Supplementary Material

Christine Peterson, Marina Bogomolov, Yoav Benjamini, Chiara Sabatti

TreeQTL: hierarchical error control for eQTL findings

In this supplement, we provide additional details on the hierarchical testing procedure, the functions available in the `TreeQTL` R package, and the example application to whole blood data from the pilot phase of the Genotype-Tissue Expression (GTEx) project.

Hierarchical testing procedure

The hierarchical testing procedure implemented in `TreeQTL` corrects for multiple comparisons in a way that respects the structure of expression quantitative trait (eQTL) analysis. In particular, in the eQTL setting, there is heterogeneity in the proportion of non-nulls across different classes of hypotheses. Firstly, there is a substantial difference in the proportion of non-null hypotheses among local vs. distal hypotheses, given that local regulation is much more common. Secondly, among the hypotheses addressing distal regulation, SNPs with any distal effects are likely to play a regulatory role for multiple genes. Since the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) for controlling the false discovery rate (FDR) is adaptive to the amount of signal in the data, applying BH the entire collection of hypotheses is sub-optimal. In addition, as demonstrated in Peterson *et al.* (2015), this strategy fails to control the FDR for the discovery of eSNPs or eGenes, key points of interest in the reporting of eQTL findings.

These observations lead us to organize the eQTL hypotheses in the tree-like structure depicted in Figure 1. In this three layer tree, the upper level branches are local and distal regulation; the middle level contains groups indexed by SNP or gene; and the leaf nodes are the individual SNP×gene hypotheses. The hypothesis L_{vg} corresponds to the null hypothesis that there is no local association between variant v and gene g; similarly, the hypothesis D_{vg} corresponds to the null hypothesis that there is no distal association between variant v and gene g. When using `TreeQTL`, the user may specify whether the identification of eGenes of eSNPs is of primary interest; however, based on the heterogeneity among the hypotheses described above, we recommend that distal analysis should be organized around the discovery of eSNPs.

In Figure 2 we illustrate the correspondence between the hypotheses shown in Figure 1 and their associated p-values. P-values for the tests are defined starting from the leaf hypotheses. The p-values for the Level 2 hypotheses L_{vg} and D_{vg} for association between SNP v and gene g are taken as input to `TreeQTL` and may be obtained via `MatrixEQTL` (Shabalin 2012) or other appropriate software. We assume that the input p-values are valid in the sense that they account for sources of confounding such as population stratification, family structure, or batch effects; we are agnostic, however, as to how they are obtained. The p-values for the Level 1 hypotheses $L_{\bullet g}$ and $H_{v\bullet}$ are then computed using the Simes' rule; `TreeQTL` also allows the option to submit alternative p-values for the Level 1 hypotheses (such as those obtained view permutation). The Simes' p-value for $L_{\bullet g}$ is defined as

$$p_{L\bullet g} = \min_{v=1,\ldots,m_g} \frac{m_g p_{L(v)g}}{v},$$

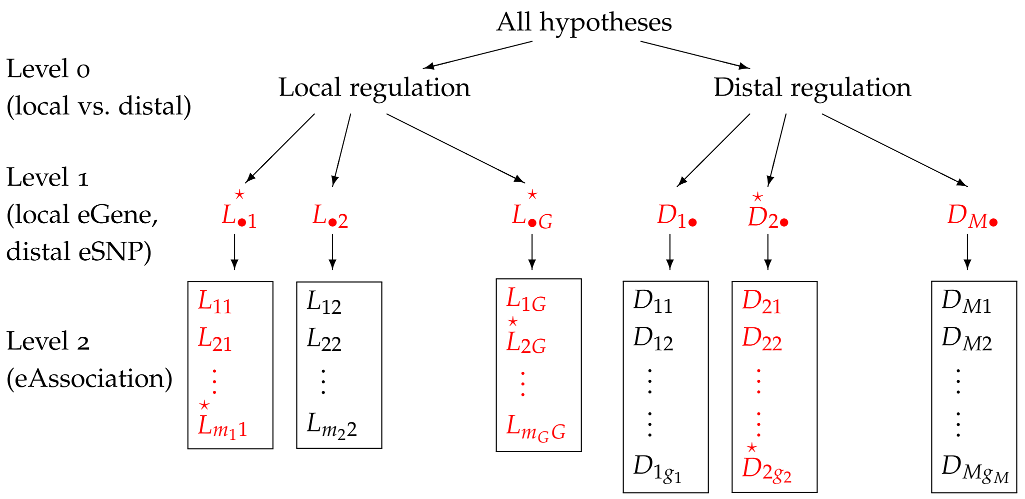

Figure 1: Organization of eQTL hypotheses in TreeQTL. Local regulation hypotheses have been grouped by gene and distal regulation hypotheses are grouped by SNP. Tested hypotheses are colored in red, and rejected hypotheses indicated with a star.

where $p_{L(v)g}$ is the vth smallest p-value among the local hypotheses referencing gene g, and m_g is the number of variants in the local region for gene g; the Simes' p-value for the other Level 1 hypotheses may be defined similarly. We choose the Simes' p-value (rather than alternatives such as Fisher's combined p-value) since it has several attractive properties in this context. Specifically, it is robust to dependence among the p-values it depends on (Benjamini and Heller, 2008). In addition, use of the Simes' rule to compute the p-values for the level 1 hypotheses guarantees that if a Level 1 hypothesis is rejected (i.e. an eSNP or eGene is discovered), then at least one eAssociation involving that eSNP or eGene will be discovered as well. Finally, we consider the Level 0 hypotheses regarding whether there is any local or any distal regulation; since we can safely reject these null hypotheses in any realistic setting, we do not explicitly compute $p_{L\bullet\bullet}$ and $p_{D\bullet\bullet}$.

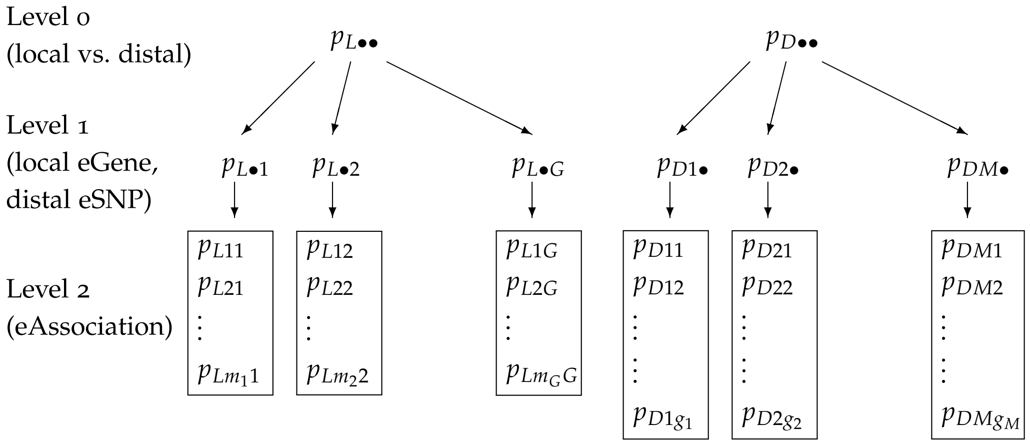

Figure 2: Hierarchical structure of the hypotheses tested

Testing for local regulation proceeds starting from Level 1 as follows:

1. Test the collection of Level 1 hypotheses for local regulation. This may be done using either the Benjamini-Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001) or

the BH procedure with target FDR level q_1. The default procedure is BY as it is more robust to dependence. A more conservative option is also available of targeting the family-wise error rate (FWER) at level q_1 via the Bonferroni procedure. (See next section for a discussion). Let R_L (R_D) represent the number of rejected local hypotheses from this stage.

2. For each rejected Level 1 hypothesis regarding local regulation, test the corresponding set of Level 2 hypotheses using BH with adjusted target level $R_L q_2 / G$ (for the selection of eGenes) or $R_L q_2 / M$ (for the selection of eSNPs).

The testing procedure for distal regulation is defined similarly. As shown in Benjamini and Bogomolov (2014), given some assumptions on the dependence among hypotheses, this procedure guarantees that the error rate for level 1 discoveries is controlled to level q_1, and that the expected average proportion of false eAssociation discoveries across the selected eSNPs or eGenes will be controlled to level q_2. Peterson *et al.* (2015) demonstrates via simulation studies that these error rates are in fact controlled under the common types of dependence found in genome-wise association studies.

There is an exact correspondence between the step-wise testing procedure described above and the functions available in the `TreeQTL` R package. Specifically, the functions `get_eGenes()` and `get_eSNPs()` can be used to identify the Level 1 rejections for either local or distal regulation. Given the return value from either of these functions, `get_eAssociations()` can then be called to obtain a full listing of the Level 2 discoveries. Note that `TreeQTL` must be run separately for local and distal analysis and expects the local and distal association SNP-gene association p-values to be stored in separate input files; these input files can be produced in a single run of `MatrixEQTL`, or constructed manually. For additional details and example code, please see the `TreeQTL` package documentation at http://bioinformatics.org/treeqtl/TreeQTL.pdf.

Illustration on GTEx data

To demonstrate the application of `TreeQTL` to real data, we used `TreeQTL` to analyze whole-blood data from the pilot phase of the GTEx project (Ardlie *et al.*, 2015). In this data set, genotype data at 6,820,472 SNPs and expression levels for 30,115 genes are available across 156 subjects. Local associations correspond to SNP-gene pairs where the SNP is within 1Mb of the transcription start site (TSS) of the gene; all other SNP-gene associations are considered distal. This definition results in approximately 142 million local tests (reflecting an average of 21 genes in the local region for each SNP) and 205 billion distal tests. Following the steps in Ardlie *et al.* (2015), p-values for local association were obtained by applying Matrix eQTL to normalized gene expression, adjusting for both known and unknown technical covariates through the inclusion of gender, 3 genotype principal components, and 15 PEER factors as covariates in the linear model. We followed the same procedure to obtain p-values for distal association, with the caveat that distal analysis is underpowered in this setting and is not attempted in Ardlie *et al.* (2015).

Here we compare the results from three different approaches: BH applied to local and distal hypotheses separately at level $q = 0.01$; `TreeQTL` defining Level 1 discoveries as eSNPs and using BH in step 1 at level $q_1 = q_2 = 0.01$; and `TreeQTL`, defining Level 1 discoveries as eSNP and using BY in step 1 at level $q_1 = q_2 = 0.01$. The number of discoveries under these methods is given in Table 1. When comparing the results of TreeQTL using BH to those from applying BH separately to all local vs. distal hypotheses, the total number of discoveries is slightly smaller and the selection of SNPs is somewhat more stringent using the hierarchical procedure; in exchange for controlling the FDR across eSNPs and average proportion of false eAssociations involving the selected eSNPs, we pay a modest price in terms of loss of overall power. The BY procedure results in a much more conservative number of eSNP discoveries.

	Local		Distal	
	eSNPs	eAssoc	eSNPs	eAssoc
BH separate	163,263	263,417	179,625	216,683
TreeQTL BH	136,609	229,821	164,860	216,933
TreeQTL BY	90,143	175,889	41,216	55,034

Table 1: Number of rejected hypotheses under the following error control approaches: BH applied to local and distal hypotheses separately at level $q = 0.01$; TreeQTL using BH in step 1 at level $q_1 = q_2 = 0.01$; and TreeQTL using BY in step 1 at level $q_1 = q_2 = 0.01$

In interpreting these eSNP results, it is important to recall the fact that dependence exists between SNPs (linkage disequilibrium) and this generates both a correlation across the test statistics even when the null hypotheses are true, and a correlation in the signal. That is, multiple SNPs neighboring a causal variant will have non zero association with the expression of a gene: for all of these SNPs one would correctly reject the null hypotheses, even if they all correspond to only one true biological discovery. We note that, on the one hand, BH, while guaranteed to control FDR under independence or special kinds of dependence, typically can handle the dependence between tests statistics due to linkage disequilibrium, so that in this case it is not necessary to resort to the more conservative BY. On the other hand, one has to be aware that the number of eSNP discoveries is going to be larger than the number of true causal variants: the signal from any true association will percolate to a number of nearby SNPs. One can argue that SNPs, especially in studies with high density genotyping as GTEx, do not represent the correct resolution to count discoveries: rather, one should consider clumps of SNPs with correlated signal as one unit of discovery. This has been argued, for example in Siegmund *et al.* (2011). Neither BH or BY simply applied to the SNP level tests would offer control of the FDR for discoveries so defined. Research on procedures that offer this control is on going and we plan to update TreeQTL to include these once solid results are available.

Partially to address this discrepancy between the number of associated SNPs and the number of underlying causal variants, the eQTL field has often focused on the notion of eGenes. This is the case of the GTEx pilot paper (Ardlie *et al.*, 2015), which utilizes a permutation-based approach to identify eGenes. Specifically, for each gene, the minimal association p-value per gene is computed; the sample labels on the expression data, gender, and PEER factors are then permuted, and the minimal association p-value is recomputed. The empirical p-value is taken to be the proportion of times the permuted p-value is less than the nominal p-value; since at most 10,000 permutations are performed, the smallest possible empirical p-value is assumed to be 0.00009999 i.e. $< 1/10{,}000$. This strategy is intended to account for the fact that multiple SNPs which are likely to be in LD are being tested for association to each gene. The final set of eGenes is then taken to be those with Storey q-value ≤ 0.05. For the whole blood pilot data set, 2,052 eGenes are identified as having Storey q-value $\leq$ 0.05 based on the empirical p-values. For comparison, 1,379 eGenes are identified when BH is applied at level 0.01 to the same set of empirical p-values.

This approach is compatible with `TreeQTL` in that the empirical p-values could be taken as input to the function `get_eGenes()`; the individual eAssociation hypotheses could then be selected on the basis of the Matrix eQTL p-values using the hierarchical procedure to adjust for selection bias. If instead we compute the gene-level p-values within `TreeQTL` using the Simes' rule and use BH for selection targeting level 0.01, we obtain 1,588 eGenes, 1,453 of which were also reported as eGenes in Ardlie *et al.* Due to the limited resolution of the permutation-based approach, direct comparison of the gene-level p-values is not possible; in fact, 1,145 of the 2,052 reported eGenes have a reported p-value of 0.00009999.

References

Ardlie,K.G. *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multi-tissue gene regulation in humans. *Science*, **8**, 648–660.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRRS B*, **57**, 289–300.

Benjamini,Y. and Heller,R. (2008) Screening for partial conjunction hypotheses. *Biometrics*, **64**, 1215–1222.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Peterson,C.B. *et al.* (2016) Many phenotypes without many false discoveries: error controlling strategies for multi-trait association studies. *Genetic Epidemiology*, **40**, 45–56.

Shabalin,A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

Siegmund, D., Yakir, B. and Zhang, N. (2011) The false discovery rate for scan statistics. *Biometrika*, **98**, 979–985.

Simes,R. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.