Identifying sample mix-ups in eQTL data

Karl Broman

Biostatistics & Medical Informatics, Univ. Wisconsin–Madison

kbroman.org
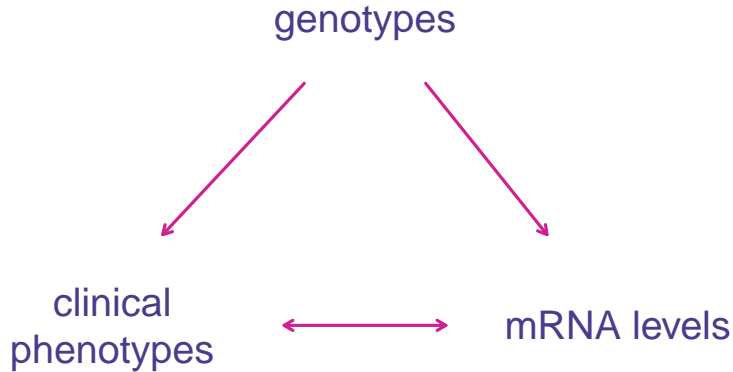github.com/kbroman
@kwbroman
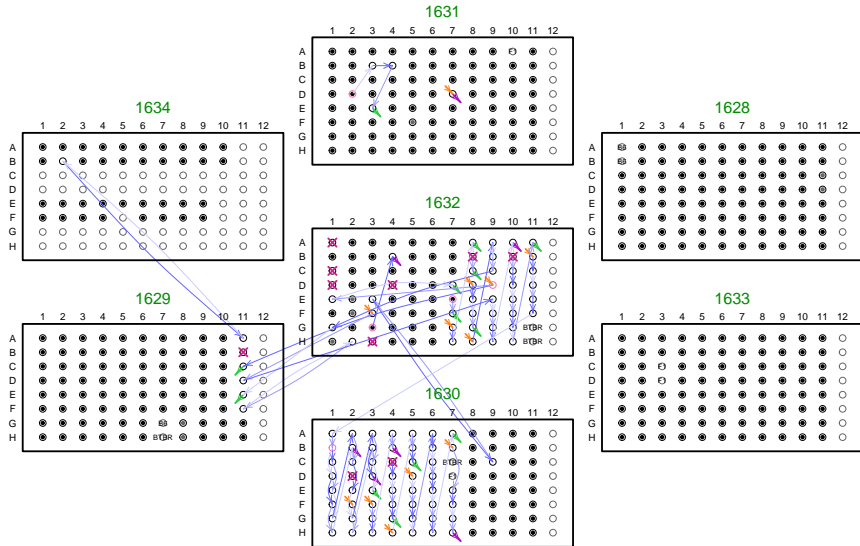Slides: kbroman.org/Talk_OSGA2021

Associations in systems genetics

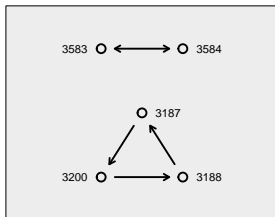genotypes

clinical phenotypes ←→ mRNA levels

Sample mix-ups

3

More sample mix-ups

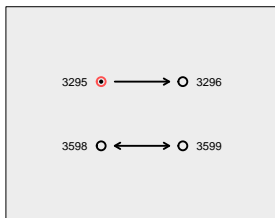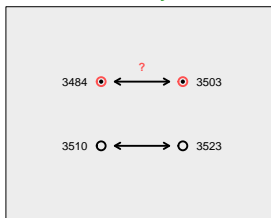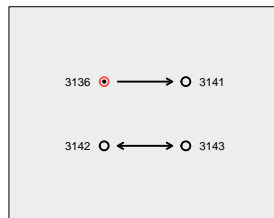# Westra et al. (2011)
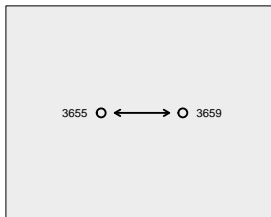
Table 2. *Cis*-eQTL mapping and sample mix-up identification results

| Stud	Population	Sample-size	Initial *cis*-eQTLs	Mix-ups detected[a] *n* (%)	Sample-size after correction *n* (%)	*cis*-eQTLs after correction *n* (%)
Choy *et al.* (2008)	CHB+JP	87	138	20 (23)	79 (90)	418 (+203)
	CE	84	558		NA	NA
	YR	85	274	2 (2)	83 (97)	287 (+5)
Stranger *et al.* (2007)	CHB+JP	90	1511		NA	NA
	CE	90	903		NA	NA
	YR	90	663	1 (1)	89 (99)	667 (+1)
Zhang *et al.* (2009)	CE	87	2581		NA	NA
	YR	89	1454	2 (2)	89 (100)	1635 (+12)
Webster *et al.* (2009)	Brai	36	1284	16 (4)	356 (98)	1367 (+6)
Heinzen *et al.* (2008)	Brai	93	349		NA	NA
	PBMC	80	297		NA	NA

Outline

► Sample duplicates

► Sex verification

► Sample mix-ups:
  – mRNA ↔ protein
  – mRNA ↔ DNA
  – protein ↔ DNA

But first

Missing Data

Percent missing genotypes
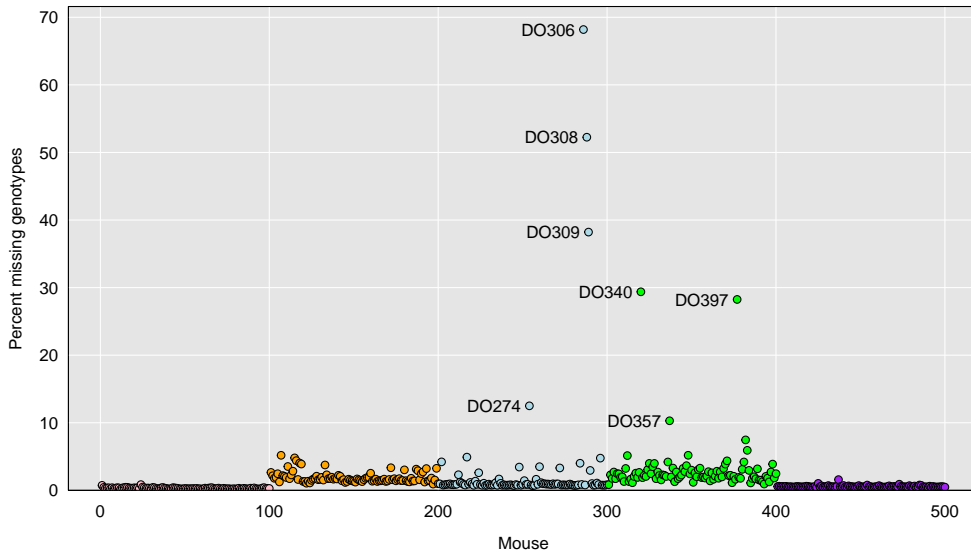
Sample duplicates

Percent matching genotypes

DO306 / DO308

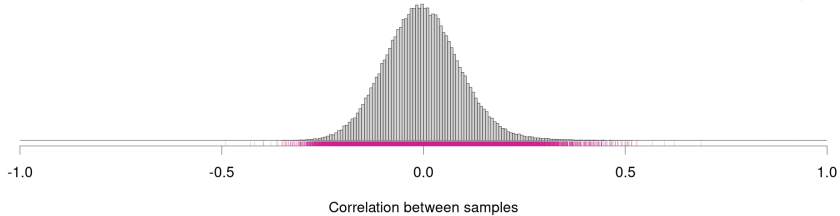Percent matching genotypes
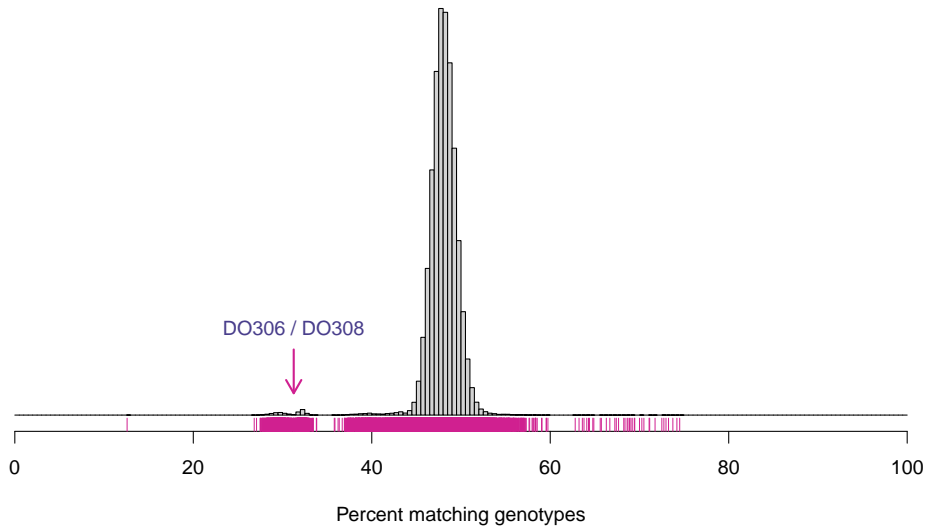
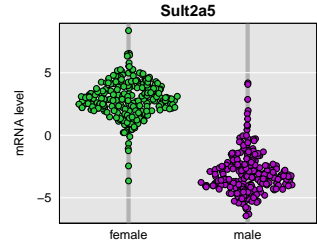# Correlation between mRNA samples
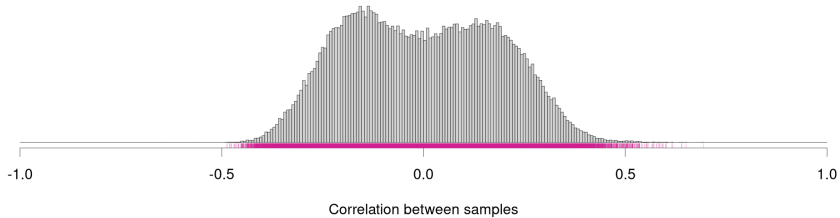
Correlation between protein samples

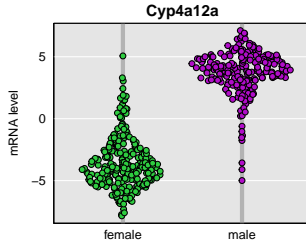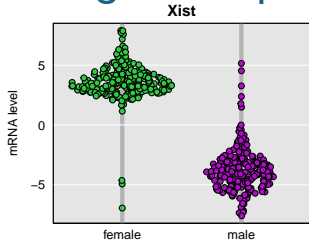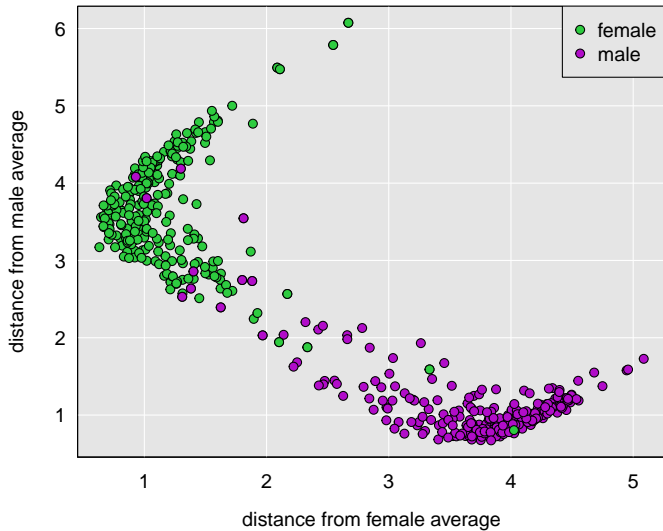# Sex verification
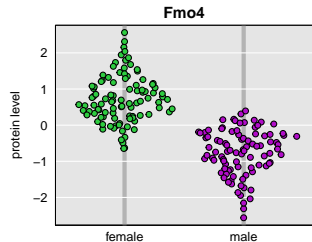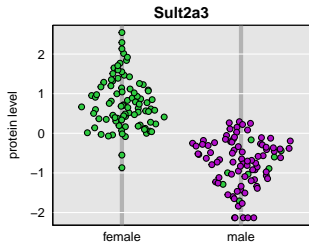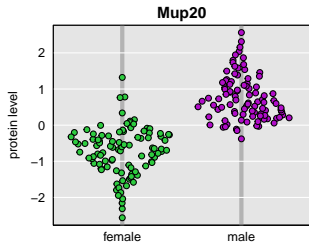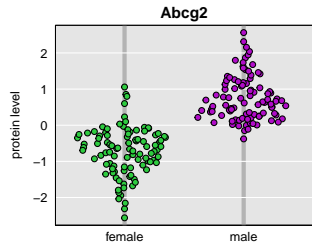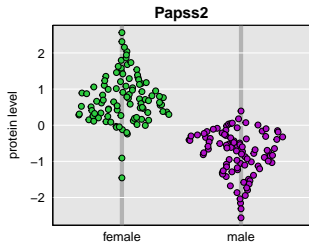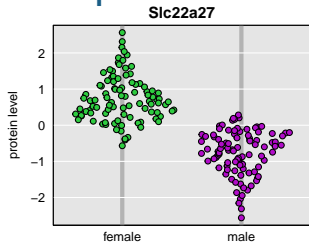
X and Y genotype dosage
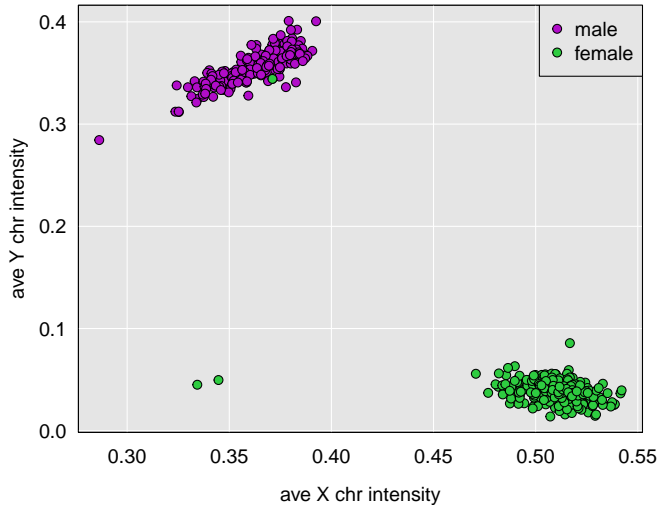
Sex and gene expression
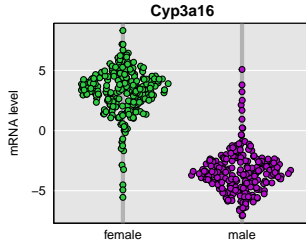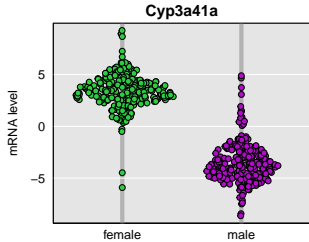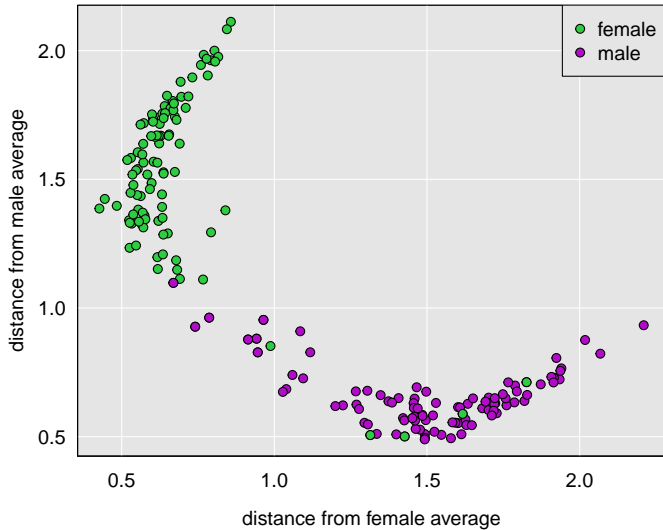
Sex and gene expression

Sex and proteins
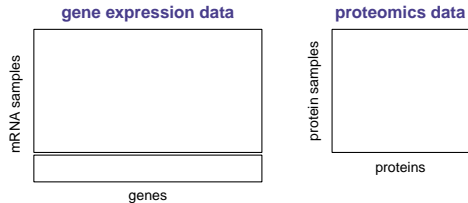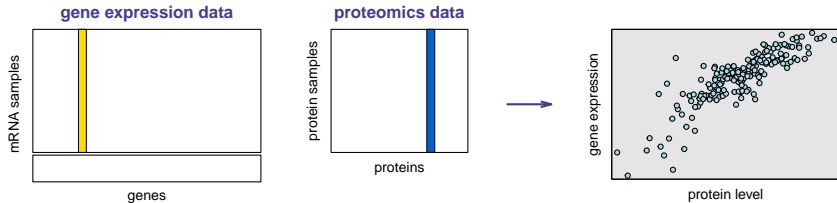
Sex and proteins
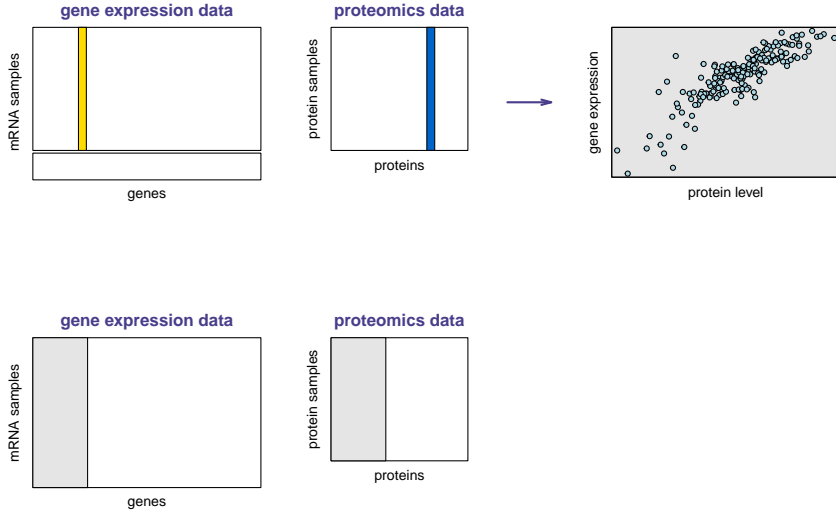
Sample mix-ups

mRNA $\leftrightarrow$ protein

mRNA ↔ protein method

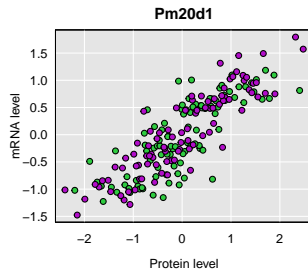**gene expression data**

mRNA samples

genes

proteomics data

protein samples

proteins

mRNA ↔ protein method

mRNA ↔ protein method

mRNA ↔ protein method
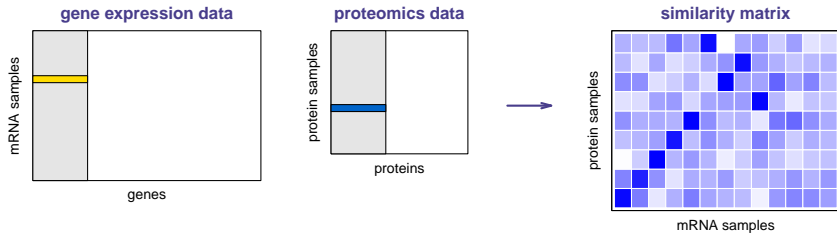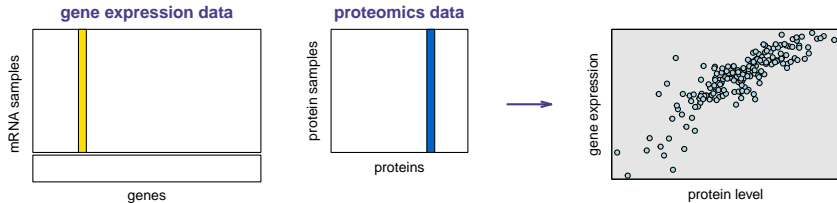
mRNA ↔ protein correlations

max cor(mRNA, protein)
adjusting for sex

Maximum protein/mRNA correlation, by protein
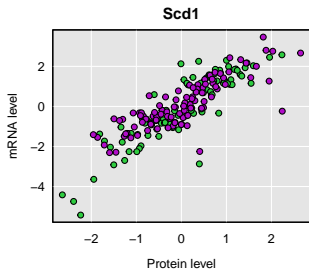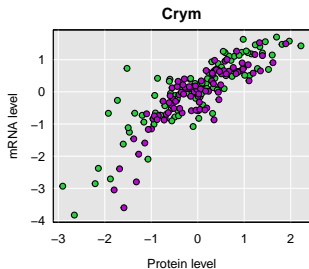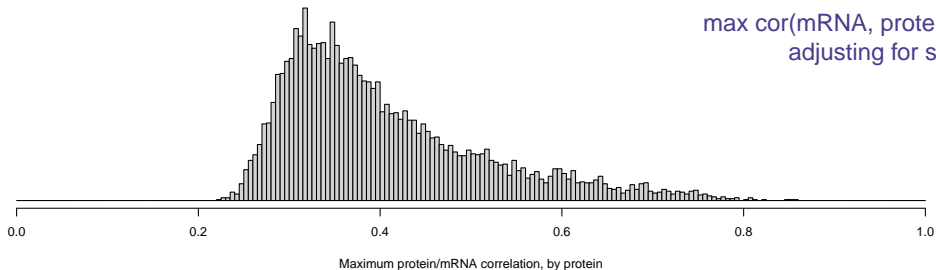
mRNA ↔ protein similarity matrix

mRNA ↔ protein similarities

Self−self
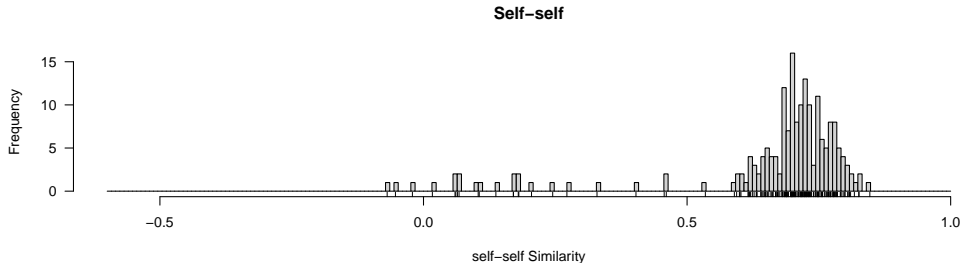
Self−nonself

mRNA ↔ protein: closest vs self
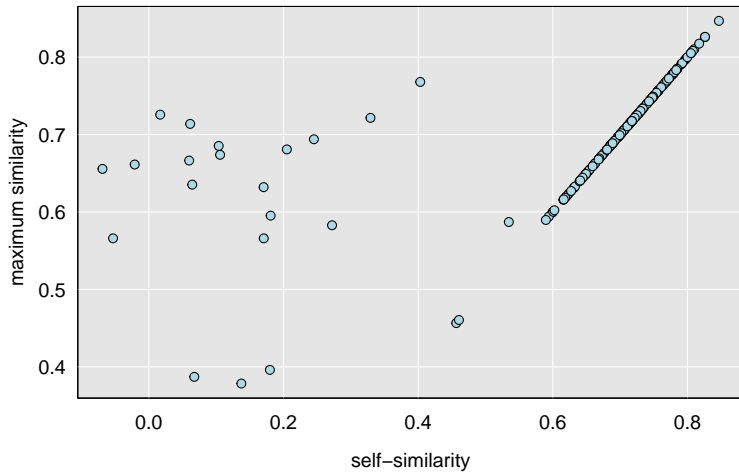
mRNA ↔ protein: selected samples
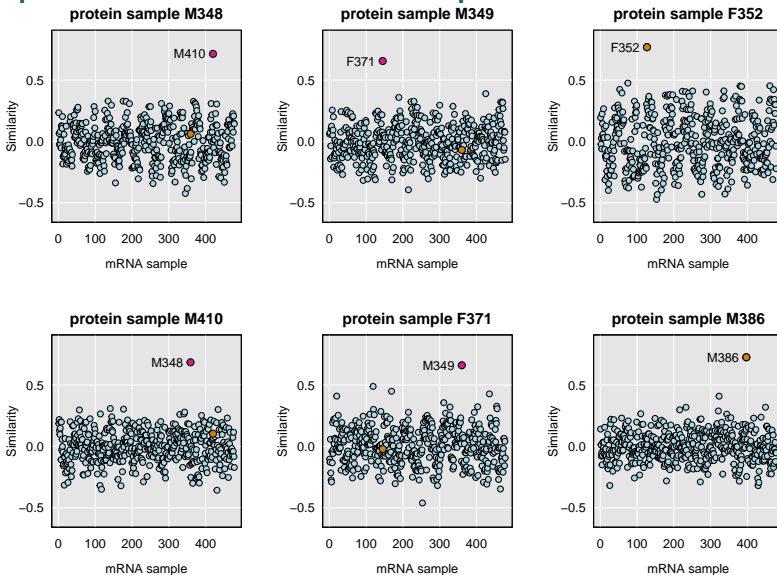
Sample mix-ups

DNA $\leftrightarrow$ mRNA

DNA ↔ mRNA method

eQTL genotypes

DNA samples

eQTL

expression traits

mRNA samples

genes with strong eQTL
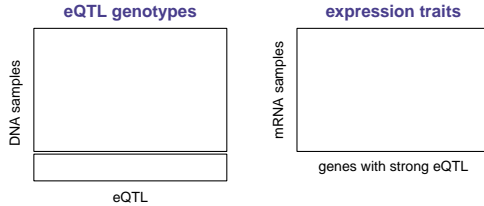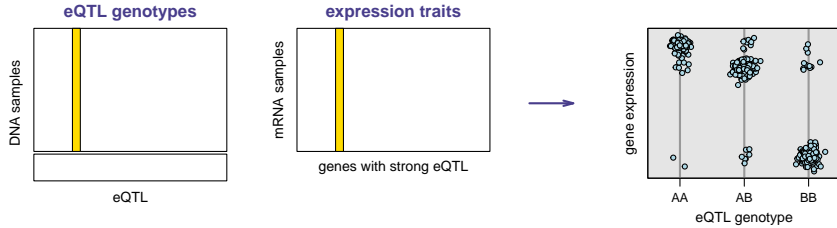
DNA ↔ mRNA method
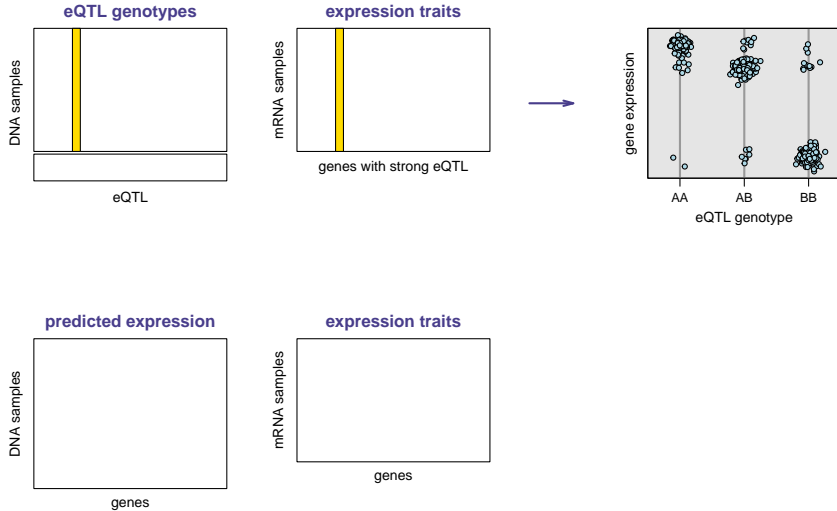
DNA ↔ mRNA method

DNA ↔ mRNA method

DNA $\leftrightarrow$ mRNA correlations

max LOD
adjusting for sex

Maximum LOD score, by mRNA
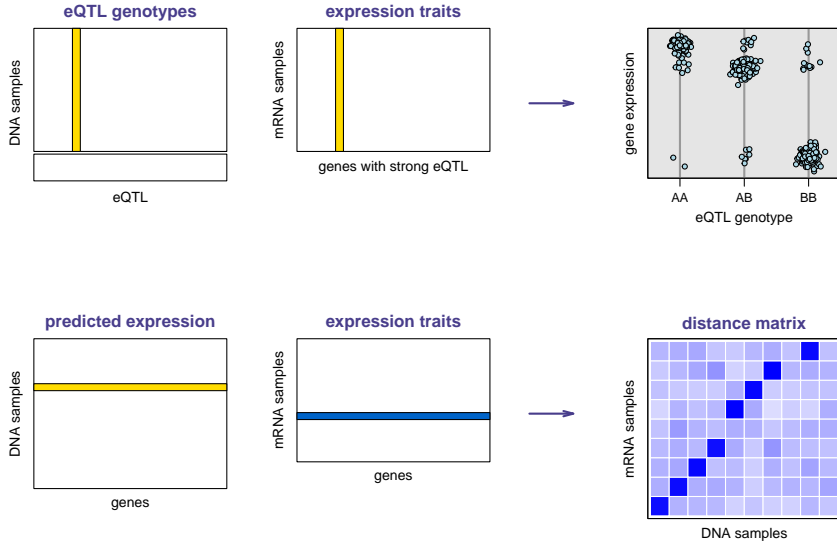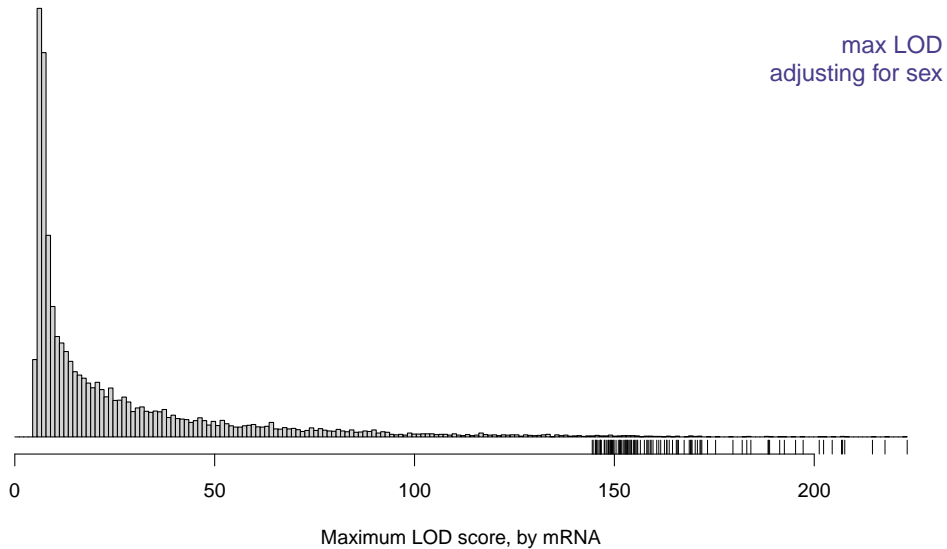
DNA ↔ mRNA similarity matrix
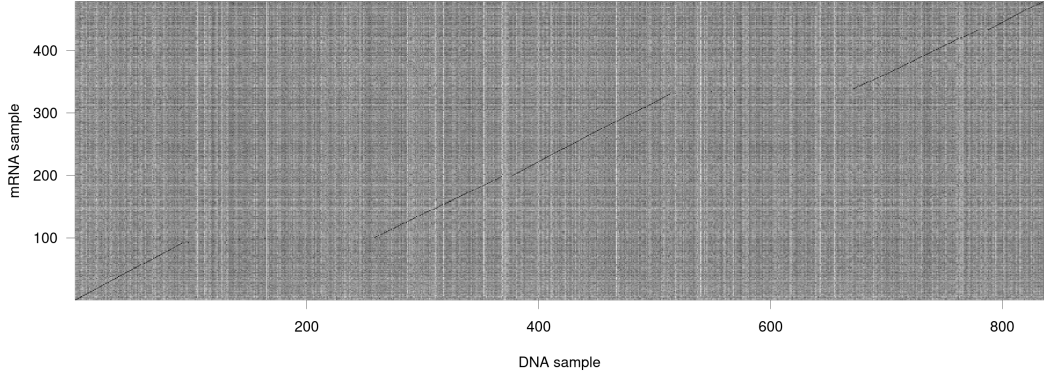
DNA ↔ mRNA similarities

Self–self

Frequency / self–self Distance

Self–nonself

Frequency / self–nonself Distance
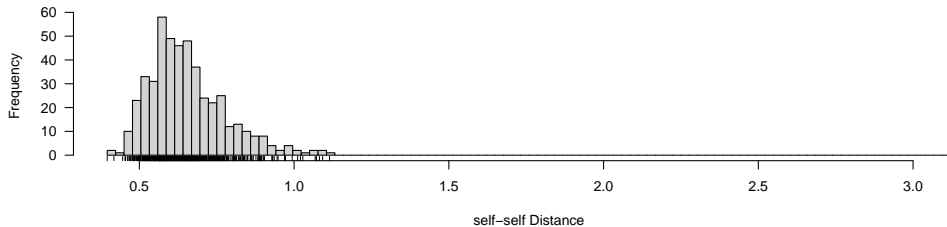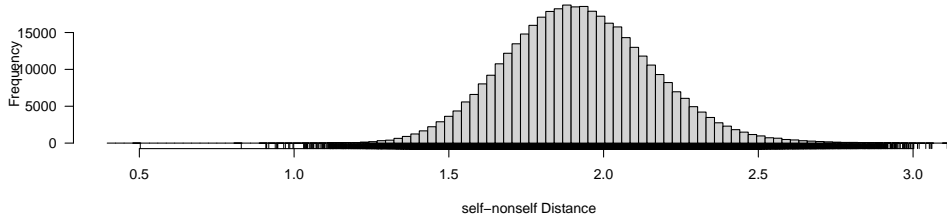
DNA ↔ mRNA: closest vs self
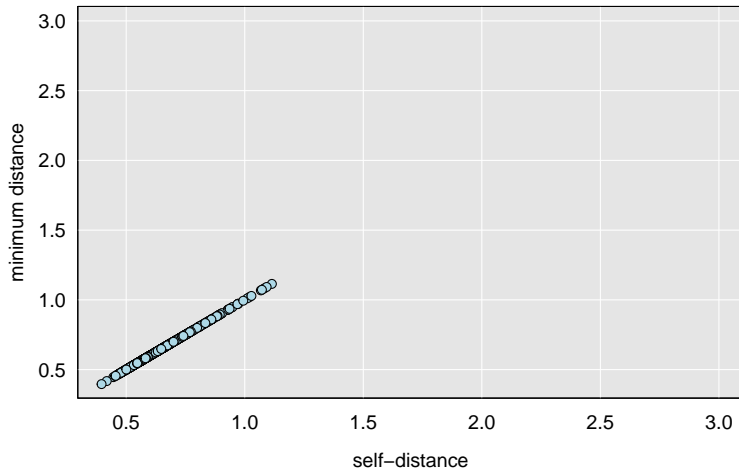
DNA ↔ mRNA: selected samples
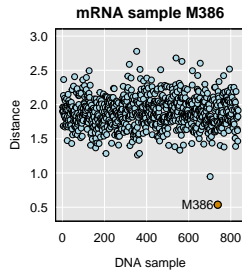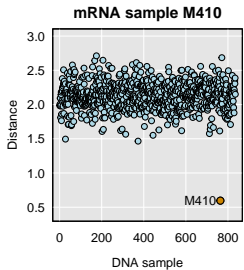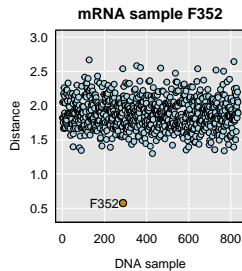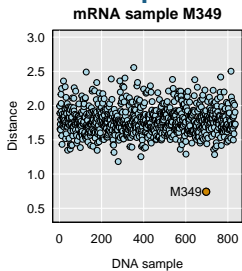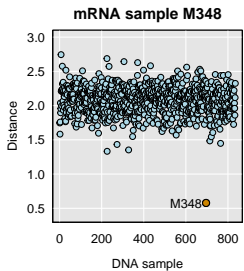
Sample mix-ups

DNA $\leftrightarrow$ protein

DNA ↔ protein method

DNA $\leftrightarrow$ protein correlations

max LOD
adjusting for sex

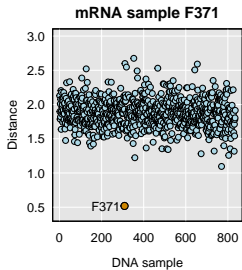Maximum LOD score, by prot
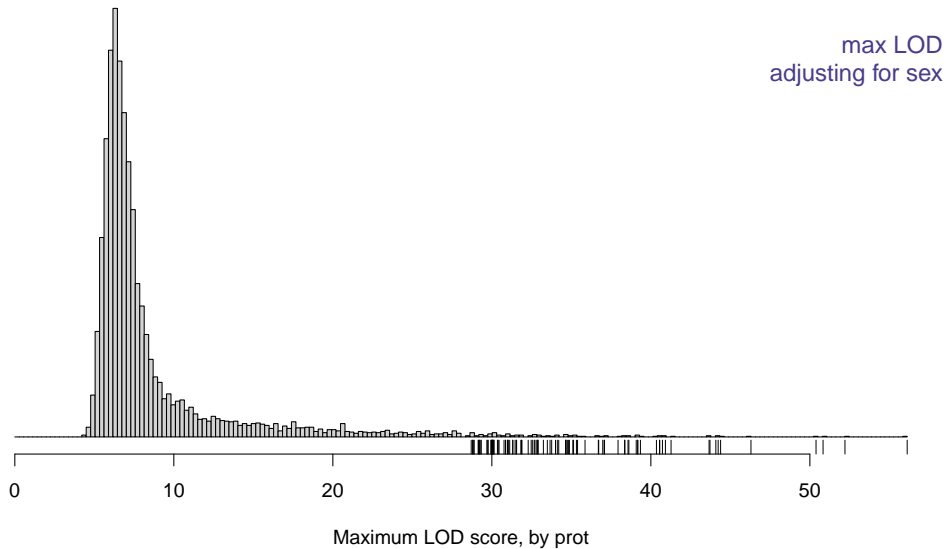
DNA ↔ protein similarity matrix
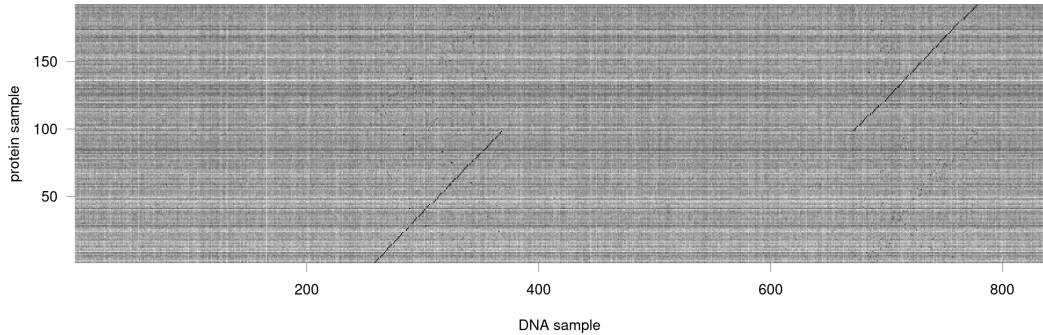
DNA ↔ protein similarities

Self−self
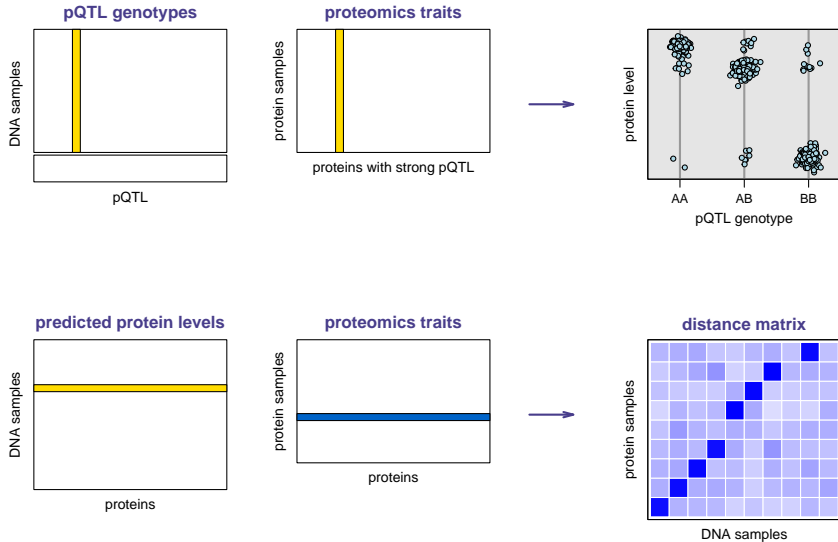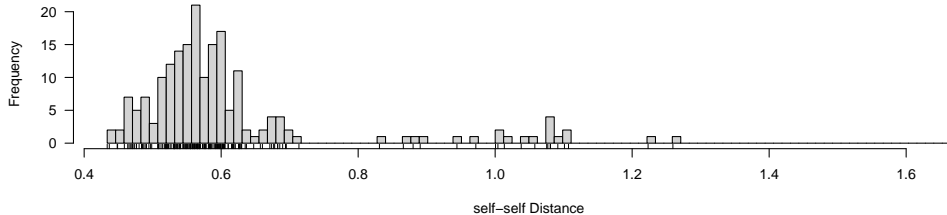
Self−nonself
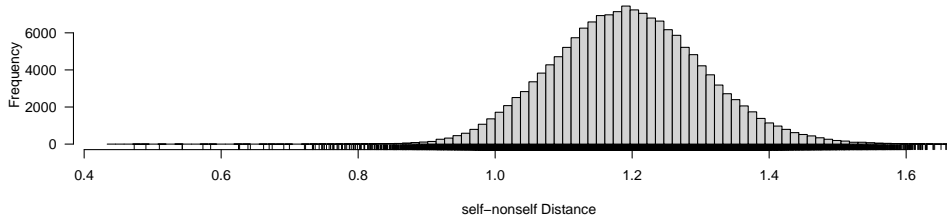
DNA ↔ protein: closest vs self

DNA $\leftrightarrow$ protein: selected samples

Summary

- This shouldn't happen.

- But if it does, you should find it.

- If two data sets have rows that correspond, you should check that they do correspond.
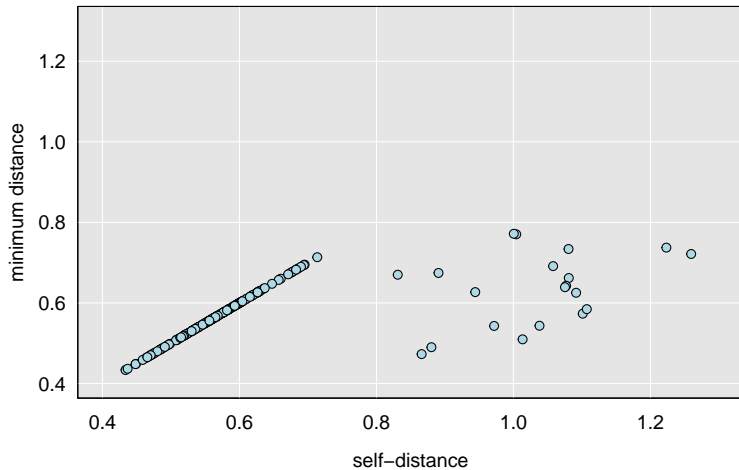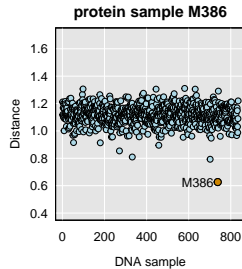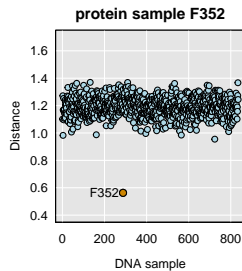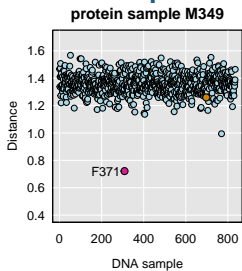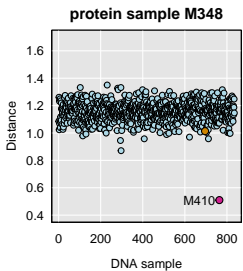
References

► Westra et al. (2011) MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics 15:2104–2111 `doi:10.1093/bioinformatics/btr323`

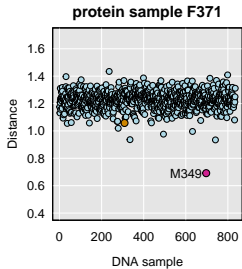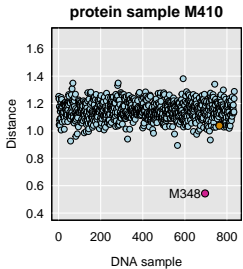► Lynch et al (2012) Calling sample mix-ups in cancer population studies. PLOS One 7:e41815 `doi:10.1371/journal.pone.0041815`

► Broman et al. (2015) Identification and correction of sample mix-ups in expression genetic data: A case study. G3 (Bethesda) 5:2177–2186 `doi:10.1534/g3.115.019778`

► Broman et al. (2019) Cleaning genotype data from Diversity Outbred mice. G3 (Bethesda) 9:1571–1579 `doi:10.1534/g3.119.400165`

Slides: kbroman.org/Talk_OSGA2021

kbroman.org

github.com/kbroman

@kwbroman