Matrix Ash

Sarah Urbut [1,2], Gao Wang 1, Matthew Stephens [1,3,‡], with the GTEX Consortium[¶]

1 Department of Human Genetics/ University of Chicago, Chicago, IL USA
2 Pritzker School of Medicine/Growth and Development Training Program/University of Chicago, Chicago, IL USA
3 Department of Statistics/ University of Chicago, Chicago, IL USA

‡These authors also contributed equally to this work.
¶Membership list can be found in the Acknowledgments section.
*** CorrespondingAuthor@institute.edu**

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Author Summary

Variation in gene expression is an important mechanism underlying susceptibility to complex disease. The simultaneous genome-wide assay of gene expression and genetic variation allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression QTLs; eQTLs). By analyzing these effects across multiple tissues, we exploit the information that the effect of the gene-snp pair in one tissue can provide about its effect in alternative tissues. Furthermore, quantifying the effect size as opposed to simply calling QTLs present or absent reveal many patterns of sharing of effects among tissues which differ in both sign and magnitude. We provide a novel framework for estimating effect sizes across multiple subgroups, considering the evidence contained in all subgroups jointly, which provides a powerful and detailed insight into quantitative heterogeneity present in the genome.

Introduction

Variation in gene expression is an important mechanism underlying susceptibility to complex disease.

The simultaneous genome-wide assay of gene expression and genetic variation allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression QTLs; eQTLs). The availability of this information provides immediate insight into a biological basis for disease associations identified through genome-wide association (GWA) studies, and can help to identify networks of genes involved in disease pathogenesis ([**?**, **?**]). Available methods are limited not only in their ability to *jointly analyze data on all tissues* to maximize power, but also in simultaneously *allowing for both qualitative and quantitative differences among eQTLs* present in each tissue.

Initial approaches to quantify the effect of a particular SNP on gene expression considered only one tissue at a time, and ignored the effect of the SNP on gene expression in other tissues. This fails to exploit the power of shared genetic variation in effects on expression - i.e. the information that the effect of the gene-snp pair in one tissue can provide about the effect in another- and limits our understanding of multiple-tissue phenotypes. Furthermore, even past attempts at quantifying heterogeneity of eQTLs using the data across tissues jointly were limited in both the number of tissues considered, and also the level of heterogeneity considered. Qualitative heterogeneity refers to calling a snp 'active' or 'inactive' in a given tissue. For example, previous work has referred to the setting in which the gene-snp pair is active in all tissues as 'shared' and active in only one as 'tissue-specific' ([**?**, **?**]). However, a QTL may be'active' in all or many tissues and with varying magnitude or sign; we refer to this as quantitative heterogeneity. Indeed, our initial motivation came from our analysis of GTEx pilot data, in which we see evidence that many (50%) QTLs are shared across all nine tissues. In this context, a QTL is called based on whether it demonstrates significant posterior probability of being active in a particular tissue. Applying our hierarchical model to the dataset from Dimas *etal* 1001[?] with 3 tissues, we found just 8% of eQTLs are specific to a single tissue, with an estimated 88% of eQTLs being common to fibroblasts, LCL cells and T-Cells [**?**]. Not all eQTLs are shared by all tissues; some tissues may share eQTLs more than others. To allow for this, our previous hierarchical model attempted to infer the extent of such sharing by estimating the proportion of eQTL which were shared in various 'configurations' or patterns of binary activity. However, these binary configurations are still *limited in their ability to capture continuous variation* in levels of activity among tissues in which the SNP is considered active. In fact, as the number of tissues considered increases, perhaps the more interesting and biologically relevant question becomes one of quantitative heterogeneity - that is, how do the patterns of effect vary across tissues in which the SNP is called 'active'. Quantifying the effect sizes of the gene-snp pair across tissues considering the evidence contained in all tissues jointly thus reveals new patterns of activity across tissues, which differ in their relationship in sign and magnitude within and between tissues. Thus effects can be 'shared' but not 'consistent' across tissues. The structure of this paper is as follows: we will describe our approach for modeling and estimating these effect sizes across tissues,

0.1 Methods Overview: Approach

We aim to learning about patterns of sharing across tissues within a SNP and among SNPs, which join to help us better understand the global and snp-specific patterns of effects of genetics on gene expression. This allows us to make comparisons among tissues in which the QTL is called active, and among gene-snp pairs with a similar degree of activity in a given tissue. Thus as an additional level of combining information, we assume that each eQTL may follow a particular pattern of activity characterized by its effects across tissues. Within these groups, the tissues exhibit characteristic patterns of sharing, which can be captured by considering the covariance structure of the genetic effects among tissues. This lends itself to a mixture model, in which we assume all the gene-snp pairs arise from a mixture of a finite number of multivariate normal (MVN) distribution, each characterised by the covariance matrix from which the vector of effects is though to arise. For each of J gene-snp pairs, we observe an R dimensional vector of standardized effect sizes $\hat{b}$ and their standard error and assume that these effects descend from some true effect size $\boldsymbol{b}$.

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

Here, the covariance matrix U_k captures the particular patterns of sharing - variation in effect sizes within and between tissues, while ω_l determines the scale of each pattern - the magnitude of the effect size. Thus we recognize that while two eQTL may obey a similar pattern or shape, the absolute scale may vary. For example, two eQTL may both have strong correlation between tissues 1 and 2 with consistently larger effects in 2, but the absolute size of the effects may vary between SNPs.

Previous work from our lab considered the idea of configuration - i.e., that a tissue was simply ‘active’ or ‘inactive’ in a particular tissues - and thus for R tissues, there were $2^Rpossibleconfigurations.whichbecomescomputationallyinfeasibleasRgrows.Furthermore,th$
$[\mathbf{?},\mathbf{?}])thesematricescontaindistinctdiagonalandoff-$
$diagonalelementswhichreflectdata-$
$specificpatternsofvariationwithinandcovariancebetweensubgroups(tissues).Thiscapturesthe$

Because we can’t know the ’true covariance matrix’ for each gene-snp pair, we aim to assemble a list which sufficiently captures the various patterns, and then ’learn’ the relative proportions of each pattern of sharing from the data. One can now model each vector of effect sizes $\boldsymbol{b}$ each as arising from a mixture that captures all the covariance patterns.

The primary novelty of this approach is *to estimate this multivariate posterior distribution on the effect size in a data-sensitive way* - i.e., using the mixture model to capture information about the covariance structure among subgroups (here, tissues). Thus we might identify a situation in which it is common to have large effects in some tissues and not others, and thus if a gene-snp pair demonstrates a small effect in one of the ’off issues’, we might be inclined to conclude that it is indeed a member of this particular class and shrink the small effect in this tissue accordingly. However, if we see the same small effect in a setting in which ’similar tissues’ have large effects, we might ’shrink’ this effect size less, due to our high prior belief in the SNP’s effectiveness garnered from adjacent tissues. Thus we deem this method ’adaptive Shrinkage’ because the appropriate amount of shrinkage is learned from the data.

Because our prior belief in consistency is strong, we identify many more ?significant associations? in setting where perhaps the observed univariate statistic in one tissue is small but otherwise large in additional tissues, nudging these effects towards something more consistent. This is in contrast to a univariate shrinkage approach, in which all effects of the same size would be ’shrunk’ equivalently, due to lack of information garnered from adjacent tissues.

In fact, Shrinkage towards 0 of small effects is a result, not a necessity - since the majority of the prior weight is on small ω components which emphasize components with small prior variance of the effect size b, many of the modest z statistics will be smoothed or shrunk towards the prior mean, 0. An additional novelty is that in learning something about the effect size in each tissue for a given gene-snp pair, we can make statements about the degree of heterogeneity - that is the proportion of the time we expect a SNP to have effects of different sign. We will be confident in our ability to identify the direction of the effect for A SNP with a large effect and relatvilely high precision, and thus we can use an estimate of the posterior mean in each component and the proportion to quantify the distribution of gene SNP pairs who have effects of opposite direction (or lack convincing evidence of effects in a consistent direction across tissues).

Materials and Methods

Let $\boldsymbol{b_j}$ represents the genetic effect of SNP-gene pair j across $R = 44$ tissues.

We assume the following mixture prior for the R dimensional vector of true effects,

$$\boldsymbol{b_j}|\boldsymbol{\pi}, \mathbf{U}, \boldsymbol{\omega} \sim \sum_{\mathbf{k,l}} \pi_{\mathbf{k,l}}\; N_{\mathbf{R}}(\mathbf{x}; \mathbf{0}, \omega_{\mathbf{l}}\mathbf{U_k}) \tag{2}$$

Where $N_R(.; \mathbf{0}, \omega_l U_k)$ denotes the density of a normal distribution with mean $\mathbf{0}$ and variance $\omega_l U_k$.

Each component of the mixture distribution is characterized by these prior covariance matrices, U_k which capture the pattern of effects across tissues. Critically, this prior distribution is the same for all J - hence the hierarchical incorporation of shared information.

0.2 Covariance Matrices

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. $U_{k=2,l} = \omega_l \mathrm{X}_z$ The (naively) estimated tissue covariance matrix as estimated from the column-centered J $\times R$ matrix of Z statistics, Z_{center}: $\frac{1}{J}\ Z_{center}{}^t\ Z_{center}$
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, i.e., the sum of the first p eigenvector approximations, where $_{1...p}$ represent the eigenvectors of the covariance matrix of tissues and $_{1...p}$ are the first p eigenvalues.
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...])
8. [1000...] or [111...] represent configurations such that given membership,$\boldsymbol{b_j}$ arise from the same prior variance.

0.3 Deconvolution

To retrieve a 'denoised' or 'deconvoluted' estimate of the non-single rank dimensional reduction matrices, we then perform deconvolution after initializing the EM algorithm with the matrices specified in (2), (3) and (5). The final results of this iterative procedure preserves the rank of the initialization matrix, and allows us to use the 'true' effect at each component component $\boldsymbol{b_j}$ as missing data in deconvoluting the prior covariance matrices. In brief, this algorithm works by treating not only the component identity but also the true effect $\boldsymbol{b_j}$ as unobserved data, and maximizing the likelihood over the expectation of the complete data likelihood, considering the values $\boldsymbol{b_j}$ as extra missing data (in addition to the indicator variables q_{ij}) (Bovy et al, 2014). This allows us to write down the 'full data' log likelihood as follows:

$$\phi = \sum_J \sum_K q_{jk} ln \alpha_k N(\hat{\boldsymbol{b_j}}|\mathit{0}, U_k + V_j)$$
$$\phi = \sum_J \sum_K q_{jk} ln \alpha_k N(\boldsymbol{b_j}|\mathit{0}, U_k)$$

(3)

Where α_k represents π_k and q_{jk} is the latent identifier variable.

0.4 Likelihood

By maximum likelihood in each tissue separately, we can easily obtain the observed estimates of the standardized genotype effect sizes, $\hat{\boldsymbol{b}}_j$, and their observed squared standard errors recorded on the diagonal of an $R \times R$ matrix noted $\hat{V}_j = (\hat{\boldsymbol{b}}_j)$. We assume that the matrix of standard errors of $\hat{\boldsymbol{b}}_j$, V_j as approximated by $\hat{V}_j$ is diagonal and that $\hat{V}_j$ is an accurate point estimate for the standard error and that these standard errors are independent between tissues.

If we now view $\hat{\boldsymbol{b}}_j$ and $\hat{V}_j$ as *observed data*, we can write a new "likelihood" using only the sufficient statistics, $\hat{\boldsymbol{b}}_j$ and $\hat{V}_j$:

$$\hat{\boldsymbol{b}}_j|\boldsymbol{b}_j \sim_R (\mathbf{x}; \mathbf{b_j}, \hat{\mathbf{V}}_\mathbf{j}) \tag{4}$$

0.5 Posterior Quantities

We aim to report posterior quantities for a given gene-snp pair **j**. We know that for a single multivariate *Normal* the posterior on $\boldsymbol{b}|U$ is simply:

$$\boldsymbol{b}|\hat{\boldsymbol{b}} \sim N_R(\tilde{\boldsymbol{\mu}}, \tilde{U})$$

where:

- $\tilde{\boldsymbol{\mu}} = \tilde{U}(\hat{V}^{-1}\hat{\boldsymbol{b}})$
- $\tilde{U} = (U^{-1} + \hat{V}^{-1})^{-1}$.

Let us concatenate the list of all KxL combinations of prior covariance matrices U_k and their scaling parameters ω_l into a KxL list and assign this length K for simplicity of notation.

Now each U_k imparts information about both *scale* and *direction* . Furthermore, a mixture-multivariate normal prior and a normal likelihood yields a mixture multivariate posterior, where the final posterior distribution is simply a weighted combination of multivariate normal distributions, where for each gene-snp pair **j** is now characterized by it's posterior mean $\tilde{\boldsymbol{\mu}}_{jk}$ and covariance $\tilde{U}_{jk} = (U_k^{-1} + \hat{V}_j^{-1})^{-1}$.

$$\boldsymbol{b}_j|\hat{\boldsymbol{b}_j}, \hat{V}_j, \hat{\boldsymbol{\pi}} \sim \sum_k^K \tilde{\pi}_{jk} N_R(\mathbf{x}; \tilde{\boldsymbol{\mu}}_{\mathbf{jk}}, \tilde{\mathbf{U}}_{\mathbf{jk}}) \tag{6}$$

Where again, $N_R(.; \mathbf{0}, \omega_l U_k)$ denotes the density of a normal distribution with mean $\tilde{\boldsymbol{\mu}}_k$ and variance $\tilde{U}_k$ and the posterior mixture weight $\tilde{\pi}_k$ is simply

$$\tilde{\pi}_{jk} = \frac{p(\hat{\boldsymbol{b}}_j|\hat{V}_j, z_j = k)\hat{\pi}_k}{\sum_{k=1}^K p(\hat{\boldsymbol{b}}_j|\hat{V}_j, z_j = k)\hat{\pi}_k} \tag{7}$$

Where $z_j = k$ is the latent variable indicator of the component identity and each $\hat{\pi}_k$ represents the Maximum Likelihood Estimate of the prior mixture weights assigned to each component.

0.6 Reported Quantities

For every gene-snp pair ‘j’, we aim to report the effect size as the posterior mean, defined as:

$$E(\boldsymbol{b}_j|\hat{\boldsymbol{b}_j}, \hat{V}_j, \hat{\boldsymbol{\pi}}) = \sum_{k}^{K} \tilde{\pi}_k \tilde{\boldsymbol{\mu}}_k \tag{8}$$

And the local false sign rate, or posterior probability of incorrectly identifying the sign of the effect for a given tissues ’r’ as :

$$P(b_{jr}) = 1 - max[\sum_k p(b_{j,r} > 0|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = k)\tilde{\pi}_k, \sum_k p(b_{j,r} < 0|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = k)\tilde{\pi}_{jk}] \tag{9}$$

1 Testing and Training

In order to determine the optimal number and rank of the covariance matrices, we divide our data set into a training and test data set, each containing 8000 genes.

In the training set, we proceed as above: choosing the top SNP for each of the 8000 genes, creating a list of covariance matrices through deconvolution and grid selection of these top 'training gene-snp' pairs.

Then, within the training data, we similarly choose a random set of gene-snp pairs (restricting our analysis to genes contained in the training set. Again, we choose 20,000 random-gene snp pairs and use the EM algorithm to learn the mixture proportions π from this data set.

We then use the KxL vector of π from the training set to estimate the log likelihood of each data point in the test data set. If our model is 'overfit' to the training data set, than a larger number of covariance matrices may actually decrease the test log-likelihood.

I found that the K=1188 set of covariance matrices containing the Identity, the denoised empirical covariance matrix, rank 5 SFA approximation and rank 3 SVD approximation as well as 5 single-rank SFA factors and the 45 *eqtl.bma.lite* configurations maximized this likelihood.

2 Training and Testing Procedure: Estimating Hierarchical Weights

We wish to choose the model which best maximizes the probability of observing the data set.

Incomplete Data likelihood:

$$L(\boldsymbol{\pi};\hat{\boldsymbol{b}}) = \prod_{j=1}^{J} \sum_{k}^{K} \pi_k P(\hat{\boldsymbol{b_j}}|z_j = k) \tag{10}$$

- To estimate the hierarchical prior weights π_k: compute the likelihood at each each gene snp pair j by evaluating the probability of observing $\hat{\boldsymbol{b}}_{\boldsymbol{j}}$ given that we know the true $\boldsymbol{b_j}$ arises from component k
- Use the EM algorithm to estimate the optimal combination of weights: How often does this particular covariance matrix occur in the data?

We then use these weights to estimate the test set log likelihood.

Results

Discussion

Supporting Information

Acknowledgments

References

1. Devaraju P, Gulati R, Antony PT, Mithun CB, Negi VS. Susceptibility to SLE in South Indian Tamils may be influenced by genetic selection pressure on TLR2

and TLR9 genes. Mol Immunol. 2014 Nov 22. pii: S0161-5890(14)00313-7. doi: 10.1016/j.molimm.2014.11.005

2. Huynen MMTE, Martens P, Hilderlink HBM. The health impacts of globalisation: a conceptual framework. Global Health. 2005;1: 14. Available: http://www.globalizationandhealth.com/content/1/1/14.