Statistical inference of eQTL sharing among many tissues

Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, Matthew Stephens

Statistical inference of eQTL sharing among a high number of tissues

T. Flutre (1,2), X. Wen (3), J. Pritchard (4,5), M. Stephens (1,6)

(1) Dept of Human Genetics, University of Chicago, (2) Dept of Plant Genetics, INRA, (3) Dept of Biostatistics, University of Michigan, (4) Depts of Genetics and Biology, Stanford University, (5) Howard Hughes Medical Institute, (6) Dept of Statistics, University of Chicago

{tflutre,mstephens}@uchicago.edu, xiaoquanw@umich.edu, pritch@stanford.edu

1. Introduction

USING gene expression as intermediate phenotype is now a common approach to interpret associations between genetic variants and organismal phenotypes. Until recently, most studies were performed on a single tissue or cell-type sampled in a hundred individuals. Statistical methods were gradually improved so as to now robustly detect genetic variants associated with changes in gene expression (eQTLs). As typical effect sizes are weak, several recent studies analyzed up to one thousand individuals, showing that most genes have at least one eQTL. However, interpretation of such associations is hampered by the fact that easy-to-sample tissues may often be irrelevant to the etiology of organismal phenotypes of interest. This prompted the NIH to fund the genotype - tissue expression pilot project (GTEx) aiming at building the largest, tissue-wide eQTL data resource to date.

IN prevision of such a data set, we recently developed a statistical framework to detect eQTLs with high power by jointly analyzing multiple tissues, and to reliably infer the proportion of tissue-consistent and tissue-specific eQTLs (Flutre *et al.*, PLoS Genetics 2013). As part of the GTEx consortium, we applied our model on the current data set of 9 tissues from 100-200 individuals. This analysis, done in close collaboration with other groups in the consortium, is planned to be presented in the GTEx session.

HOWEVER, most current methods are unable to efficiently cope with the larger number of tissues that will be available in the future. The reason stems from the use of configurations, binary vectors representing activity patterns of eQTLs among tissues. Indeed, a data set of 20 tissues generates 2^{20} configurations ($> 10^6$). Instead of considering each of them, our improved model learns only the subset of configurations present in the data. Moreover, we dont restrict ourselves anymore to genes expressed in all tissues, by directly analyzing gene expression in terms of read counts. The results obtained with this new model will be presented on the 9 tissues comprising the current GTEx data set. To alleviate the need for permutations in such large-scale studies, we also developed an efficient, yet conservative procedure to control the Bayesian FDR, and evaluate it rigorously by doing permutations.

2. Learning clusters of configurations

LET'S imagine a study totalizing N individuals, genotyped at P common snps, and, for each individual, mRNA expression levels for G genes were measured in S tissues (more generally, subgroups). In Flutre *et al.*, we proposed a statistical framework to analyze all tissues jointly. A gene-snp pair is modeled in each tissue by a linear regression:

$$\mathbf{y}_s = \mu_s \mathbf{1} + \beta_s \mathbf{x}_s + \mathbf{e}_s \text{ with } \mathbf{e}_s \sim N(0, \sigma_s^2 I) \quad (1)$$

Information is borrowed across tissues via the prior $\beta_s \sim \mathcal{N}(\bar{\beta}, \psi^2)$ and $\bar{\beta} \sim \mathcal{N}(0, w^2)$ which allows for heterogeneity in genotype effects among tissues (when $\psi^2 \neq 0$).

IN case the individuals are shared between tissues, partially or not, the errors are also allowed to be correlated between tissues. The genes are assumed to be expressed in all tissues and, beforehand, their expression levels per tissue are tranformed to a standard Normal. Other covariates can be included, such as sex, genotype PCs, inferred expression confounders, etc.

TO test if a snp is an eQTL or not and, if yes, in which tissue(s), we use the notion of *configuration*. For S tissues, there are $J = 2^S$ configurations. Each is an S-dimensional binary vector, $\boldsymbol{\gamma}_j$, which s^{th} element is 1 if the eQTL is active in the s^{th} tissue, and 0 otherwise.

USING the Bayes Factor from Wen & Stephens (arXiv, 2011), we can compute the support in the data for each configuration, BF_j. Then, using the hierarchical model from Flutre *et al.*, we can borrow information across genes to estimate by maximum likelihood the prior probability of each configuration, η_j ("empirical Bayes").

HOWEVER, the space of all configurations increases geometrically with S. Statistically speaking, this means that the estimate of each configuration probability, $\hat{\eta}_j$, will be unreliable. Furthermore, it will be computationally intractable to simply compute all Bayes Factors BF_j.

THEREFORE, instead of explicitly considering each configuration, we can attempt to cluster them into a set of K *types* (K to be determined) while accounting for uncertainty in which gene-snp pairs are eQTLs. Each type has a corresponding S-dimensional vector $\boldsymbol{q}_k$ such that q_{ks} is the probability that an eQTL is active in tissue s given that it is of type k. Moreover we assume that, given a type, the activity of an eQTL in a given subgroup is independent from its activity in other subgroups. Formally, we hence get the following mixture of Bernoulli vectors:

$$p(\boldsymbol{\gamma}_j | \boldsymbol{\pi}, Q) = \sum_{k=1}^{K} \left[\pi_k \left(\prod_{s=1}^{S} q_{ks}^{\gamma_{js}} (1 - q_{ks})^{1-\gamma_{js}} \right) \right] \quad (2)$$

Both $\boldsymbol{\pi}$ and Q can be estimated by maximum likelihood as before, but now the number of parameters is $K \times (S+1)$ instead of 2^S.

3. Assessing the inference via simulations

WE simulated data with $N = 1000$ individuals, $G = 3000$ genes, with one SNP per gene, $\pi_0 = 0.3$ of them being eQTLs, and a realistic expected proportion of variance explained by any individual eQTL, $h = 0.2$. For the moment, we fixed the number of subgroups at $S = 10$ and chose $K = 4$ types. Based on what we know, we choose the first type to correspond to the consistent configuration, such that $\forall s \; q_{1s} \sim \mathcal{U}([0.9; 1])$. In order to make the other types sufficiently different from each other, we borrowed the idea of the "F" model in the analysis of population structure, such that $\forall k > 2$ and $\forall s \; \boldsymbol{q}_k \sim \mathcal{D}((1-F)/0.5F)$ with $p = 0.5$ and $F \in \{0.2, 0.5, 0.8\}$.

FROM $R = 10$ replicates, each with $I = 10$ different random initializations to avoid local maxima, for which we estimate the π_k's and the $\boldsymbol{q}_k$'s, helding all other parameters fixed, we can assess the binary classification of "eGenes" using the posterior of each gene to have an eQTL:

model	TPR	FPR	FDR	FNR
config	0.8989	8.10^{-5}	6.10^{-5}	0.1286
type	0.8632	0.0	0.0	0.1665

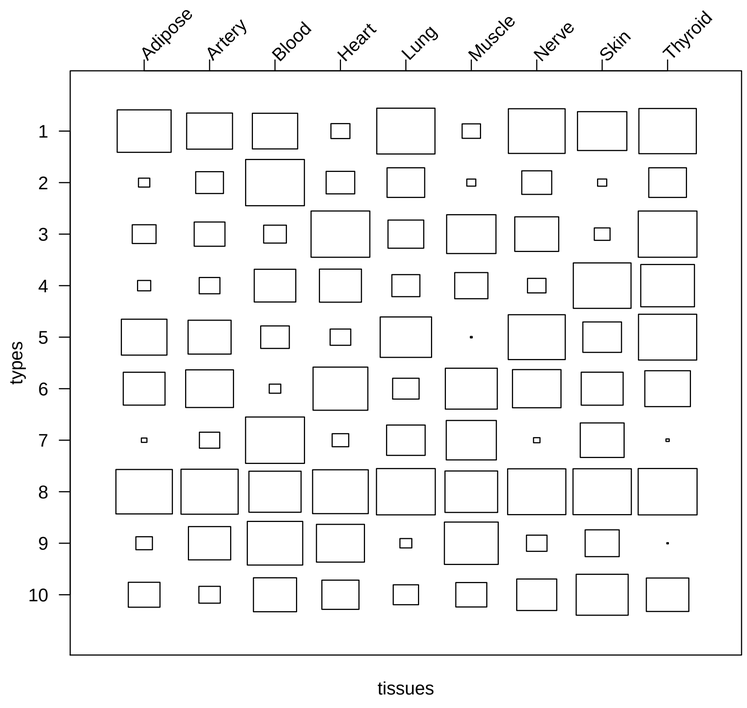
We can also assess the binary classification of eQTL per subgroup using the posterior for a SNP to be active in this subgroup, for instance in subgroup 3:

model	TPR	FPR	FDR	FNR
config	0.9223	0.0031	0.00177	0.1281
type	0.8694	0.06887	0.03927	0.2098

The "type" model hence displays comparable performances as the "configuration" model from Flutre *et al.* (given all the assumptions used in these simulations).

4. Exploratory analysis on GTEx pilot data

AS part of the GTEx consortium, we applied the model of Flutre *et al.* on the 9 tissues from the pilot GTEx project. Some individuals have missing tissues but, overall, expression levels are available for 22,000 genes in 100-200 individuals. Focusing on $\pm$100 Kb around the transcription start site corresponds to 10.5 millions gene-snp pairs. Using the EBF procedure from Wen to avoid permutations, we estimated $\hat{\pi}_0 \approx 0.55$. This procedure is very quick, yet it can be shown to be conservative (article soon on arXiv). Using this estimate to control the Bayesian FDR, we can confidently call 10,030 "eGenes".



APPLYING the "type" model on this data set with $K = 10$ gives the matrix of estimated type probabilities $\hat{q}_{ks}$ shown on the left. It also gives a very similar estimate $\hat{\pi}_0 \approx 0.58$ and calls 7,477 eGenes. This analysis is exploratory (e.g. choice of K), but shows promising results for the future GTEx data set potentially collecting 30-50 tissues in 900-1,000 individuals.

5. Perspectives

THIS project is still preliminary and numerous improvements are planned: (i) take into account an average effect size, as in the initial model, to improve power; (ii) assess how well K is chosen (plateau of observed log-likelihood); (iii) explore a higher number of subgroups (say, $S = 30$); (iv) assess the count-data likelihood on configuration and type estimates; (v) integrate the notion of "types" into the multi-eQTL model from Wen (arXiv); (vi) take into account genomic annotations in the prior for a SNP to be an eQTL.

6. Software and Acknowledgments

THE `eQtlBma` package from Flutre *et al.* is available on Github. The new functionalities introduced in this poster will be part of it in the future. The GTEx data analysis was greatly eased thanks to all the people in the GTEx consortium. This work was completed in part with resources provided by the University of Chicago Research Computing Center, and is supported by the NIH (grant MH090951). T. Flutre is partially funded by the INRA.