

Correction for Hidden Confounders in the Genetic Analysis of Gene Expression

Jennifer Listgarten
Microsoft Research
Los Angeles, CA 90024

Carl Kadie
Microsoft Research
Redmond, WA 98052

Eric E. Schadt
Pacific Biosciences
Menlo Park, CA 94025

David Heckerman
Microsoft Research
Los Angeles, CA 90024

Abstract

Understanding the genetic underpinnings of disease is important for screening, treatment, drug development, and basic biological insight. One way of getting at such an understanding is to find out which parts of our DNA, such as single-nucleotide polymorphisms (SNPs), affect particular intermediary processes such as gene expression. Naively, such associations can be identified using a simple statistical test on all paired combinations of genetic variants and gene transcripts. However, a wide variety of confounders lie hidden in the data, leading to both spurious associations and missed associations if not properly addressed. We present a statistical model that jointly corrects for two particular kinds of hidden structure: genetic or population structure (e.g., race, family-relatedness), and microarray expression artifacts (e.g., batch effects), when these confounders are unknown. Applying our method to both real and synthetic, human and mouse data, we demonstrate the need for such a joint correction of confounders, and also the disadvantages of other possible approaches based on those in the current literature.

Our approach uses a *linear mixed-effects model* depicted in Figure 1. Variable y_g^i , the expression level of gene probe g for individual i , is a linear function of (1) the *fixed effects* \vec{x}^i consisting of one or more SNPs, covariates for that individual, and a bias/offset term, (2) a *random effect* (i.e., hidden variable) u_g^i , which can be thought of as where that individual lies in a one-dimension confounder space, and (3) Gaussian noise. In particular,

$$y_g^i \sim N(\vec{x}^i \vec{\beta}_g + \tau_g u_g^i, \sigma_g^2).$$

The hidden variables \vec{u}_g have a normal distribution with covariance K , where K_{ij} is some measure of similarity between individuals i and j . Because we address confounding from both genetic structure and expression artifacts, we define K to be a mixture of similari-

ties, one based on the SNPs of the individuals, and one based on the gene expressions of those individuals. Importantly, these similarities are shared across all gene probes, allowing us to learn K via (e.g.) maximum likelihood.

Our software is available from <http://www.microsoft.com/science>. The full paper is available as Listgarten et al. (2010).

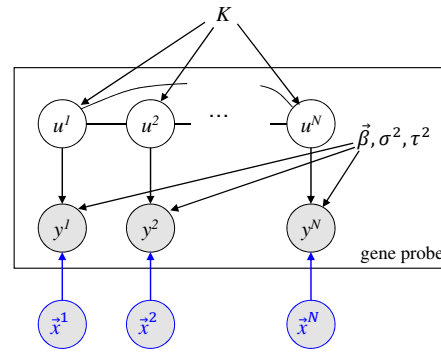


Figure 1: Our linear mixed-effects model. The shaded nodes are observed. Not shown is the mixing weight for the two components of K , which can depend on gene probe.

References

J. Listgarten, C. Kadie, E. E. Schadt, and D. Heckerman (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **107**:16465–16470.