Identifying and correcting sample mix-ups in eQTL data

Karl W Broman[1], Mark P Keller[2], Aimee Teo Broman[1], Danielle M Greenawalt[5],
Christina Kendziorski[1], Eric E Schadt[6], Śaunak Sen[7], Brian S Yandell[3,4], and Alan D Attie[2]

[1]Biostatistics and Medical Informatics, [2]Biochemistry, [3]Statistics, [4]Horticulture, UW-Madison; [5]Merck & Co., Inc.; [6]Pacific Biosciences; [7]UC–San Francisco

Abstract

In a mouse intercross with more than 500 animals and genome-wide gene expression data on six tissues, we identified a high proportion of sample mix-ups in the genotype data, on the order of 15%.

Local eQTL (genetic loci influencing gene expression) with extremely large effect may be used to form a classifier for predicting an individual's eQTL genotype from its gene expression value. By considering multiple eQTL and their related transcripts, we identified numerous individuals whose predicted eQTL genotypes (based on their expression data) did not match their observed genotypes, and then went on to identify other individuals whose genotypes did match the predicted eQTL genotypes.

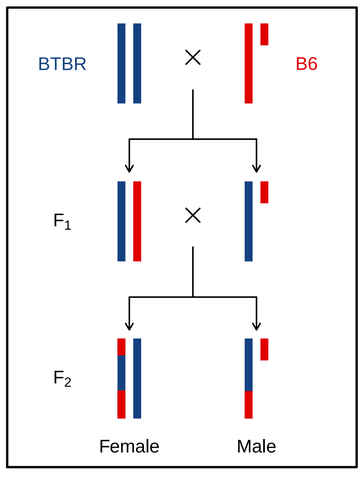
The concordance of predictions across six tissues indicated that the problem was due to mix-ups in the genotypes. Consideration of the plate positions of the samples indicated a number of off-by-one and off-by-two errors, likely the result of pipetting errors.

Such sample mix-ups can be a problem in any genetic study. As we show, eQTL data allow us to identify, and even correct, such problems.

Data

- ∼500 B6 × BTBR intercross mice, all ob/ob
- Genotypes at 2057 SNPs [Affymetrix chips]
- 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)

Initial observation: Sex swaps



We should have:

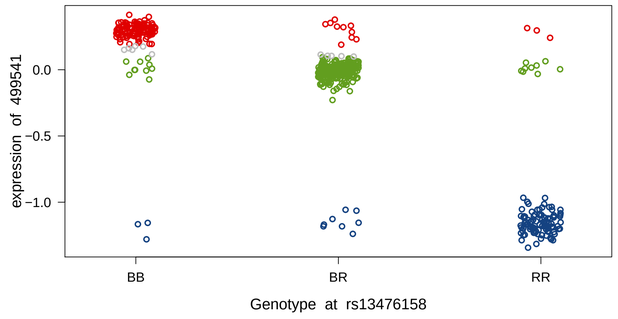
F_2 females:	R/R or B/R
F_2 males:	hemizygous B or R

But 35 mice had X chromosome genotype that conflicted with their sex.

Which are correct: genotypes or sexes?

- We could look for a transcript (e.g., Xist) whose expression level is diagnostic for sex.
- Even better, we can look at transcripts with strong local eQTL (for which genotype is strongly associated with expression level).
- Transcripts with strong local eQTL are diagnostic for genotype. By considering multiple such transcripts across the genome, we can form a DNA fingerprint.
  - ▶ QTL = quantitative trait locus: a genomic region that influences a quantitative trait
  - ▶ eQTL = expression QTL: a QTL that influences the level of expression of a gene

A diagnostic transcript

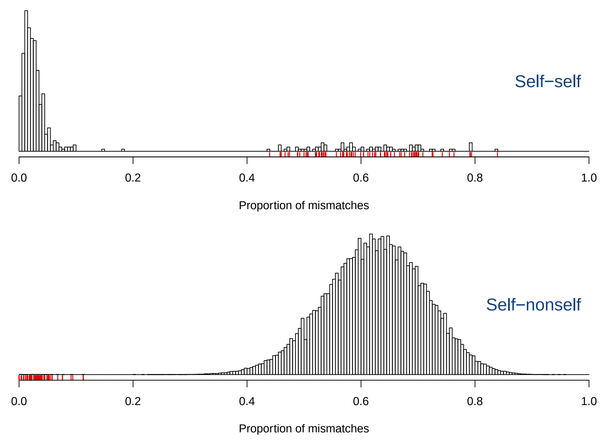


Colors indicate the inferred eQTL genotype according to a k-nearest neighbor classifier, with gray points not called.

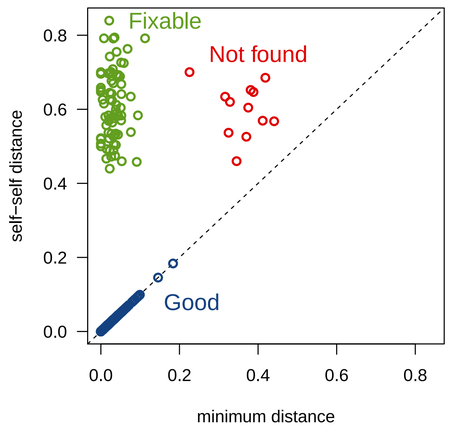
The method

- Identify expression traits with strong local eQTL (that is, for which genotype at the transcript's genomic position is strongly associated with its expression level)
- For each trait, create a classifier for predicting eQTL genotype from expression phenotype
- For each pair of mice, calculate the proportion of mismatches between the observed eQTL genotypes of one mouse and the inferred eQTL genotypes of the other

Proportions of mismatches in eQTL genotypes



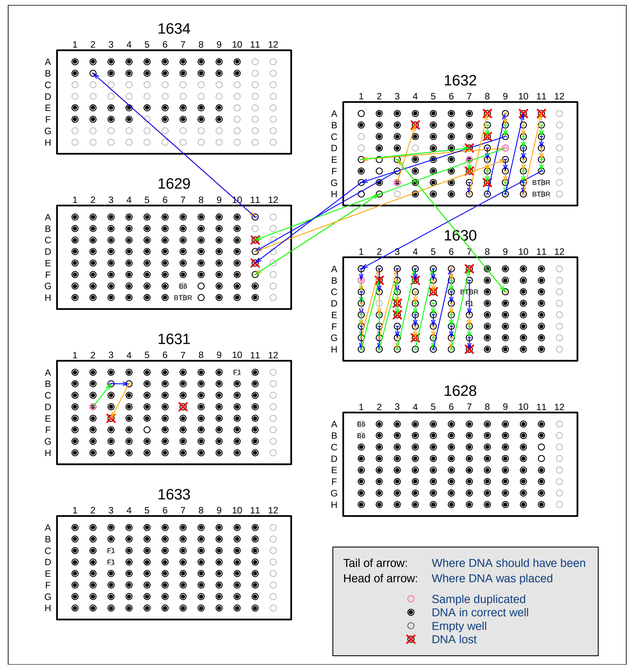
Decisions



There were ∼550 mice with genotypes and ∼500 with expression data.

For each mouse, we plot the proportion of mismatches, between its observed genotype data and the genotypes inferred from the corresponding gene expression data, against the minimal such proportion of mismatches, comparing that observed genotype data to each set of inferred genotypes.

Inferred genotype mix-ups

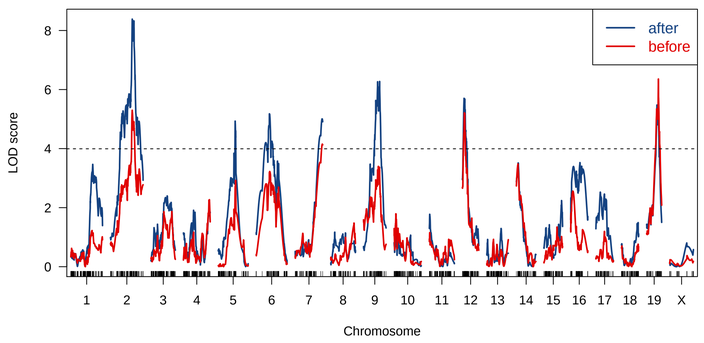


The gene expression data from the multiple tissues were concordant and indicated that the problems were in the genotype data.

(We did, however, identify and correct a small number of sample mix-ups within each of the six sets of gene expression arrays. This was done by considering pairs of tissues and measuring the correlation in a mouse's gene expression across tissues.)

The bulk of the problems concerned apparent pipetting errors in the genotyping plates: a series of off-by-one and off-by-two errors covering half of each of two plates.

Improved results!



LOD curves for insulin, indicating the evidence for QTL, before and after correcting the sample mix-ups. The corrected data give stronger evidence and more QTL.

Summary

- Sample mix-ups happen
- With eQTL data, we can both identify and correct mix-ups
- The general idea here has wide application for high-throughput data
- R package: `http://github.com/kbroman/lineup`
- Very similar to MixupMapper (Westra et al., Bioinformatics 27:2104–2111, 2011)

Contact

Karl Broman
kbroman@biostat.wisc.edu
http://www.biostat.wisc.edu/~kbroman

This work was supported in part by NIH grants GM074244 (to KWB) and DK066369 (to ADA).