Matrix ASH: Modeling Genetic Effects Across Multiple Subgroups

Sarah Urbut, Gao Wang, Matthew Stephens

Department of Human Genetics and University of Chicago

Introduction

- Variation in gene expression is an important mechanism underlying susceptibility to complex disease.
- However, most studies to date have been conducted in a single immortalized peripheral cell type or single tissue framework
- The solution: GTEx! by 2016: 900 post-mortem donors, with approximately 30 tissues collected from each donor
- Our mission: *jointly analyze data on all tissues* to maximize power, and to identify and quantify the variability in effect sizes.

Objectives

- Combine information across tissues
- Capture distinct variation in effect sizes within and between subgroups: 'patterns of sharing'

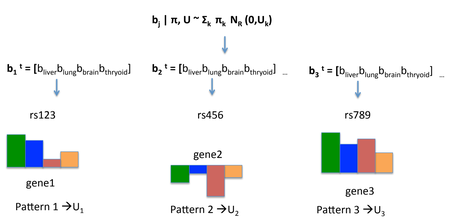


Figure 1: Figure caption

- Assume that each eQTL belongs to a group characterized by its effects across tissues.
- Within these groups, the tissues exhibit characteristic patterns of sharing
- Natural mixture model, in which we assume all the gene-snp pairs arise from a mixture of a finite number of Gaussian distributions
- Each component of the mixture is defined by the prior covariance matrix U_k from which the vector of standardized effect sizes $\boldsymbol{b}_j$ is drawn.
- 'Learn' the relative proportions of each pattern of sharing from the data
- **Key: Distinct data-sensitive diagonal and off-diagonal elements capture a wide array of patterns of sharing**

Mathematical Section

We assume the following mixture prior for the R dimensional vector of true effects, $\boldsymbol{b_j}$ represents the genetic effect of SNP-gene pair j across $R = 44$ tissues:

$$\boldsymbol{b_j}|\boldsymbol{\pi}, \mathbf{U} \sim \sum_{\mathbf{k,l}} \pi_{\mathbf{k,l}}\, N_{\mathbf{R}}(\mathbf{0}, \omega_{\mathbf{l}}\mathbf{U_k}) \qquad (1)$$

For a given ω_l, we specify 4 'types' of RxR prior covariance matrices $U_{k,l}$.

1. $U_{k=1,l} = \omega_l\, \mathbf{I}_R$
2. $U_{k=2,l} = \omega_l\, \frac{1}{J}\, Z^t_{center}\, Z_{center}$
3. $U_{k=3,l} = \omega_l\, \frac{1}{J}\, V_{1..p}\, d^2_{1..p}\, V^t_{1..p}$ is the rank p eigenvector approximation of the tissue covariance matrices
4. $U_{k=4:4+Q-1,l} = \frac{1}{J}((\Lambda\mathbf{F})^t\Lambda\mathbf{F})_q$ corresponding to the q_{th} sparse factor representation of the tissue covariance matrix
5. $U_{k=4+Q,l} = \frac{1}{J}\,(\Lambda\mathbf{F})^t\Lambda\mathbf{F}$ is the sparse factor representation of the tissue covariance matrix, estimated using all q factors.
6. $U_{k=5+Q:R+4+Q,l} = \frac{1}{J}\,([100..]'[100...]$
7. $U_{k=R+5+Q,l} = \frac{1}{J}\,([111...]'[111...])$

- Compute the mixture weights $\boldsymbol{\pi_{kl}}$ hierarchically
- For a given gene-snp pair, the Likelihood on $\boldsymbol{b}$:

$$\hat{\boldsymbol{b}}|\boldsymbol{b} \sim N_R(\boldsymbol{b}, \hat{V}) \qquad (2)$$

- We know that for a single multivariate *Normal* the posterior on $\boldsymbol{b}|U_0$ is simply:

$$\boldsymbol{b}|\hat{\boldsymbol{b}} \sim N_R(\boldsymbol{\mu}_1, U_1)$$

$$p(\boldsymbol{b}|\hat{\boldsymbol{b}}, \hat{V}, \hat{\boldsymbol{\pi}}) = \sum_{k=1,l=1}^{K,L} \sim N_R(\boldsymbol{\mu}_{1kl}, U_{1kl})\tilde{\pi}_{k,l} \qquad (3)$$

- Since the prior weights are computed from 'mostly null data' the prior weights will heavily weight the components with small posterior means (as determined by small prior variance in U_k).
- $\tilde{\pi}_{k,l} = P(Component|Data) \propto P(Data|Comp.) \times P(Comp.)$: Combine hierarchical and snp-specific information
- Allows pair to find its true match!

Results

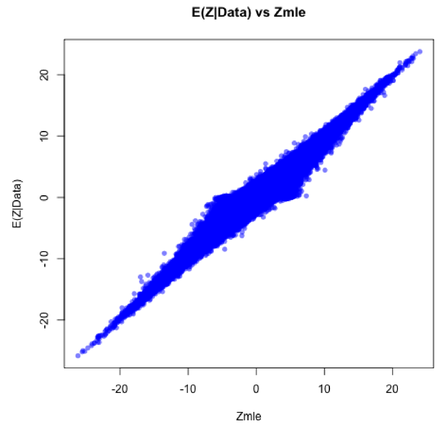


Figure 2: E(Z|Data) vs MLE: Using information gleamed from our high 'prior' probability of observing a particular effect due to the effects other tissues leads us to shrink some summary statistics more than others

Identify strong tissue specific QTLs as gene-snp pairs in which *LFSR* is low in only one tissue.

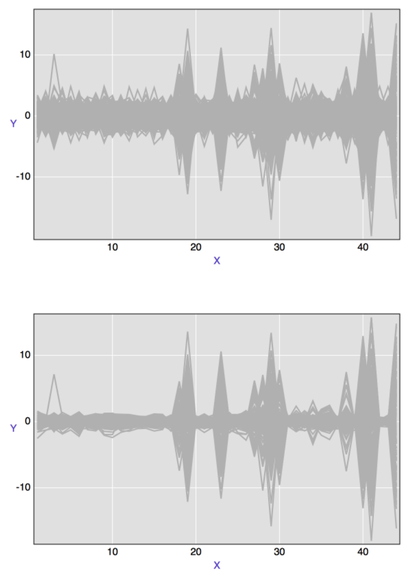
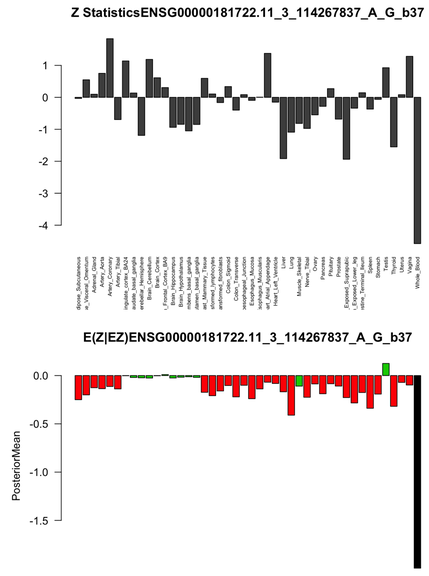


Figure 3: Small statistics are 'denoised' without losing strong tissue-specific effects

Conclusion



- Small positive effects in several tissues are pushed towards the overall negative effect
- Tissue specific effect in whole blood is allowed to remain: high prior belief in tissue-specific effects here.

... but we don't shrink strong effects - large effects preserved (see extremes of Figure 2)

Additional Information

- Learn about the overall heterogeneity of the data set
- For each gene snp pair, *Pr(Inconsistent Signs)* in 100 simulations from (3) according to $\tilde{\pi}$