Identifying and correcting sample mix-ups in eQTL data

Karl W Broman¹, Mark P Keller², Aimee Teo Broman¹, Danielle M Greenawalt⁶,
Christina Kendzierski¹, Eric E Schadt⁶, Šaunak Sen⁷, Brian S Yandell³,⁴, and Alan D Attie²

¹Biostatistics and Medical Informatics, ²Biochemistry, ³Statistics, ⁴Horticulture, UW-Madison; ⁵Merck & Co., Inc.; ⁶Pacific Biosciences; ⁷UC–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:
F₂ females: R/R or B/R
F₂ males: hemizygous B or R

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

Proportions of mismatches in eQTL genotypes

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

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/lin eup
- Very similar to MixupMapper (Wisstra 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 KB) and DK06869 (to ADA).