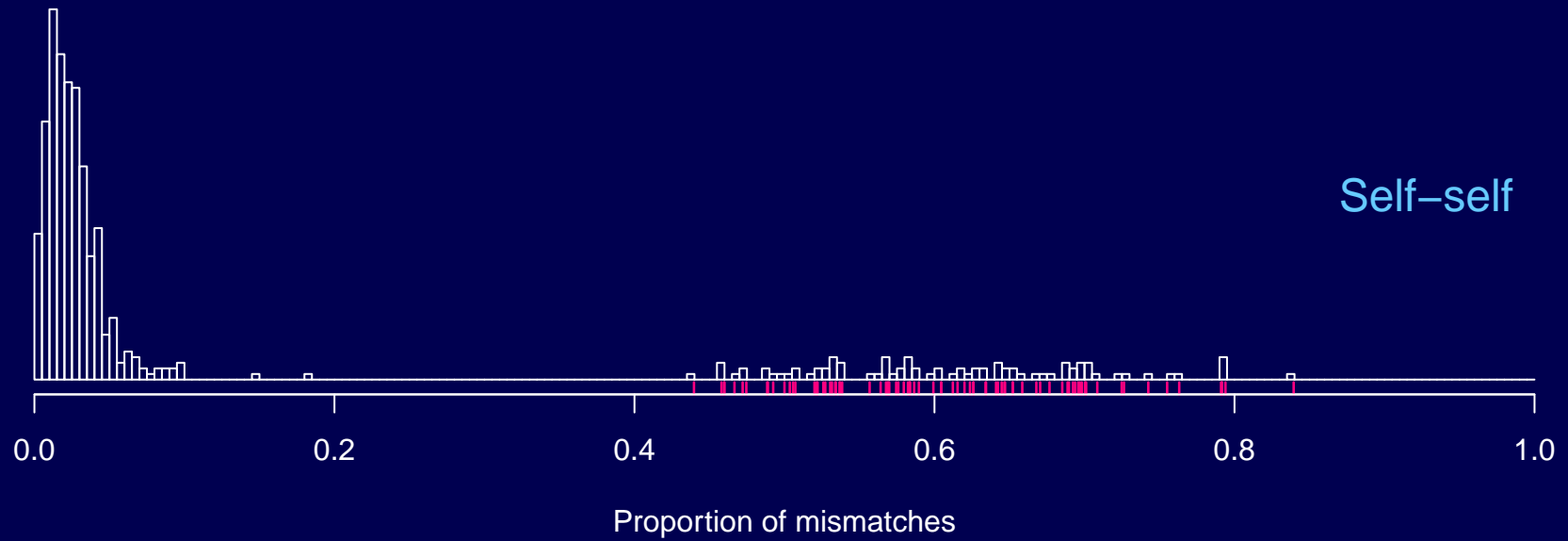
Prop'n mismatches



Prop'n mismatches

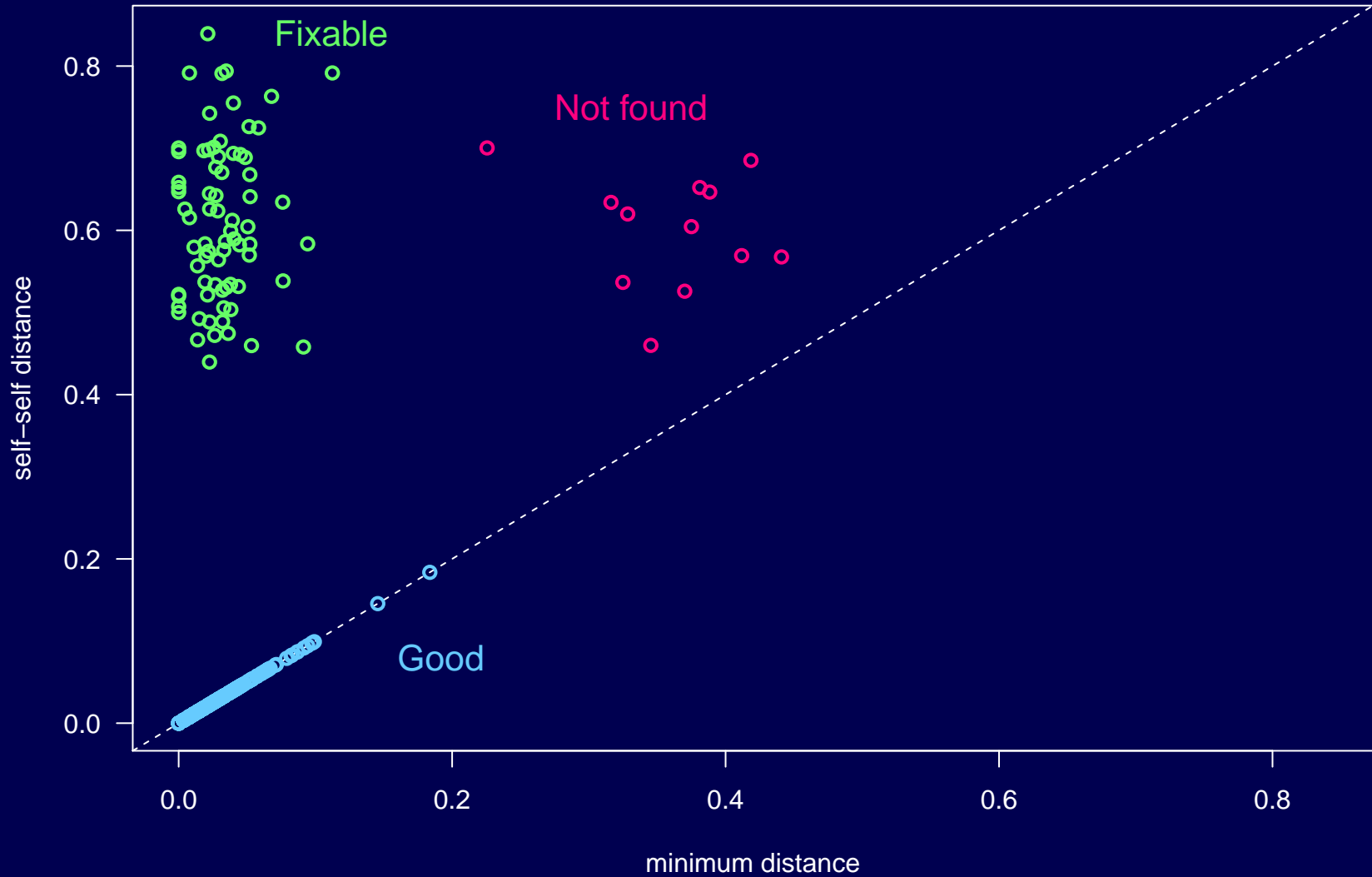


Prop'n mismatches



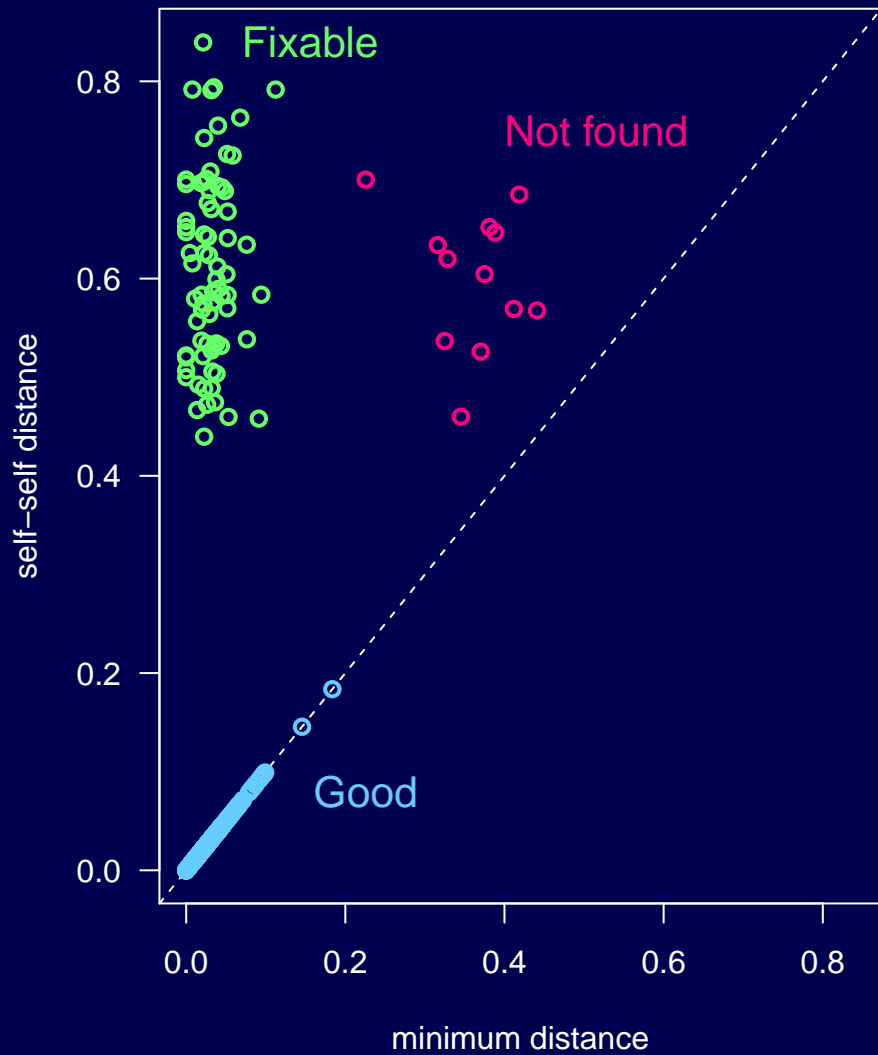
Decisions

Self vs best

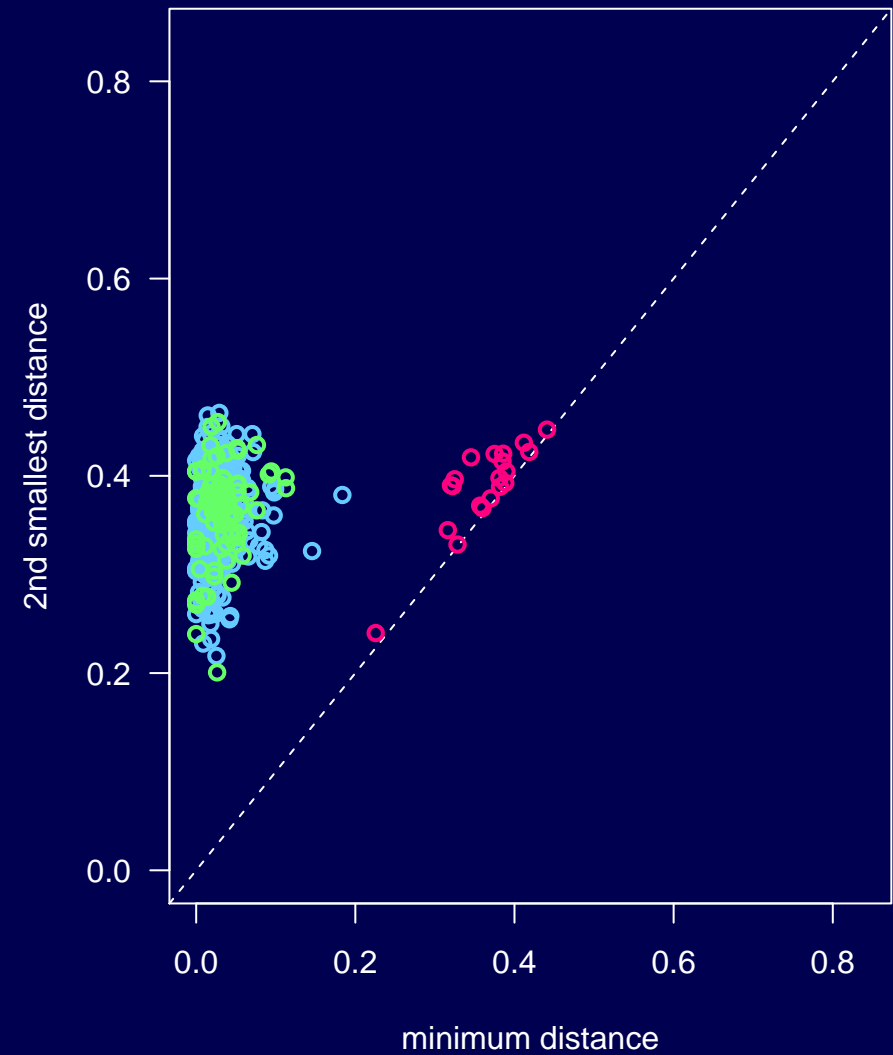


Decisions

Self vs best



Next-best vs best

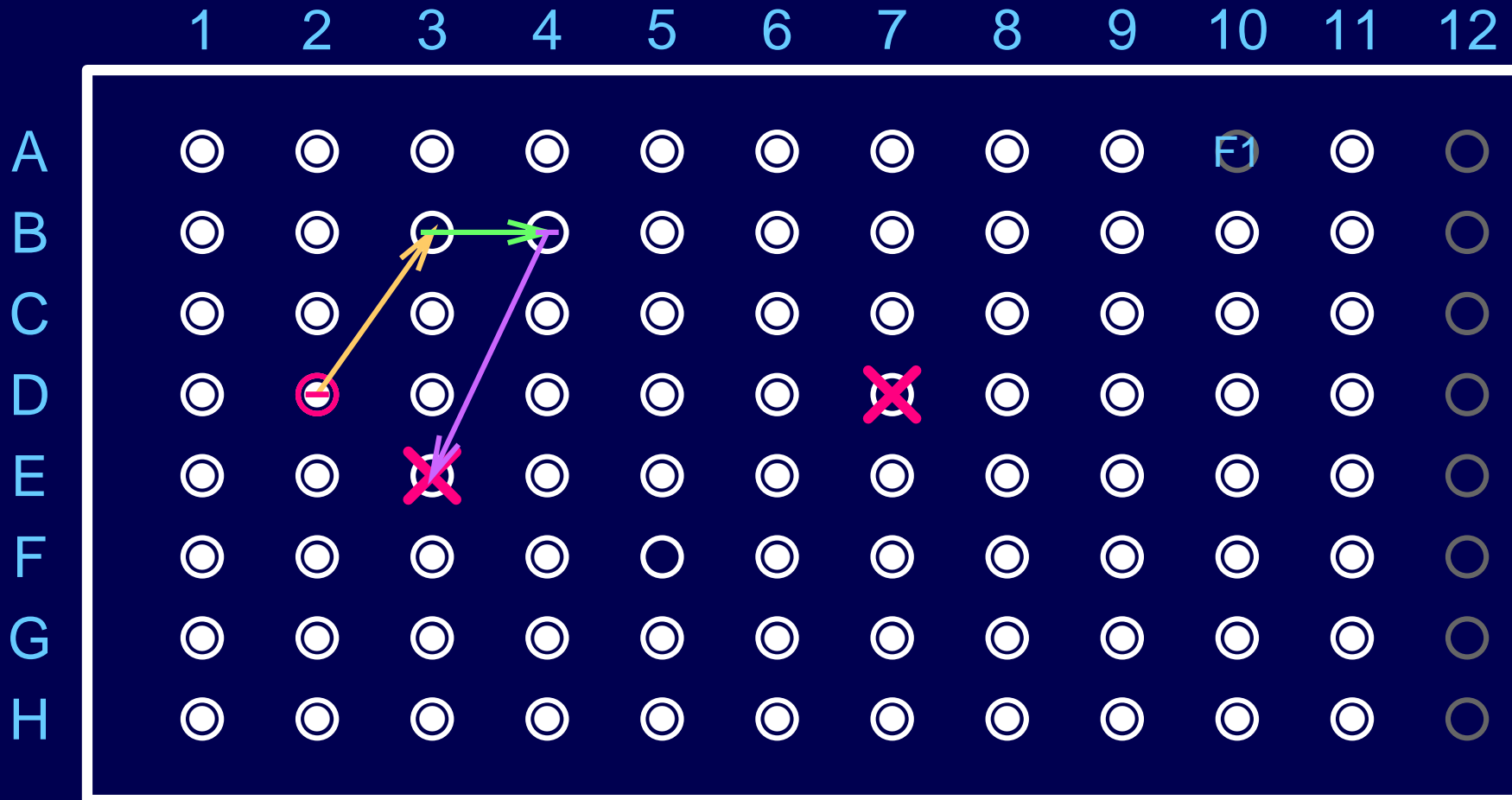


Genotype mix-ups



Plate 1631

1631



Plates 1632 and 1630

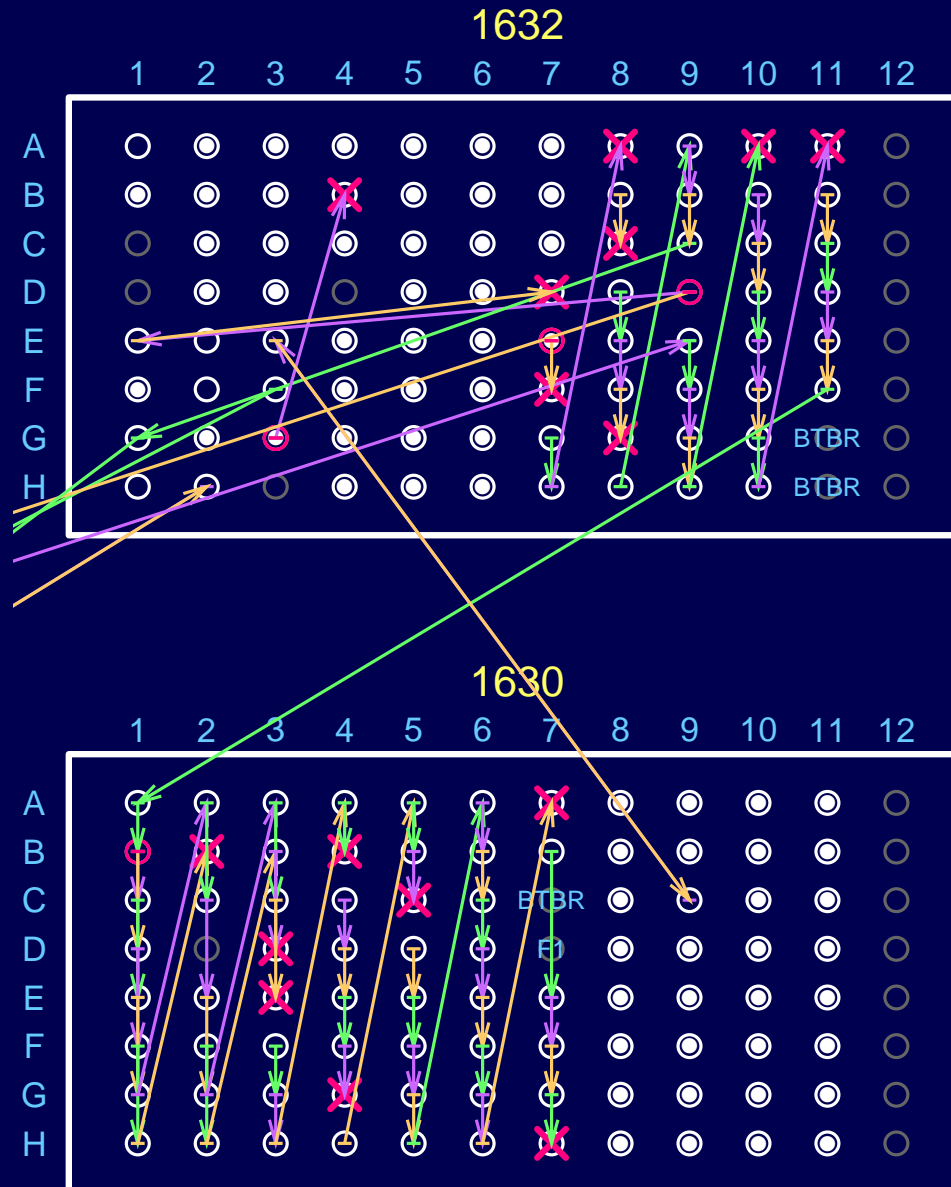
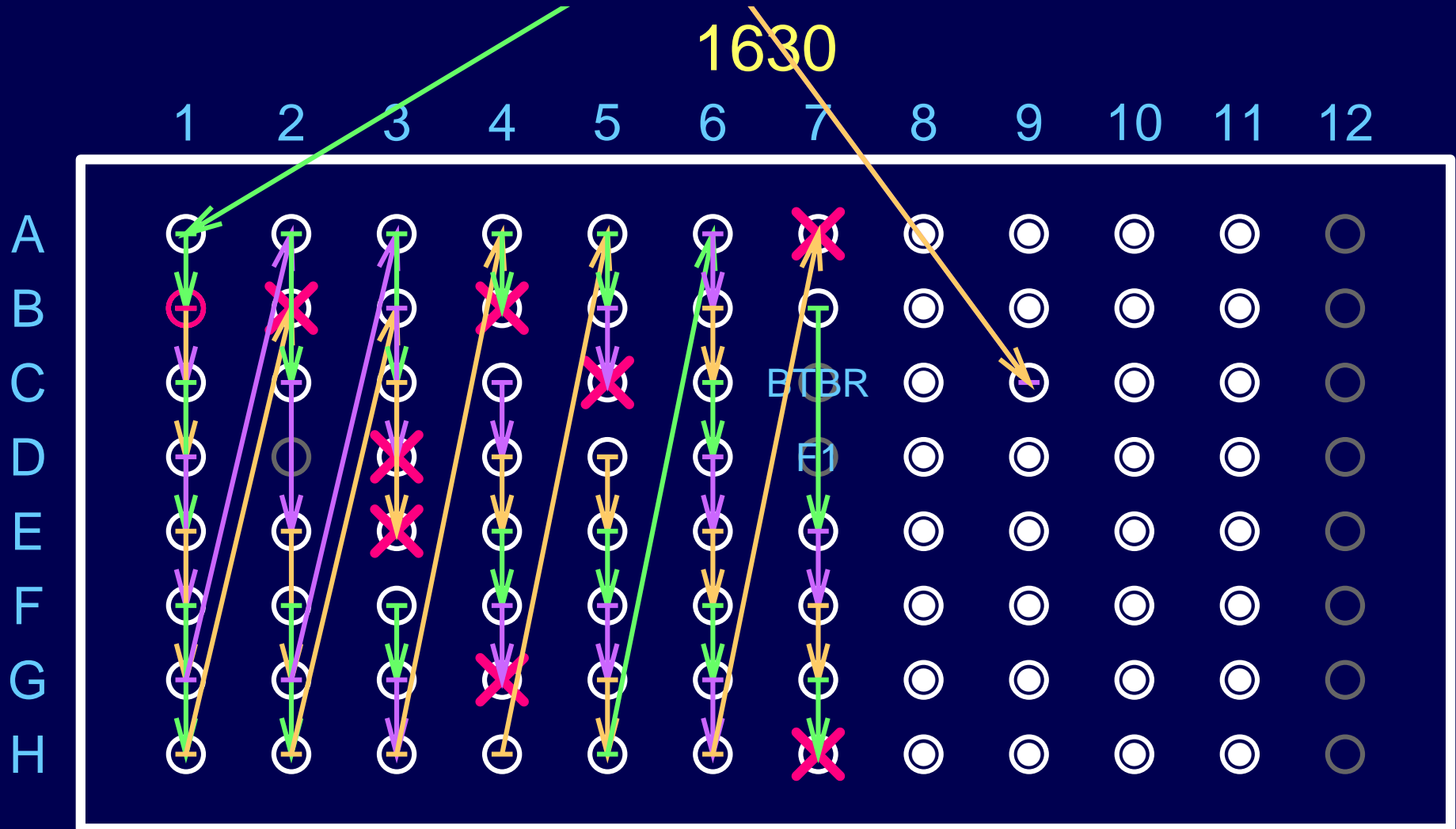


Plate 1630



E vs E

expression in islet

mice



transcripts

expression in liver

mice

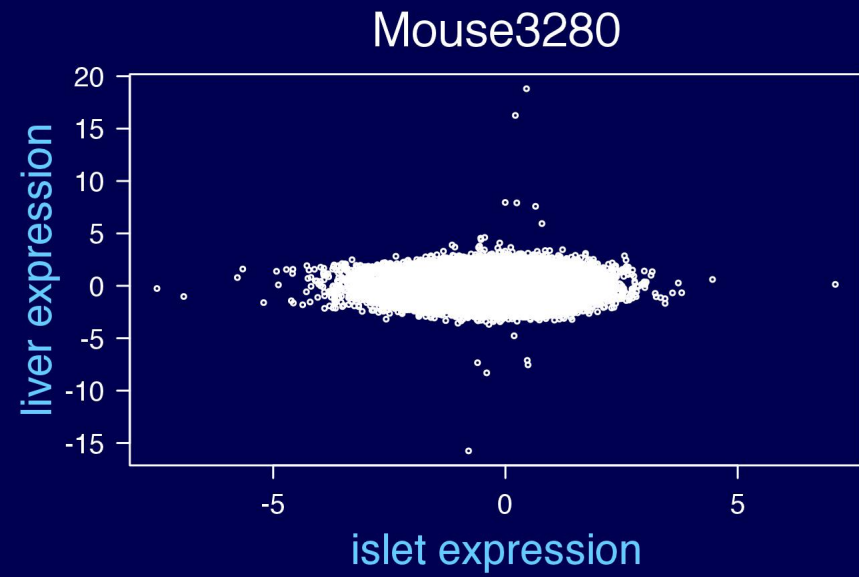


transcripts

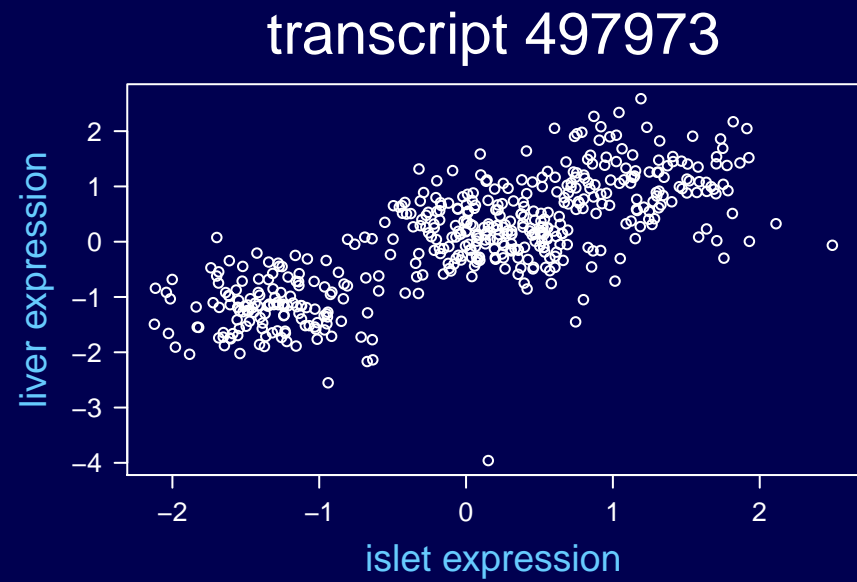
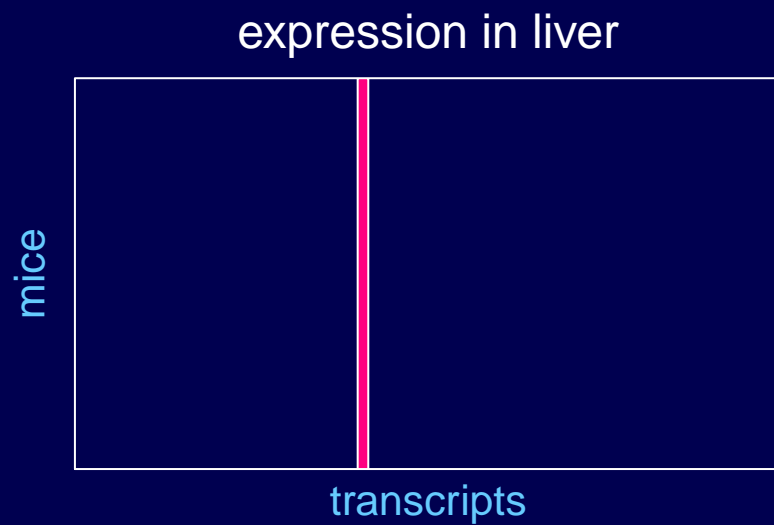
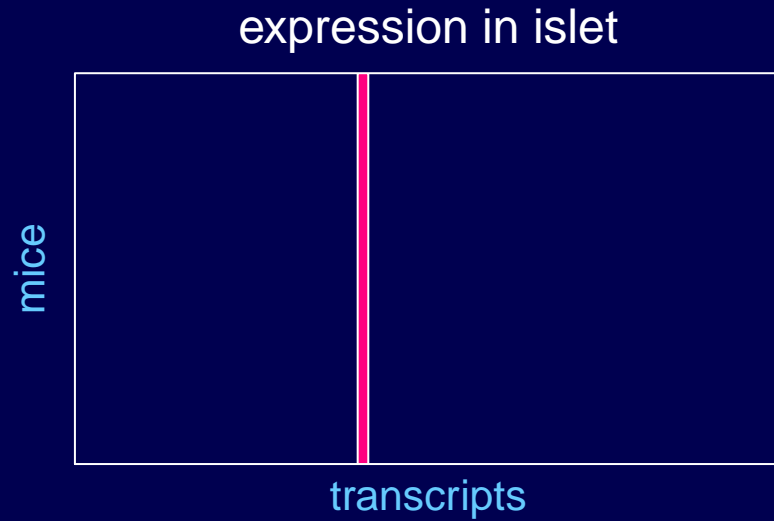
E vs E



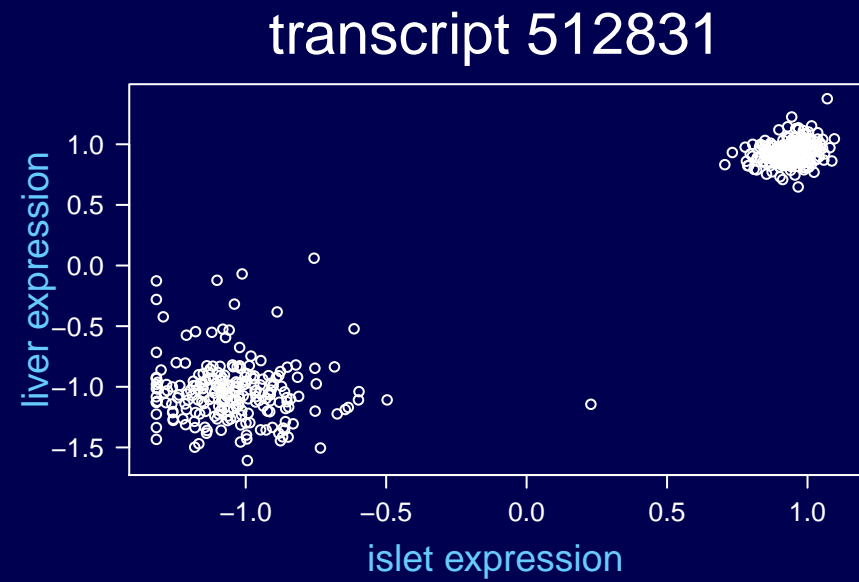
E vs E



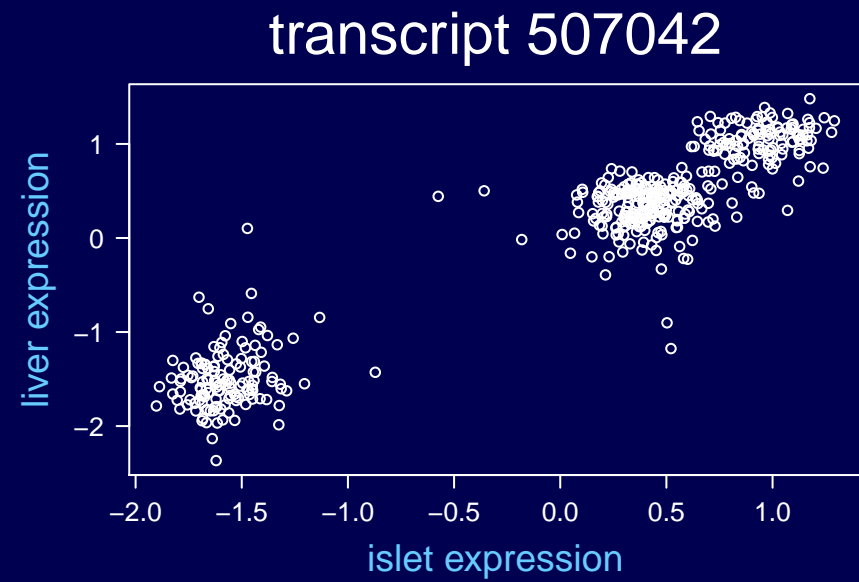
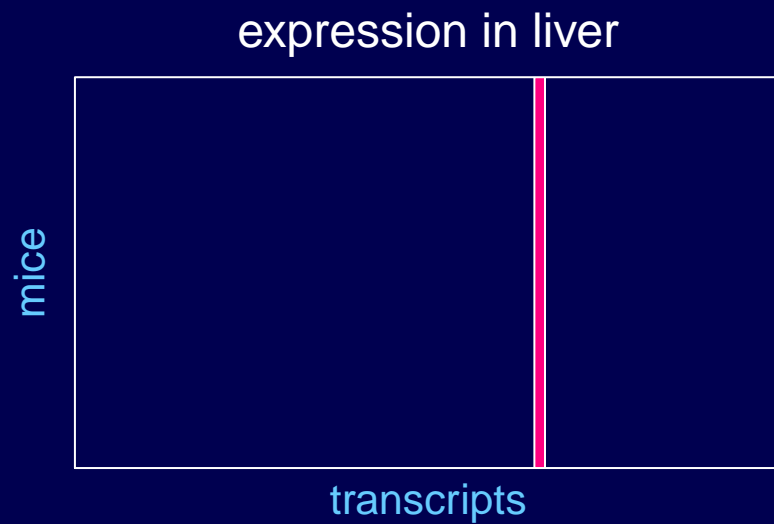
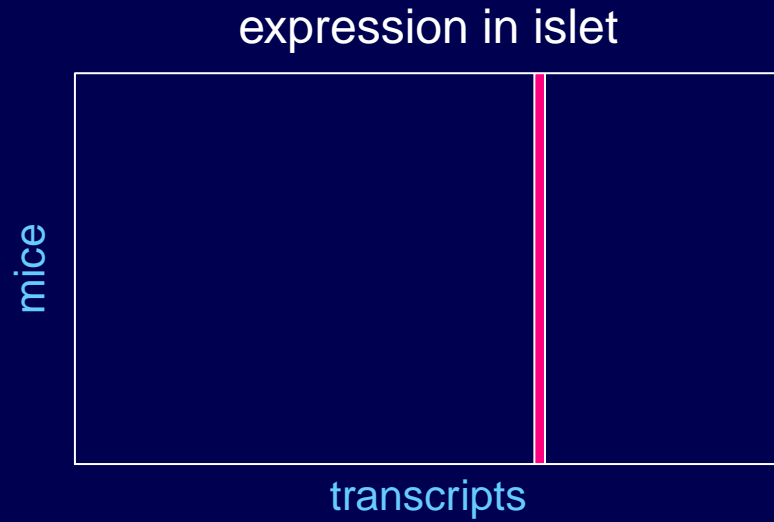
E vs E



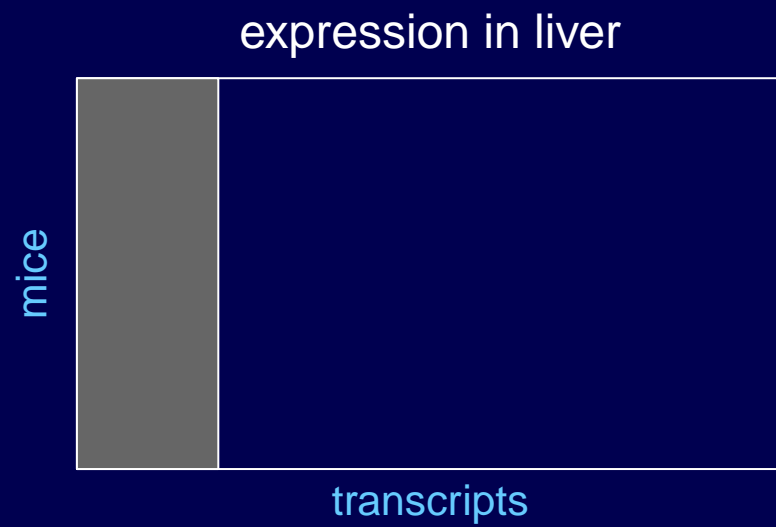
E vs E



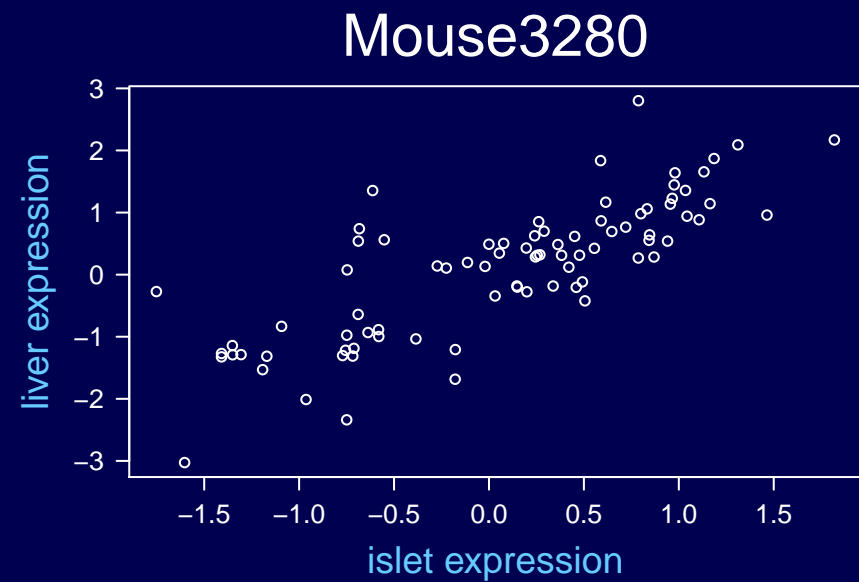
E vs E



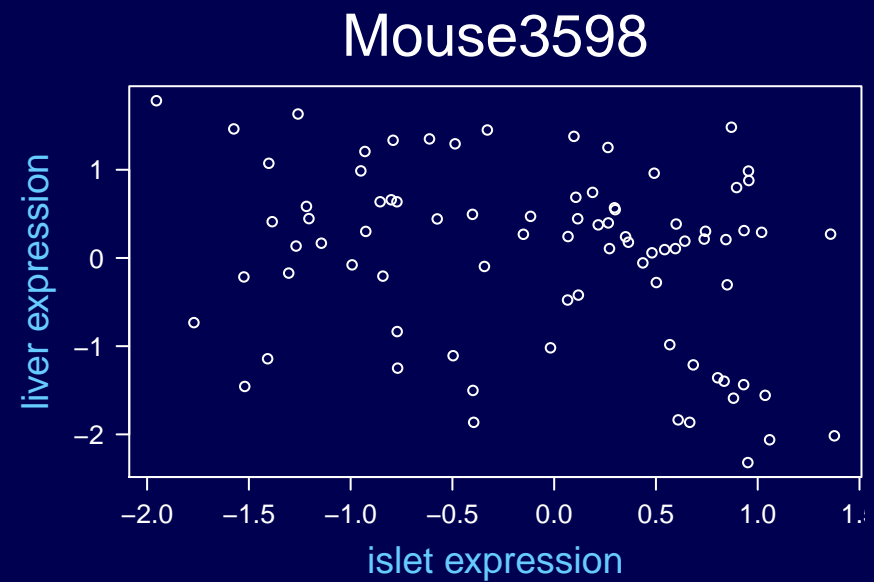
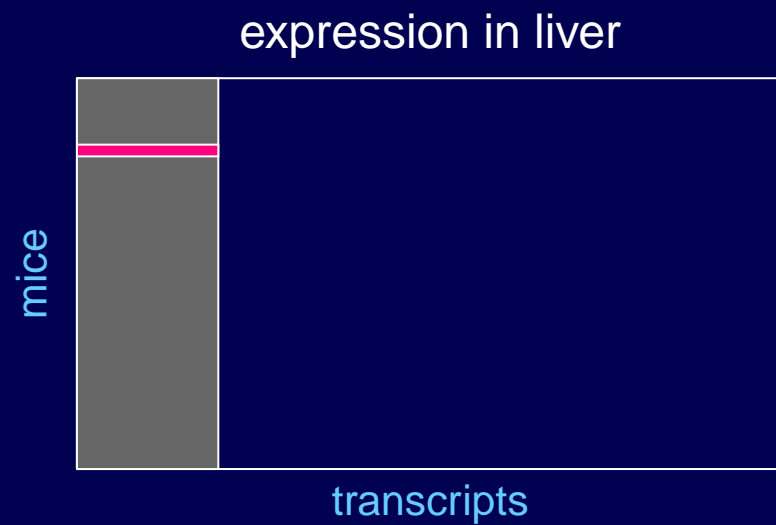
E vs E



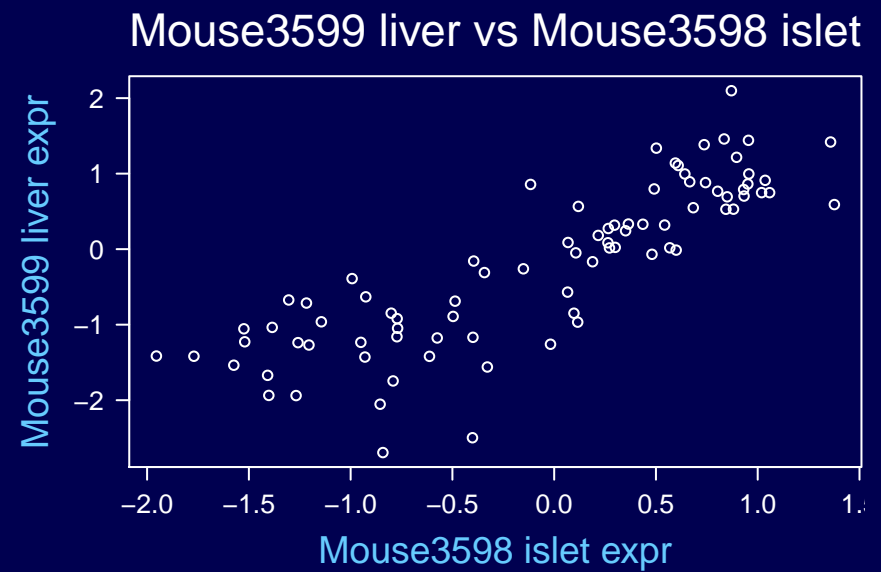
E vs E



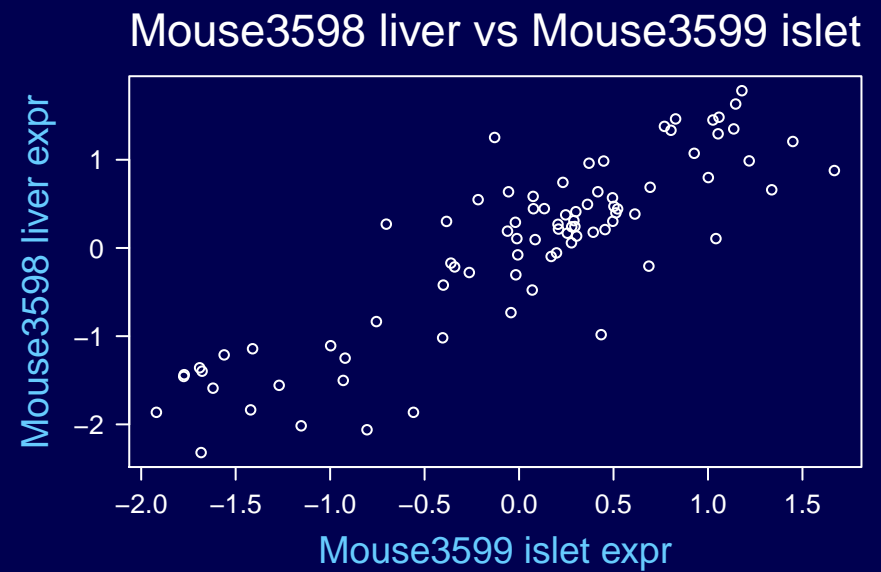
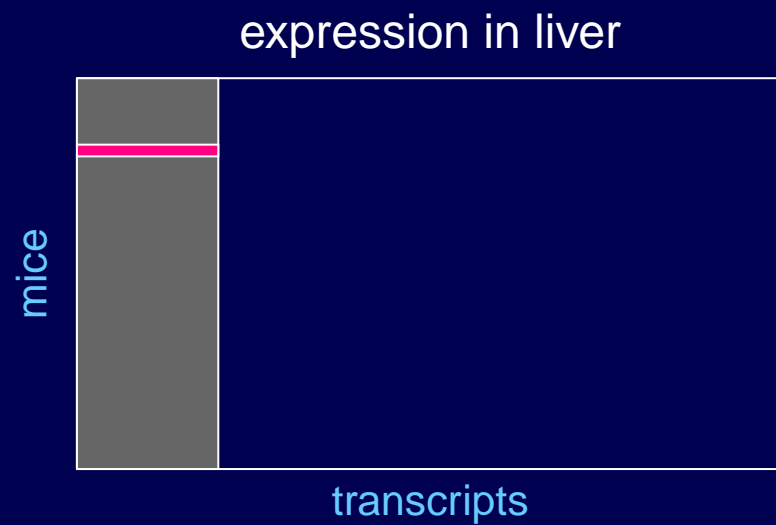
E vs E



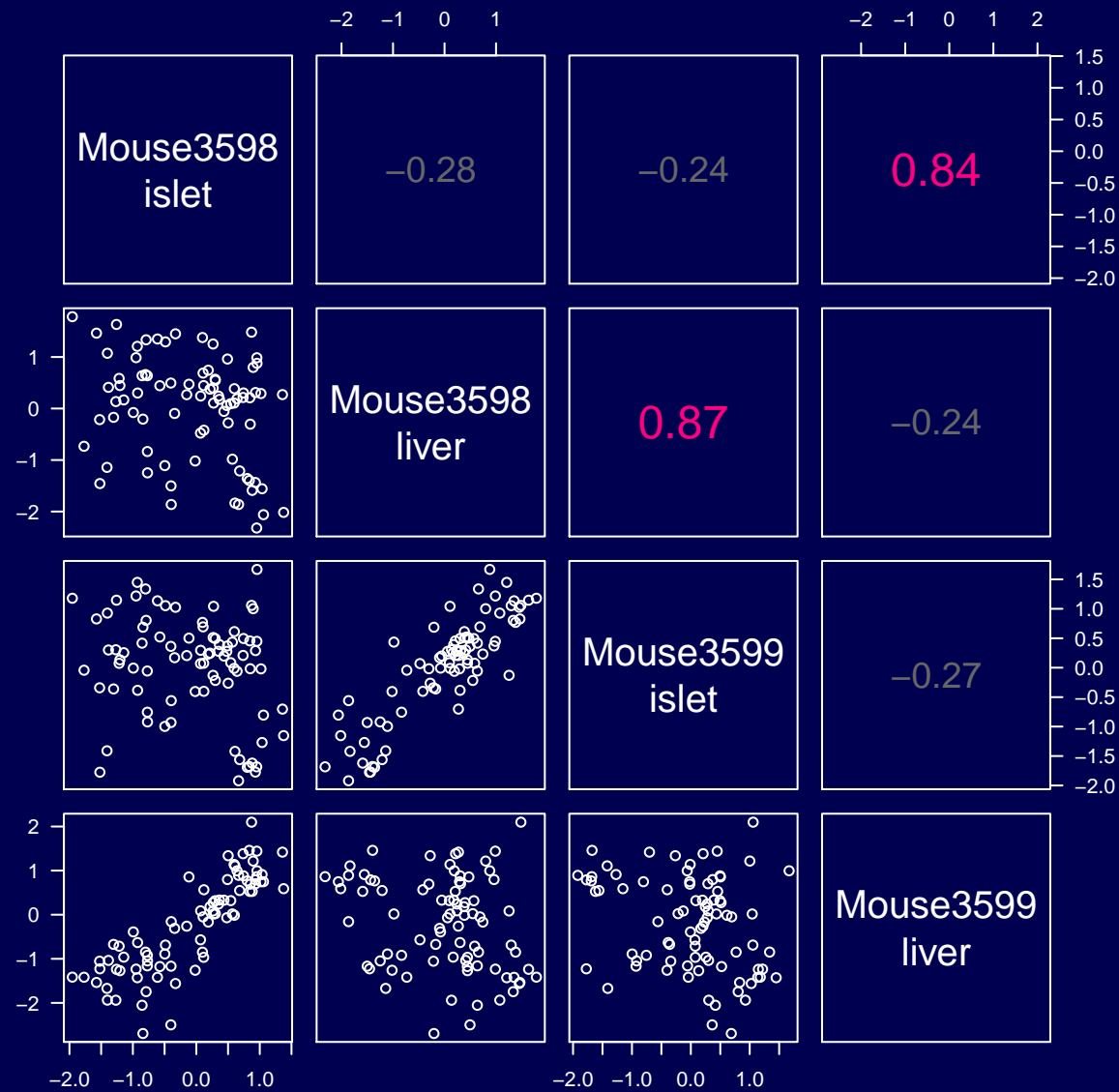
E vs E



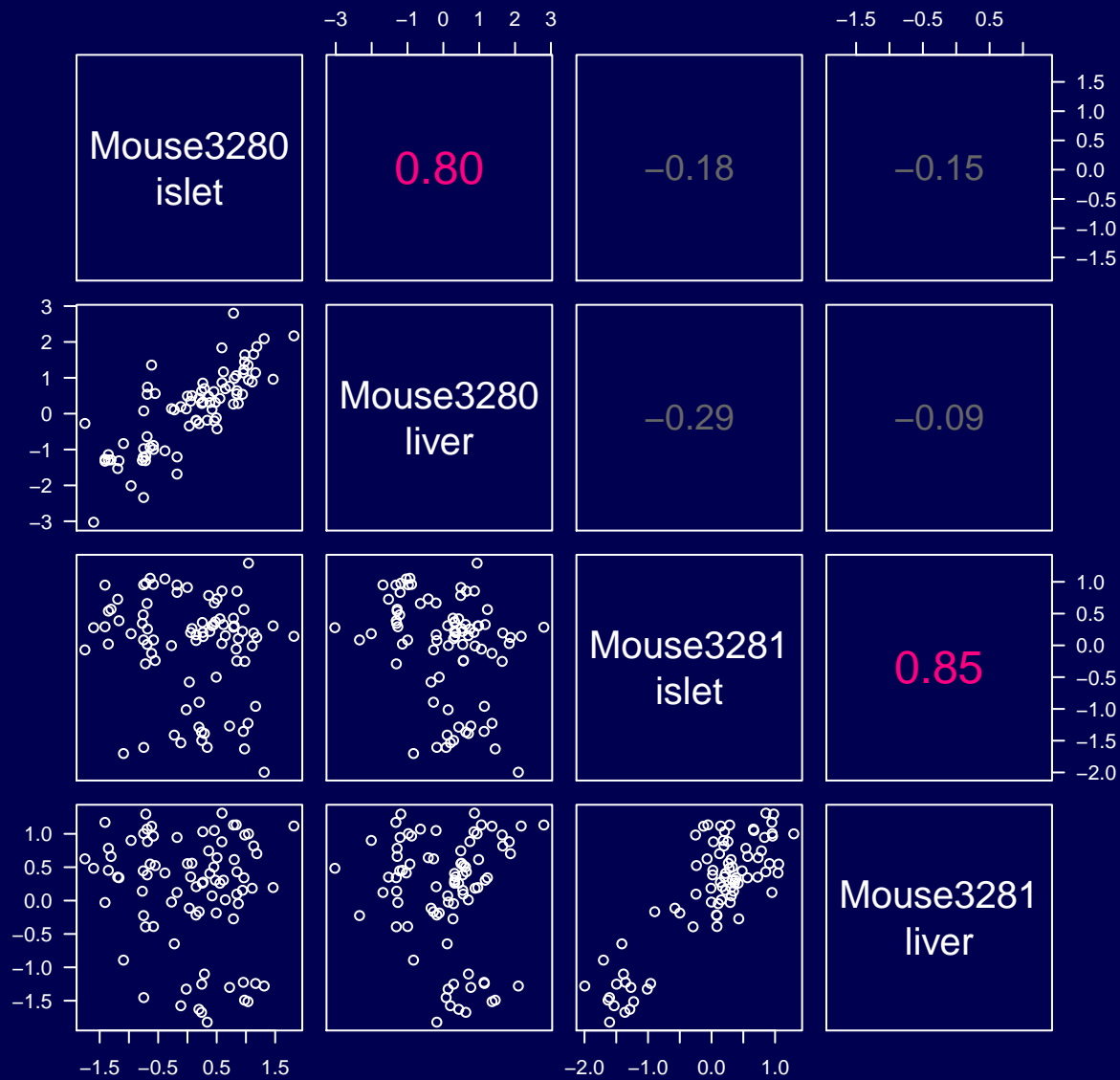
E vs E



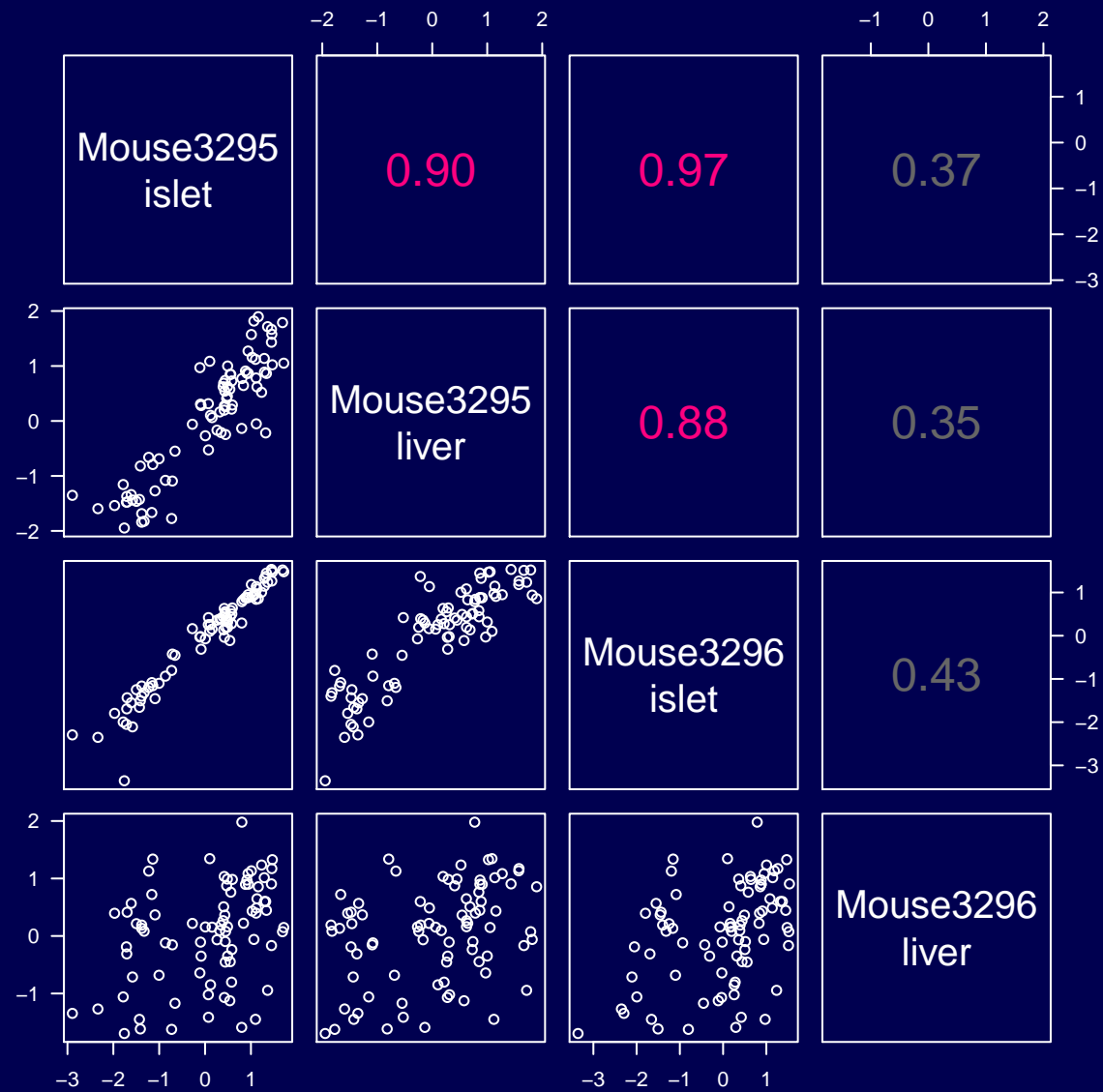
E vs E



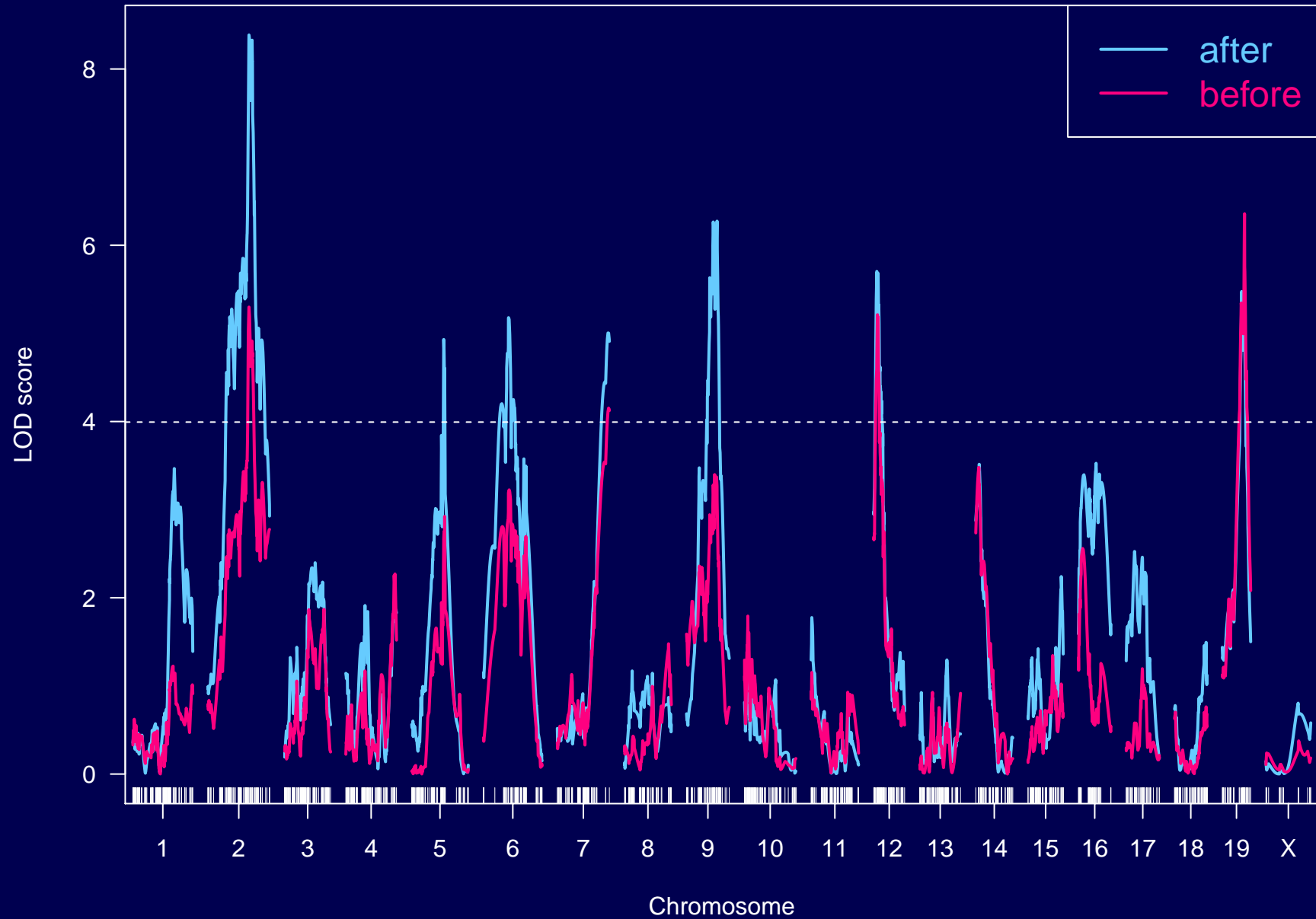
E vs E



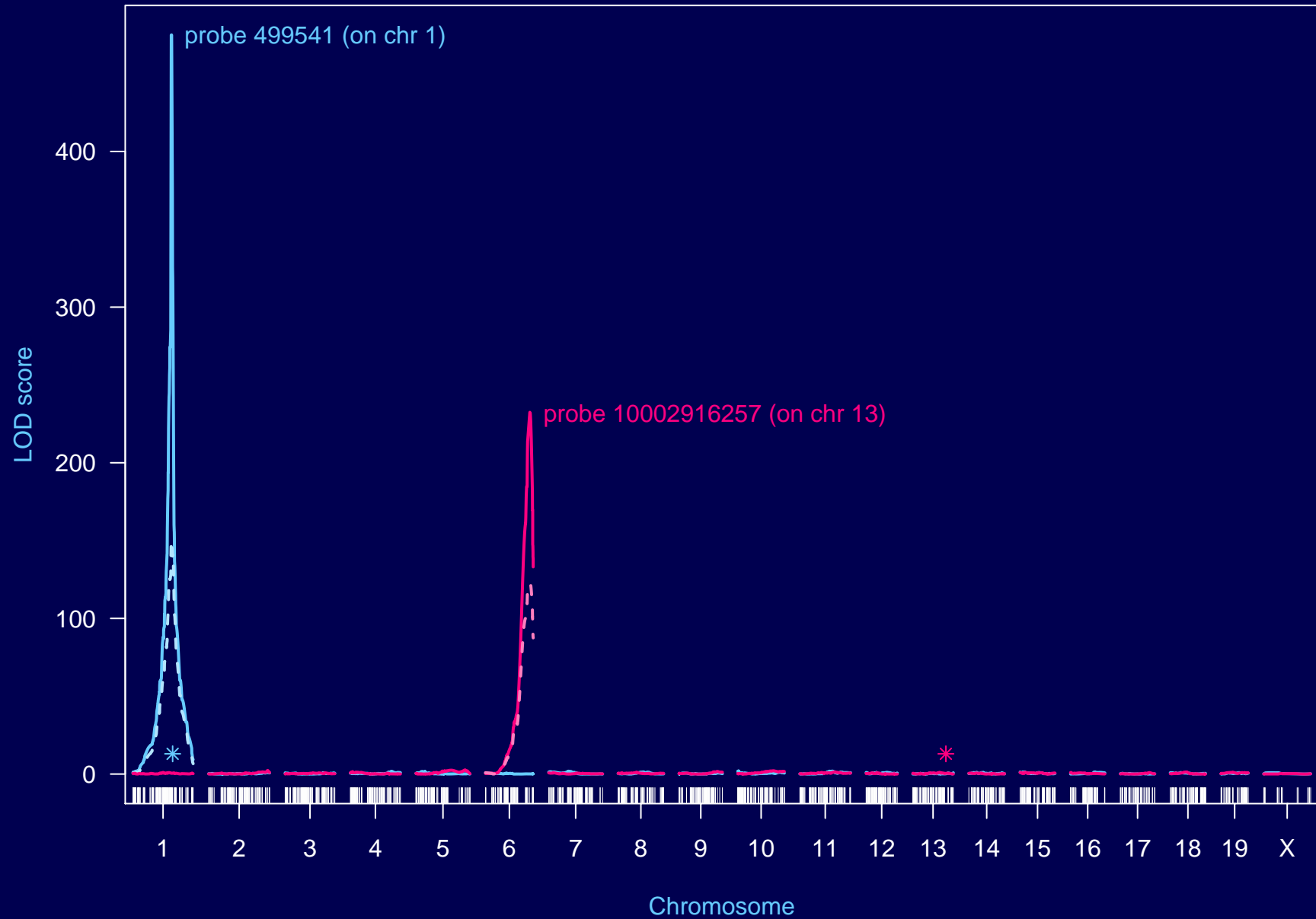
E vs E



Insulin QTL



Strong eQTL



Summary

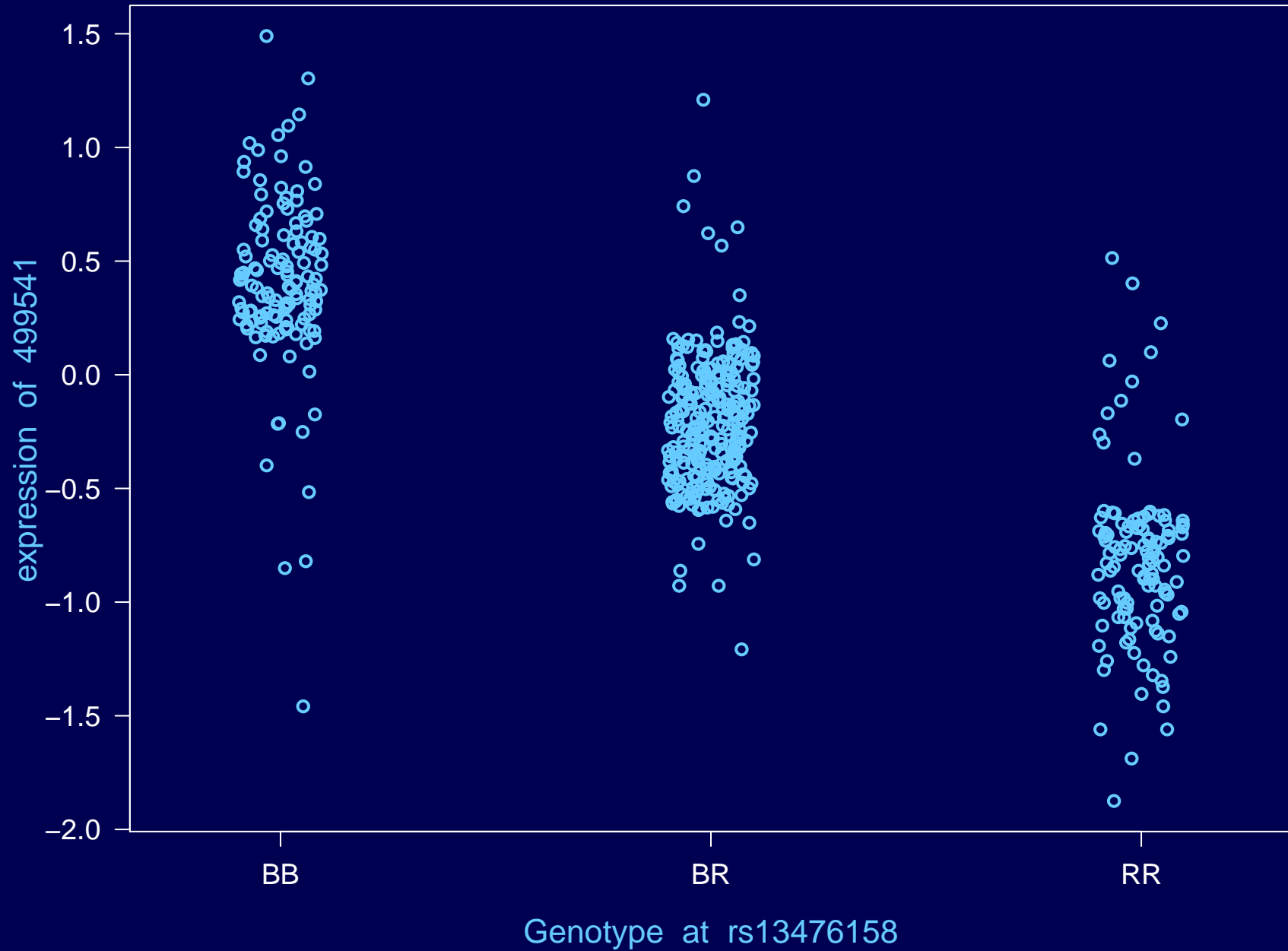
- Sample mix-ups happen
- With eQTL data, we can both identify and correct mix-ups
- There is great value in having expression on multiple tissues
- The general idea here has wide application for high-throughput data
- Related work:
 - Westra et al. (2011) *Bioinformatics* 27:2104–2111
 - Schadt et al. (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?
- Have a system for keeping track of everything
 - Files, versions of files, analyses, . . .
 - Like a lab notebook

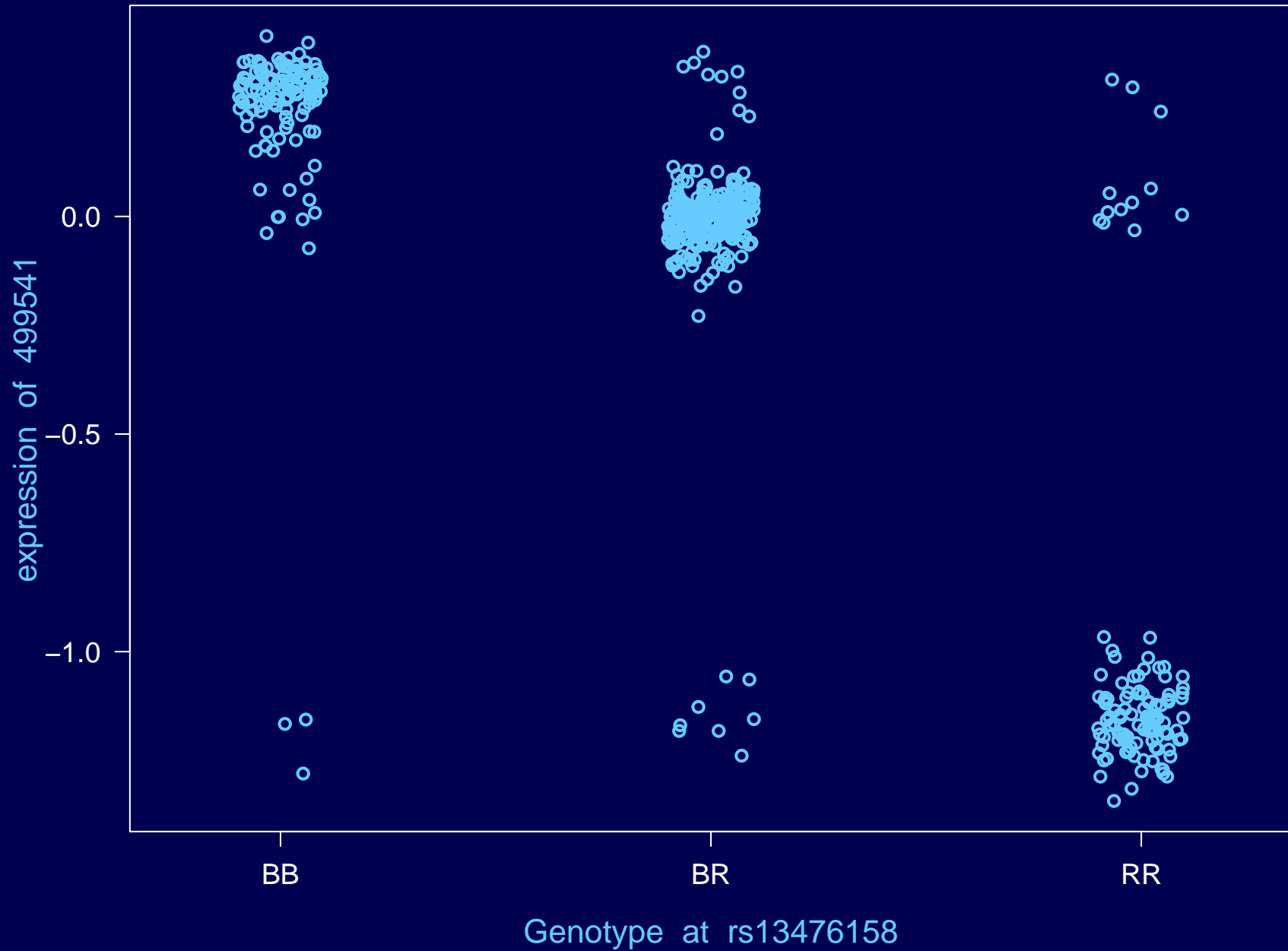
E vs G

transformed scale



E vs G

original scale



Lessons

- Don't fully trust anyone
 - Including yourself
- Make lots of plots
 - Don't rely on summary statistics, like LOD scores
 - Look responses on the original scale
- Follow up all aberrations
- Take your time with data cleaning
 - A month, two months, a year?
- Have a system for keeping track of everything
 - Files, versions of files, analyses, . . .
 - Like a lab notebook

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