Bioinformatics, 32, 2016, i156–i163
doi: 10.1093/bioinformatics/btw272
ISMB 2016

OXFORD

Using genomic annotations increases statistical power to detect eGenes

Dat Duong[1], Jennifer Zou[1], Farhad Hormozdiari[1], Jae Hoon Sul[4], Jason Ernst[1,2], Buhm Han[5,*] and Eleazar Eskin[1,3,*]

[1]Department of Computer Science, [2]Department of Biological Chemistry, [3]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA, [4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA and [5]Department of Convergence Medicine, University of Ulsan College of Medicine & Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

*To whom correspondence should be addressed.

Abstract

Motivation: Expression quantitative trait loci (eQTLs) are genetic variants that affect gene expression. In eQTL studies, one important task is to find eGenes or genes whose expressions are associated with at least one eQTL. The standard statistical method to determine whether a gene is an eGene requires association testing at all nearby variants and the permutation test to correct for multiple testing. The standard method however does not consider genomic annotation of the variants. In practice, variants near gene transcription start sites (TSSs) or certain histone modifications are likely to regulate gene expression. In this article, we introduce a novel eGene detection method that considers this empirical evidence and thereby increases the statistical power.

Results: We applied our method to the liver Genotype-Tissue Expression (GTEx) data using distance from TSSs, DNase hypersensitivity sites, and six histone modifications as the genomic annotations for the variants. Each of these annotations helped us detected more candidate eGenes. Distance from TSS appears to be the most important annotation; specifically, using this annotation, our method discovered 50% more candidate eGenes than the standard permutation method.

Contact: buhm.han@amc.seoul.kr or eeskin@cs.ucla.edu

1 Introduction

Many studies over the past decade examined the contribution of genetic loci to phenotypic variation in complex traits. Genetic loci that are associated with gene expression are called expression quantitative trait loci (eQTLs) (Brem and Kruglyak, 2005; Gilad *et al.*, 2008; The GTEx Consortium, 2015). One important task in eQTL studies is to find eGenes or genes whose expressions are associated with at least one genetic variant. The standard method to determine whether a gene is an eGene requires association testing at all variants near the gene (cis-variants) and the permutation test to correct for multiple testing. The permutation test is the gold standard for multiple-testing correction, as it properly accounts for the linkage disequilibrium (LD) structure in the genome.

However, this standard method does not consider which variants are more likely to regulate gene expression. In order to better detect eGenes, we can increase the statistical power of this standard method by using annotation data. In practice, regulatory variants found near the transcription start sites (TSSs) and certain histone modifications are more likely to be associated with gene expression (van de Geijn *et al.*, 2015). Additionally, recent large-scale genomics studies have annotated regions of the genome that are likely to alter gene expression in individuals (Ernst and Kellis, 2015; The Roadmap Epigenomics Mapping Consortium, 2015). For example, almost 80% of the chip-based heritability of disease risk for 11 human diseases examined in the Wellcome Trust Case Control Consortium (WTCCC) can be explained by genome variation in DNase I hypersensitivity sites. These variations are likely to regulate chromatin accessibility and thus transcription (Gusev *et al.*, 2014). These genomic annotations for the variants can be used to increase the power to detect eGenes.

Although several methods were recently developed to address challenges in multiple-testing correction in eQTL studies, these methods do not improve statistical power in comparison to the standard method. Sul *et al.* (2015) improved the runtime of the standard permutation test by replacing the permutation procedure with sampling from the multivariate normal distribution (Mvn). Davis *et al.* (2016) further improved this runtime by estimating the effective number of tests based on the eigen-decomposition of the genotype correlation matrix. These methods aim to reduce runtime but do not attempt to increase statistical power of the standard

approach. Therefore, these methods are not capable of detecting more eGenes.

In this article, we introduce a new method for discovering eGenes (Func-eGene) that uses genomic annotation of the variants to increase statistical power. Although gene expression can be affected by trans-variants (Bryois *et al.*, 2014), in this article, we focus on methods to detect cis-acting eQTLs. We rely on the multi-threshold association test that specifies different significance thresholds for the variants when correcting for multiple testing (Darnell *et al.*, 2012; Eskin, 2008). In this multi-threshold association study mindset, we can assign less stringent significance thresholds to variants that have a high propensity to contribute to gene expression, thereby increasing power. If an appropriate prior is provided, this multi-threshold association method has a closed-form solution that guarantees the best statistical power for the association test (Darnell *et al.*, 2012). However, there are two key difficulties we encounter when directly applying this multi-threshold association study method to discover eGenes. First, the multi-threshold association test depends on a permutation test to correct for LD. This permutation test is slow when applying to a large dataset. Second, we rarely know of an appropriate prior based on the annotation for genetic variants under study.

Our new method Func-eGene avoids these difficulties. To reduce runtime, we replace the permutation test with the sampling procedure in Sul *et al.* (2015). To find an appropriate prior, we do a grid search over possible sets of scores assigned to annotation categories. The goal of our search is to find the set of scores that maximizes the number of eGenes. To avoid data re-use and over-fitting, we use a cross-validation strategy.

We applied our method to the liver dataset from the Genotype Tissue Expression (GTEx) Consortium. First, we used the distance from the TSS as a genomic annotation because variants near the TSS are likely to affect gene expression. Using this annotation alone, our method Func-eGene increased the number of candidate eGenes by 50% compared with the standard method. Then, we added DNase hypersensitivity sites as a second genomic annotation. The TSS and DNase annotations together did not discover more candidate eGenes than the TSS annotation alone. Third, we separately applied the binding sites for histones H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K9ac and H3K9me3 as genomic annotations. These histone marks found more candidate eGenes than the standard method but less than the number reported by using the TSS annotation alone. Distance from TSS appears to create the most informative prior for detecting eGenes.

2 Methods

2.1 Standard method to detect eGenes

An eGene is a gene that has an associated eQTL (Sul *et al.* 2015). Let s be a test statistic such as a t-distributed statistic from Pearson or Spearman correlation between a genetic variant and gene expression. Define $F(s)$ as its cumulative density function. Suppose α_s is the desired false-positive rate. In a two-tailed hypothesis test, assuming that the distribution is symmetric, $2(1-F(|s|))$ is the P-value, and $\pm F^{-1}(1-\alpha_s/2)$ defines the rejection region for s. We use $(x)_{p\times q}$ to specify a matrix x of dimension $p \times q$. Parentheses and subscripts are omitted whenever the context is clear.

2.1.1 Association test

Suppose that there are N individuals. In the simplest scenario, which considers only one variant and one gene y, the hypothesis test tests whether gene expression $(e_y)_{N\times 1}$ of y is independent of the variant genotype $g_{N\times 1}$. If they are independent, the variant is not an eQTL and y is not an eGene. The null hypothesis H_0 is that y is not an eGene, and the alternative hypothesis H_1 is that y is an eGene. To conduct this single hypothesis test, the standard method assumes a linear model

$$e_y = Xb + g\beta + \epsilon \tag{1}$$

In Equation 1, the design matrix $X_{N\times P}$ contains P fixed effects (i.e. gender, ethnicity, age, etc.). The vector $b_{P\times 1}$ is their regression coefficients. β is the regression coefficient of the variant genotypes. $\epsilon_{N\times 1}$ is a vector of independent sampling errors that is normally distributed ($\epsilon_{N\times 1} \sim N(\boldsymbol{0}, I\sigma^2)$). In Equation 1, linear regression is used to estimate $\widehat{\beta}$ and its variance $\widehat{\sigma}^2_{\widehat{\beta}}$.

Our test statistic s is the normalized $\widehat{\beta}$ ($s = \widehat{\beta}/\widehat{\sigma}_{\widehat{\beta}}$). Under the null hypothesis, s follows a t-distribution with $N-P-1$ degrees of freedom. If we suppose N is large, then $F(s) \approx \Phi_\mu(s)$ where Φ_μ is a normal cumulative density with mean μ and variance one. This mean μ is also known as the z-score non-centrality parameter. Our null and alternative hypotheses can then be written as

$$H_0: \text{ Gene } y \text{ is not an eGene} \leftrightarrow H_0: \mu = 0$$

$$H_1: \text{ Gene } y \text{ is an eGene} \leftrightarrow H_1: \mu = w \text{ where } w \neq 0$$

H_0 is rejected if the P-value is less than α. This P-value is named eGene P-value (Sul *et al.*, 2015).

2.1.2 Multi-association test

In a more common scenario, many variants in-cis with gene y are tested. In this case, the test consists of M univariate association tests. The hypothesis test tests whether the expression $(e_y)_{N\times 1}$ of y is independent of all variant genotypes $(g_i)_{N\times 1}$ $(i = 1 \ldots M)$. As before, one assumes

$$e_y = Xb + g_i\beta_i + \epsilon_i \quad i = 1 \ldots M \tag{2}$$

In Equation 2, $X_{N\times P}$ contains P fixed effects, and $b_{P\times 1}$ is their regression coefficients. β_i is the regression coefficient for the genotypes of variant i. $(\epsilon_i)_{N\times 1}$ is a vector of independent sampling errors, and follows $(\epsilon_i)_{N\times 1} \sim N(\boldsymbol{0}, I\sigma^2)$. In Equation 2, linear regression is again used to estimate $\widehat{\beta}_i$ and its variance $\widehat{\sigma}^2_{\widehat{\beta}_i}$.

Our test statistic s_i is the normalized $\widehat{\beta}_i$ ($s_i = \widehat{\beta}_i/\widehat{\sigma}_{\widehat{\beta}_i}$). Let μ_i be the expected value of each test statistic s_i. We write the hypothesis as

$$\begin{aligned} &H_0: \text{ Gene } y \text{ is not an eGene} \leftrightarrow H_0: \mu_i = 0 \text{ for all } i \\ &H_1: \text{ Gene } y \text{ is an eGene } \leftrightarrow H_1: \mu_i = w_i \\ &\qquad\qquad \text{where } w_i \neq 0 \text{ for some } i \in \{1 \ldots M\} \end{aligned} \tag{3}$$

We then compute the P-value at each variant i and reject H_0 if their minimum P-value $\mathscr{P}$ is less than α_c (Sul *et al.*, 2015). α_c is the false-positive rate adjusted for multiple testing. For example, if Bonferonni correction is applied, $\alpha_c = \alpha/M$. Bonferonni correction is conservative because it ignores linkage-disequilibrium among the variants. The permutation test is thus the gold standard method (Sul *et al.*, 2015).

2.2 Functional annotation-based multi-threshold eGene (Func-eGene)

The standard method applies an identical univariate association test to each variant and uses the minimum P-value as a test statistic. This is equivalent to assigning to all variants a uniform prior of being associated with the gene expression. However, we often have

annotations that specify whether the variants are in some regulatory regions of the genome. We developed a new method named Func-eGene that considers this evidence to increase statistical power to find candidate eGenes.

2.2.1 Multi-threshold association test

Our Func-eGene is built upon the multi-threshold association test method that assigns different significance thresholds to different hypothesis tests (Darnell *et al.*, 2012; Eskin, 2008). We briefly describe their method here, assuming that we have data about the relative importance of the genetic variants in our study. Let this information about the M variants be given as $c_1 \dots c_M$, where $\sum_{i=1}^{M} c_i = 1$ and $c_i \geq 0$ for all i. For example, regulatory variants can be assigned higher c_i than those which are not. We assume that $c_1 \dots c_M$ are given beforehand. Later, we drop this assumption and determine an appropriate $c_1 \dots c_M$ from the data.

Darnell *et al.* (2012) and Eskin (2008) use this data to create a non-uniform prior in the hypothesis test by giving the M observed statistics $s_1 \dots s_M$ their own thresholds $t_1 \dots t_M$, which are functions of $c_1 \dots c_M$. These t_i must be corrected for multiple testing by using the constraint $\sum_{i=1}^{M} t_i = \alpha$, where α is the genome-wide false-positive rate. This constraint holds only if the variants are independent. Later, we remove this requirement and properly account for LD. We then maximize the statistical power of the hypothesis test, which is a function of t_i.

Statistical power is defined as the probability of an observed statistic being significant when the alternative hypothesis is true. In the simplest case, there is only one variant i and the gene y. The power of the two-tail hypothesis test denoted as $P_s(t_i, \mu_i)$ is the probability that $|s_i| > F^{-1}(1 - t_i/2)$ when the true mean of s_i is not zero. Suppose N is large so that $P_s(t_i, \mu_i)$ can be approximated using the standard normal cumulative density Φ.

$$P_s(t_i, \mu_i) = \mathbb{P}(|s_i| > F^{-1}(1 - t_i/2)) \tag{4a}$$

$$= \Phi(\Phi^{-1}(t_i/2) - \mu_i) + 1 - \Phi(\Phi^{-1}(1 - t_i/2) - \mu_i) \tag{4b}$$

In a more common scenario, one considers M variants in-cis with gene y. The statistical power to detect y being an eGene denoted as $P(t_1 \dots t_M)$ is the weighted average over M variants, $P(t_1 \dots t_M) = \sum_{i=1}^{M} c_i P_s(t_i, \mu_i)$. For a more detailed discussion of this definition, see Eskin (2008) and Rubin *et al.* (2006).

Darnell *et al.* (2012) and Eskin (2008) find $t_1 \dots t_M$ so that the statistical power $P(t_1 \dots t_M)$ is maximized under the constraint $\sum_{i=1}^{M} t_i = \alpha$ and $t_i \geq 0$ for all i. This solution is obtained by taking the gradient of the Lagrangian $\mathcal{L}(\ell, t_1 \dots t_M)$ with respect to the Lagrangian multiplier ℓ, and the unknown variables $t_1 \dots t_M$.

Optimal solution is achieved when $\nabla_{t_i} L(\ell, t_1 \dots t_M) = \nabla_{t_j} L(\ell, t_1 \dots t_M)$ for all $i, j \in \{1 \dots M\}$ (Eskin, 2008). Moreover, this equality has a closed form (Darnell *et al.*, 2012)

$$\frac{c_i\left(\phi_{\mu_i}(\Phi^{-1}(t_i/2)) + \phi_{-\mu_i}(\Phi^{-1}(t_i/2))\right)/2}{\phi_0(\Phi^{-1}(t_i/2))} = \frac{c_j\left(\phi_{\mu_j}(\Phi^{-1}(t_j/2)) + \phi_{-\mu_j}(\Phi^{-1}(t_j/2))\right)/2}{\phi_0(\Phi^{-1}(t_j/2))} \quad \forall i, j \in \{1 \dots M\} \tag{5}$$

The symbol ϕ_z is the probability density function of a normal distribution having mean z and variance one. $\Phi^{-1}(w)$ is the quantile of w under a normal distribution of mean zero and variance one.

Once the observed statistics $s_1 \dots s_M$ are estimated, we compare their P-values $p_i = 2(1 - F(|s_i|))$ to $t_1 \dots t_M$. If $p_i < t_i$, then variant i is an eQTL. If at least one variant is an eQTL, then the gene is an eGene.

This method can be used to calculate the multiple-testing-corrected P-values $p_1^* \dots p_M^*$. In fact, finding per-marker significance thresholds and computing the corrected P-value are two related tasks in multiple-testing correction. Briefly, to obtain the corrected P-value p_i^*, we begin with a small α^*, find optimal $t_1^* \dots t_m^*$ with constraint $\sum_{i=1}^{M} t_i^* = \alpha^*$ in which case $t_i^* < p_i$ due to small α^*. We repeat this analysis while increasing α^* until $t_i^* = p_i$. α^* will be our corrected P-value p_i^*. If $p_i^* < \alpha$, then variant i is an eQTL. If any of the variants is an eQTL, then the gene is an eGene. The multiple-testing-corrected eGene P-value becomes

$$p^*_{\text{eGene}} = \min\{p_i^*\}_{i=1}^{M} \tag{6}$$

Comparing p_i against t_i and comparing p_i^* against α give identical eQTLs and eGenes. These are two different viewpoints of the same multiple-testing correction.

2.2.2 LD-corrected *P*-value

When LD among the variants is unignorable, corrected P-values p_i^* and eGene P-value p^*_{eGene} violate the independence assumption and become conservative. To avoid this, Darnell *et al.* (2012) and Eskin (2008) suggested using a permutation test to compute p^*_{eGene}.

Because these studies did not describe in detail how a permutation test is done, we explain the procedure here. We do one permutation by permuting the expression measurements among the individuals while keeping their genotype data unchanged so that LD is retained. Leaving the LD intact keeps the correlation between the genotypes of the individuals which then retains the correlation of the test statistics. Suppose that we do such permutation B times. In the j-th permutation, we find M corrected P-values $p^*_{i,j}$ and their eGene P-value as $p^*_{\text{eGene},j} = \min\{p^*_{i,j}\}_{i=1}^{M}$ Let $p^*_{\text{eGene},obs}$ be the eGene P-value in the observed data. Define the LD-corrected eGene P-value as

$$p^*_{\text{eGene,LD-corrected}} = \frac{\sum_{j=1}^{B} \mathbb{1}(p^*_{\text{eGene},obs} \geq p^*_{\text{eGene},j})}{B} \tag{7}$$

where $\mathbb{1}$ is an indicator function. This LD-corrected eGene P-value is not conservative and has correct false-positive rate. This permutation however is time-consuming. In each permutation, we need to find the corrected P-values p_i^* which requires a search for α^* as described above. Repeating this search B times makes the permutation test very time-consuming.

2.2.3 LR-based permutation test

To speed up the permutation test, Func-eGene uses the likelihood ratio (LR). The permutation using P-values in Darnell *et al.* (2012) and Eskin (2008) is slow because every permutation finds the corrected P-value p_i^*. Func-eGene uses a test statistic that does not require p_i^*. To do this, we interpret Equation 5 as a LR multiplied by a prior probability. Define $g_i(s_i)$ such that

$$g_i(s_i) = \frac{c_i\left(\phi_{\mu_i}(s_i) + \phi_{-\mu_i}(s_i)\right)/2}{\phi_0(s_i)} \tag{8}$$

Equation 8 becomes a LR evaluated at s_i where $H_0 : \mathbb{E}(s_i) = 0$ and H_1 is an average of two two-tail hypotheses $H_1 : \mathbb{E}(s_i) = \mu_i$ and $H_1 : \mathbb{E}(s_i) = -\mu_i$.

Here $g_i(s_i)$ is a monotonic increasing in $|s_i|$. The key concept is that we can replace p_i^* in Equation 6 with $g_i(s_i)$. Define the eGene LR to be

$$g_{\text{eGene}} = \max\{g_i(s_i)\}_{i=1}^{M} \tag{9}$$

Let $g_{\text{eGene},obs}$ be the observed eGene LR in the data computed as in Equation 9 using the observed test statistics $s_1 \dots s_M$. Define $g_{\text{eGene},j}$ as the eGene LR computed as in Equation 9 using the test statistics $s_{1,j} \dots s_{M,j}$ in the j-th permutation. The LD-corrected P-value becomes

$$p^*_{\text{eGene,LD-corrected}} = \frac{\sum_{j=1}^{B} \mathbb{1}(g_{\text{eGene},obs} \leq g_{\text{eGene},j})}{B} \tag{10}$$

and is equivalent to Equation 7. By using LR instead of P-value, we avoid finding p^* and make the permutation test faster.

2.2.4 LR-based Mvn sampling

Even with the LR-based permuation test, the method's runtime depends on the number of individuals. To overcome this problem, we observe that Equation 8 is a simple function in s_i. Applying the method in Sul *et al.* (2015), we use the Mvn density instead of the permutation test. We sample $s^0_1 \dots s^0_M$ from a null density $\text{Mvn}(\mathbf{0}, \Sigma)$. $\Sigma_{M\times M}$ measures the LD among the M variants, and is computed as $\frac{1}{N} G^\top G$ where $G_{N\times M}$ is the genotype data. When there are few individuals, the null density in the permutation test and one formed by $s^0_1 \dots s^0_M$ are mismatched. Following Sul *et al.* (2015), we correct $s^0_1 \dots s^0_M$ using the variant minor allele frequencies and the sample size. The corrected $s^0_1 \dots s^0_M$ are used to compute $g_{\text{eGene},j}$ in Equation 10.

2.2.5 Finding appropriate priors using genomic annotations

We demonstrated that Func-eGene can maximize statistical power and approximate eGene P-values when the functional priors are specified beforehand. However, these priors are hardly known a priori. In this section, we introduce a heuristic data-adaptive procedure to determine an appropriate prior that can yield the most number of candidate eGenes. Suppose we categorize the cis-variants using J different annotations. Each annotation j is given a score b_j. A variant belonging to j inherits its score. Define $k_{ij} \in \{0, 1\}$ so that $k_{ij} = 1$ if the variant i belongs to j. The score u_i at a variant i is defined as

$$u_i = \sum_{j=1}^{J} b_j k_{ij} \tag{11}$$

The normalized prior c_i at variant i is

$$c_i = \frac{u_i}{\sum_{m=1}^{M} u_m} \tag{12}$$

where M is the number of variants. Equation 11 assumes that the functional annotations behave in an additive manner. It is possible to include interaction terms among the annotations, and the optimization procedure below remains applicable. Equations 11 and 12 imply that Equations 8, 9 and 10 are functions of the annotation score $\mathbf{b} = (b_1 \dots b_J)$.

Using these priors in Func-eGene, we calculate Equation 10. It is important to see that Equation 8 is the product of c_i and the function $h(s_i) = \frac{\phi_{\mu_i}(s_i) + \phi_{-\mu_i}(s_i)}{2\phi_0(s_i)}$. Thus, we compute $h(s_i)$ from the observed data only once, and then use them when evaluating eGene LR at each possible $\mathbf{b}$ to get $p^*_{\text{eGene,LD-corrected}}$.

Our objective is to determine the optimal score $\mathbf{b}^* = (b^*_1 \dots b^*_J)$ for the annotations that yield the most number of candidate eGenes. One immediate but impractical solution is to search all possible $\mathbf{b}$. To do this, Func-eGene must be run for each $\mathbf{b}$ to get the threshold for the observed eGene LR in Equation 9, which corresponds to the specified significance threshold α. This threshold is some upper α quantile under a null density of g_{eGene}. This quantile depends on $\mathbf{b}$. Let $Q_y(\mathbf{b}^{(k)})$ be the quantile threshold of g_{eGene} at gene y, using some k-th choice of $\mathbf{b}$ denoted as $\mathbf{b}^{(k)}$. If $g_{\text{eGene}} > Q_y(\mathbf{b}^{(k)})$, then y is an eGene. Using these quantile thresholds, the number of eGenes can be estimated for a choice $\mathbf{b}^{(k)}$.

Running Func-eGene for all $\mathbf{b}$ is time demanding. Thus, for finding a good choice of $\mathbf{b}$, we use the following procedure. We aim to approximate quantile thresholds of all $\mathbf{b}$, while implementing Func-eGene only once. One can pick a starting $\mathbf{b}^{(0)}$, compute $Q_y(\mathbf{b}^{(0)})$ for all genes y in the data. In each subsequent choice k, find $Q_y(\mathbf{b}^{(k)})$ for only a subset of genes, and estimate the ratio change of the quantile threshold

$$\delta_y(\mathbf{b}^{(k)}, \mathbf{b}^{(0)}) = \frac{Q_y(\mathbf{b}^{(k)})}{Q_y(\mathbf{b}^{(0)})} \tag{13}$$

Determine the average $\bar{\delta}(\mathbf{b}^{(0)}, \mathbf{b}^{(k)})$ using the $\delta_y(\mathbf{b}^{(k)}, \mathbf{b}^{(0)})$ computed over the subset. Use $Q_y(\mathbf{b}^{(0)})$ and $\bar{\delta}(\mathbf{b}^{(0)}, \mathbf{b}^{(k)})$ to estimate $Q_y(\mathbf{b}^{(k)})$ for all genes y in the data, assuming that the ratio Equation 13 changes only slightly for all the genes.

This procedure quickly calculates the observed eGene LR and its threshold at all $\mathbf{b}$, using the permutation test or the sampling scheme in Sul *et al.* (2015) only once. We emphasize that this approximation is based on a subset of genes and is best used for finding a good choice $\mathbf{b}^*$. After we find $\mathbf{b}^*$, ideally, we would apply a complete Func-eGene run using $\mathbf{b}^*$ to determine the number of eGenes.

Because we determine good choices for $\mathbf{b}$ from the data, data re-use and over-fitting are two issues which can inflate the false-positive rate. To avoid this, we use a cross-validation method that divides the data into two subsets. We obtain best scores from one set and apply these scores to find eGenes in the other set, and vice versa.

3 Results

We applied our method Func-eGene to the GTEx dataset. The GTEx pilot study collected 9365 tissue samples from more than 30 distinct tissues from 237 post-mortem donors and performed RNA-seq to quantify gene expression in those tissues (The GTEx Consortium, 2015). We used the liver tissue data that has 97 samples. All individuals were genotyped at 5M SNPs and imputed with 1000 Genomes Phase I as the reference panel. The number of genes expressed in this tissue is 21 868.

3.1 Func-eGene controls false-positive rate

There are two ways to apply Func-eGene. Permutation Func-eGene relies on the traditional permutation test to calculate the null density of the observed statistic, whereas Mvn Func-eGene relies on the Mvn-sampling procedure in Sul *et al.* (2015).

Simulations demonstrate that both the permutation and the Mvn Func-eGene control the false-positive rate well. The gene `ENSG00000204219.5` expressed in the liver tissue is chosen as an example. This gene belongs to chromosome 1 and has 3872 cis-variants of which 431 are in the TSS region. For the sake of simplicity, variants in the TSS region are assumed 100 times more likely to affect gene expression. The non-uniform priors are then specified such that the ratio of priors for variants inside and outside TSS is 100/1. This ratio is reset to 1/1 in the uniform prior. In the alternative hypothesis, our method requires the true effects (i.e. z-score non-centrality parameters) of the variants as input parameters μ_i's.

In this experiment and onward, in the alternative hypothesis, we use $\mu_i = 3.5$—this choice is addressed in Section 4.

To simulate gene expression data under the null hypothesis, we permuted the gene expression measurements among the 97 individuals and kept the genotype data unchanged. In each simulation, we computed the eGene *P*-value using permutation and Mvn Func-eGene with both uniform and non-uniform priors. We applied the permutation procedure using 10^4 permutations and the Mvn method using 10^6 samplings. Simulated eGene *P*-values under 0.05 are considered significant. Permutation and Mvn Func-eGene have false-positive rates of 0.046 and 0.044, respectively, using a uniform prior. Both have false-positive rates of 0.051 and 0.052, respectively, using a non-uniform prior. *Q*–*Q* plots against the uniform density illustrate that the simulated densities under the null hypothesis match the uniform distribution well (Fig. 1(a)).

3.2 Func-eGene increases statistical power

An appropriate prior when applied in Func-eGene increases statistical power to detect candidate eGenes. To demonstrate this, we conducted simulation studies where there exists one variant that induces the gene expression. The gene `ENSG00000204219.5` is again chosen as an example.

To simulate the gene expression measurements for 97 individuals, we consider only variants within 150 kb up- and down-stream of the TSS, and presume that the maximum absolute effects of these variants (denoted as $\beta_{\max}$) to be the only genetic contribution toward the gene expression.

Gene expression measurements are thus simulated as $e_y = g\beta_{\max} + \epsilon$ where $(e_y)_{97\times1}$ is the simulated gene expression measurements, $g_{97\times1}$ is the genotype of the variant corresponding to $\beta_{\max}$, and $\epsilon_{97\times1}$ is sampled from $\text{Mvn}(0, I\sigma^2)$.

In our experiment, we vary this standard deviation σ from 0.00 to 1.50 At each instance of σ, we simulate 200 sets of gene expression data and compute the eGene *P*-value for each of them. The simulated power at this σ is the fraction of eGene *P*-values from the 200 simulated datasets that are less than 0.05. As we increase σ, the randomness effect dominates the effects of variants and the statistical power to discover any association between a variant and the gene expression diminishes.

We applied the uniform and non-uniform priors to the simulated data. In the non-uniform prior, the ratio of prior for a variant inside and outside the TSS region is 100/1. This non-uniform prior reflects the fact that eQTLs tend to reside near the TSS. Figure 1(b) indicates that the non-uniform prior increases statistical power in both the permutation and Mvn Func-eGene and that conditioned on the priors, permutation and Mvn Func-eGene archive equivalent statistical power.

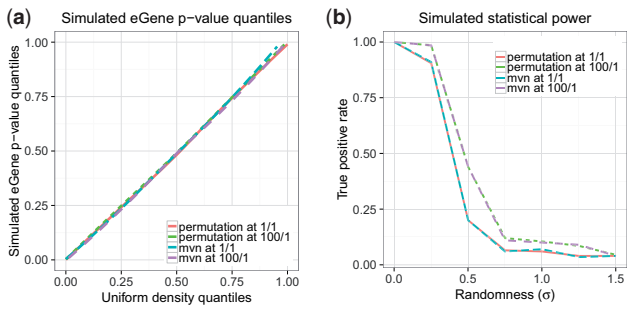

Fig. 1. (a) *Q*–*Q* plots of the uniform density quantiles against the simulated eGene *P*-value quantiles using Func-eGene at the gene `ENSG00000204219.5` under the null hypothesis. (b) Func-eGene simulated statistical power at the gene `ENSG00000204219.5`

3.3 Func-eGene discovers more candidate eGenes in the liver GTEx data

Permutation and Mvn Func-eGene are applied to the liver GTEx data which contains 21 868 quantile normalized gene expression measurements across 22 autosomal chromosomes in 97 individuals. Both uniform and non-uniform priors are tested. Our goal in this experiment is to demonstrate that: (i) using an informative non-uniform prior increases the number of candidate eGenes and (ii) Mvn Func-eGene is a good estimation of the gold-standard permutation test.

We computed eGene *P*-values by using 10^6 samplings for Mvn Func-eGene and 10^4 permutations for the permutation Func-eGene as indicated in the GTEx pilot analysis (The GTEx Consortium, 2015). The efficiency gain of Mvn sampling over the permutation test diminishes when the number of cis-variants for a gene is much greater than the number of sample size. Following Sul *et al.* (2015), we divide the cis-variants for a gene into blocks of size 500 kb, and apply Mvn sampling separately to each block. The most significant *P*-value taken across these blocks is the eGene *P*-value.

Cis-variants are variants located within the 1 Megabase up- and down-stream of TSS of a gene (The GTEx Consortium, 2015). In the liver GTEx data, the average number of cis-variants per gene is 4681. We define gene TSS-region to be 150 kb up- and down-stream of the gene TSS. The average fraction of variants inside this region is 14.74%.

Spurious effects on gene expression might dominate the effects of the cis-variants. To remove them, we regress out the following covariates: the first three genotyping principal components, the first 15 expression Probabilistic Estimation of Expression Residuals (PEER) factors, and gender (Stegle *et al.*, 2012; The GTEx Consortium, 2015). To be consistent with the GTEx pilot analysis, we transform eGene *P*-values into *Q*-values to control the false discovery rate over the entire sample. Genes having *Q*-values under 0.05 are considered eGenes (Storey and Tibshirani, 2003; The GTEx Consortium, 2015).

For the sake of simplicity, non-uniform priors are assigned so that the ratio of prior for a variant inside and outside the TSS region is 100/1, an assumption we address later.

Using an informative non-uniform prior increases the number of candidate eGenes in both permutation and Mvn Func-eGene (Table 1). The number of candidate eGenes has increased by 50.4%

Table 1. Number of candidate eGenes discovered by permutation and Mvn Func-eGene using uniform and non-uniform priors for 21 868 genes in the liver tissue GTEx data

	Permutation		Mvn		Overlap[a]	
eGene[b]	Yes	No	Yes	No	Yes	No
Uniform[c]	1626	20 242	1582	20 286	1549	20 209
Non-uniform[d]	2445	19 423	2449	19 419	2379	19 353
Overlap[e]	1484	19 281	1457	19294		

[a]Condition on uniform prior or non-uniform prior and count the number of eGenes (or not eGenes) that permutation Func-eGene agrees with Mvn Func-eGene.

[b]Indicates the number of genes detected to be eGenes.

[c]Uniform prior uses the prior ratio 1/1 for all variants.

[d]Non-uniform prior uses the prior ratio 100/1 so that variants in the TSS are 100 times more likely to affect gene expression.

[e]Condition on permutation or Mvn Func-eGene and count the number of eGenes (or not eGenes) that the uniform prior agrees with non-uniform prior.

Table 2. The number of eGene discovered at 19 annotation score ratios

Row	Ratio[a]	eGene [b]	Row	Ratio[a]	eGene[b]
1	1:1:1	2057	10	1:1:10	1579
2	1:10:1	2032	11	1:10:10	1747
3	1:100:1	1890 (1834)	12	1:100:10	1904
4	10:1:1	2450	13	10:1:10	2060
5	10:10:1	2331	14	100:1:10	2473 (2413)
6	10:100:1	1991	15	1:1:100	1280
7	100:1:1	2493 (2449)	16	1:10:100	1391
8	100:10:1	2489 (2449)	17	1:100:100	1673
9	100:100:1	2329	18	10:1:100	1548
			19	100:1:100	2014

[a]Prior ratios of variants inside and outside an annotation. These ratios are in order $A_{\text{TSS}} : A_{\text{DNase}}$: other.

[b]The numbers are obtained using the approximation method in Section 2.2.5. Numbers in parentheses are obtained using Mvn Func-eGene

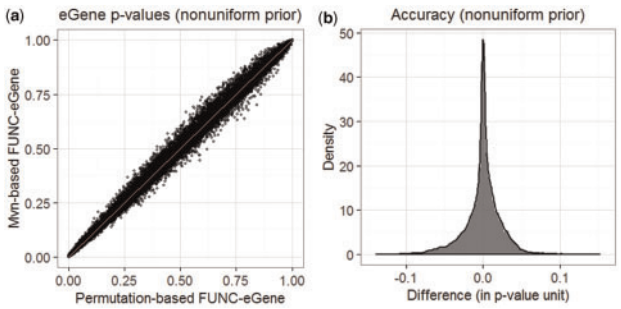

Fig. 2. (a) Scatter plot of eGene *P*-values using Mvn Func-eGene and the permutation test. (b) Histogram of the difference between eGene *P*-values using Mvn Func-eGene and the permutation test

in permutation Func-eGene (from 1626 to 2445) and by 54.8% in Mvn Func-eGene (from 1582 to 2449) by applying non-uniform priors. The numbers were comparable between the two implementations, indicating that Mvn approximates the null distribution of observed statistics well.

Sul *et al.* (2015) have shown that Mvn sampling and permutation test have comparable eGene *P*-values without using any prior. Here we show that the results are also comparable when using a non-uniform prior. Figure 2(a) compares Mvn eGene *P*-values against those in the permutation test, and Figure 2(b) indicates the Mvn method is at most ±0.10 units from the gold-standard permutation test *P*-values. Our Mvn method and the permutation test agree on 2379 candidate eGenes. Because Mvn method is an approximation to the permutation test, we analyze the candidate eGenes that the Mvn method misses and falsely discovers. Of the 66 missed eGenes, the maximum *Q*-value is 0.079 and the median is 0.053. Of the 70 falsely discovered eGenes, the minimum *Q*-value is 0.028 and the median 0.045. Thus, these misclassified eGenes are genes with borderline *Q*-values.

The function *Q*-value in *R* requires that the distribution of input *P*-values has a fairly flat right tail (Storey and Tibshirani, 2003). Our eGene *P*-values meet this condition (Fig. 3).

Figure 4(a) compares our eGene *Q*-values against those in the permutation test, and Figure 4(b) indicates that accuracy of Mvn Func-eGene is at most ±0.05 units from the gold-standard permutation test *Q*-values.

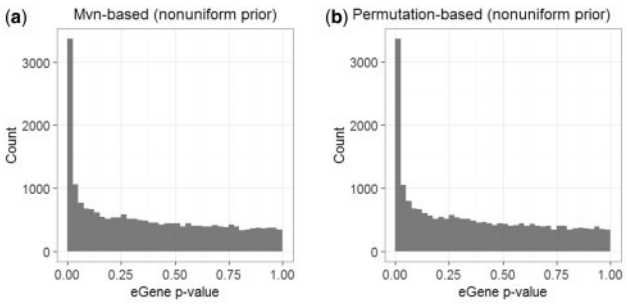

Fig. 3. Histogram of the eGene *P*-values using Mvn Func-eGene and the permutation test

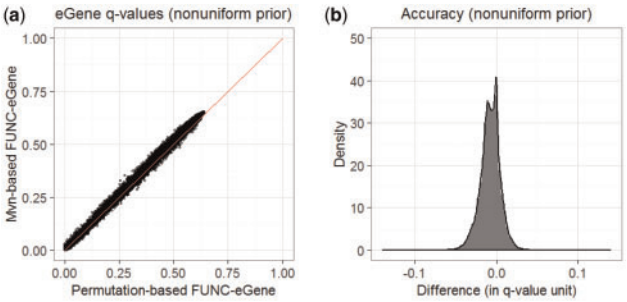

Fig. 4. (a) Scatter plot of the eGene *Q*-values using Mvn Func-eGene and the permutation test. (b) Histogram of the difference between eGene *Q*-values using Mvn Func-eGene and the permutation test

3.4 Func-eGene chooses appropriate priors

So far, we have assumed that the priors are defined a priori. In this section, we applied the method in Section 2.2.5 to determine appropriate priors from the data. To demonstrate this method, we consider two functional annotations A_{TSS} and A_{DNase}. A_{TSS} contains variants within 150 kb of the TSS. A_{DNase} contains variants within the E066 DNase hypersensitivity narrow gapped peaks (http://egg2.wustl.edu/roadmap/data) in the liver tissue. Across 21 868 autosomal genes, the average fraction of cis-variants belonging in the A_{TSS} and the A_{DNase} are 14.74% and 4.66%, respectively. The overlap between them is 0.88%.

The relative prior ratio between annotations can be represented by three numbers, b_1, b_2 and b_3, which correspond to the scores for $A_{\text{TSS}}, A_{\text{DNase}}$ and neither. For example, a variant in A_{TSS} has b_1/b_2 times higher score than a variant in A_{DNase}. Since the scores are assumed to be additive in Equation 11, a variant in both A_{TSS} and A_{DNase} has $(b_1 + b_2)/b_3$ times higher score than a variant in neither classes. We constrained each of b_1, b_2, and b_3 to be between 100 times greater and 100 times smaller than the other two. Only the relative ratios of the scores matter. Given this constraint, we did a grid search over the parameter space evaluating a total of 441 choices of score $\mathbf{b} = (b_1, b_2, b_3)$.

We implemented Mvn Func-eGene only once with the functional scores $\mathbf{b}^{(0)} = (1, 1, 1)$. At each gene, we recorded the upper 1% quantile of the observed statistics, which corresponds to the significance threshold $\alpha = 0.01$. We used this *P*-value threshold because this threshold roughly corresponds to the maximum eGene *P*-value of genes whose *Q*-values are under the threshold 0.05 in our data of Section 3.3. Using this complete Mvn Func-eGene run at $\mathbf{b}^{(0)}$, we computed new observed statistics and their thresholds at another choice $\mathbf{b}^{(k)}$ using the approximation method in Section 2.2.5.

Table 3. The number of candidate eGenes detected by Mvn Func-eGene at the best priors in each annotation

Annotation	(%)[a]	Ratio[b]	eGene[c]
H3K27ac[d]	12.25	40/1	1944
H3K27me3[e]	7.26	1/70	1880
H3K4me1[d]	16.38	80/1	1858
H3K4me3[d]	7.73	50/1	1917
H3K9ac[d]	9.74	100/1	1861
H3K9me3[e]	11.92	1/50	1879
TSS	14.74	60/1	2479
DNase	4.66	100/1	1834
Uniform	100	1/1	1592

[a]Average percent of variants in an annotation.
[b]Best prior ratios of variants inside and outside an annotation given by the method in Section 2.2.5.
[c]Numbers are obtained by using Mvn Func-eGene at the best ratios.
[d]Associated with gene activation.
[e]Associated with gene suppression

We tested 441 choices of **b**. Table 2 displays 19 of these 441 choices and the number of candidate eGenes discovered using these scores. The score $\mathbf{b} = (100, 1, 1)$ had the most candidate eGenes, indicating that using A_{TSS} alone is enough to increase the number of eGene discovered. For a few choices of **b**, we ran the Mvn Func-eGene without using the approximation method, and the results are comparable.

3.5 Evaluation using histone marks

Because there exist other genomic annotations, we hope to evaluate their effects on eGene detection. Ideally, we would use all annotations simultaneously in the model and find the optimal prior parameters. Unfortunately, our method uses grid search which is prohibitively time-consuming in high dimensional space. For this reason, we evaluate each annotation separately, which at least can provide an overview of which annotation contains useful information for eGene discovery.

We applied the optimization to six histone marks, A_{TSS}, and A_{DNase}. In each annotation, we find the best prior ratio of the variants inside and outside the annotated site. Table 3 indicates that using these annotations can increase the number of candidate eGenes. These numbers are from a complete Mvn Func-eGene run using all genes in the liver tissue data and not from the method in Section 2.2.5. The marks associated with activation of gene expression (H3K27ac, H3K4me1, H3K4me3, H3K9ac) have prior ratios more than one, whereas the marks associated with suppression of gene expression (H3K27me3, H3K9me3) have prior ratios less than one. All of the histone marks yield more candidate eGenes than the uniform prior.

In Table 3, TSS has the best prior ratio at 60/1. This prior gives more eGenes than the prior 100/1 reported in Table 1. Figure 5(a) indicates that the false-positive rate at the gene `ENSG00000204219.5` using prior 60/1 matches well to the uniform density and is not inflated. Figure 5(b) indicates that the simulated statistical power at this gene using Mvn Func-eGene with the prior ratio 60/1 is not worse than the guessed ratio 100/1. It is important to stress that the ratio 60/1 is best with respect to all the genes in liver tissue data and not this particular gene.

3.6 Mvn Func-eGene has better runtime than permutation method

In this section, we compare the runtime of the permutation and Mvn Func-eGene. In this experiment, the Mvn method uses 1000 samples, and the permutation-based procedure uses 1000 permutations. In both cases the number of individuals is 97. The standard permutation method computes one eGene *P*-value in time linear in the number of cis-variants for a gene (Sul *et al.*, 2015). The permutation-based Func-eGene relies on this basic permutation procedure and has almost identical runtime to the standard permutation method. Figure 6 displays the runtime for a few randomly chosen genes from the GTEx liver tissue data. Mvn Func-eGene is considerably faster than permutation approaches but runs in polynomial time. In the GTEx liver tissue data, the 5% upper quantile of cis-variants per gene is 6833 variants. Thus in practice the polynomial nature of the Mvn Func-eGene does not impede its application.

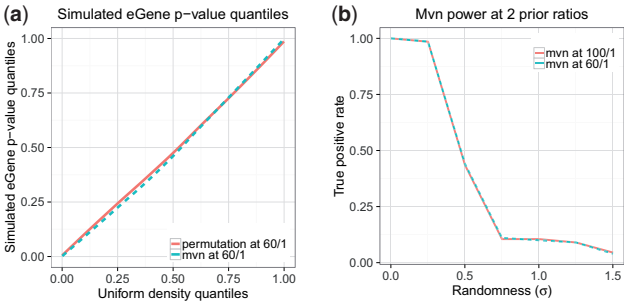

Fig. 5. (a) False-positive rate for permutation and Mvn Func-eGene using prior ratio 60/1. (b) Mvn Func-eGene statistical power at gene `ENSG00000204219.5` using ratio 60/1 is not worse than the ratio 100/1

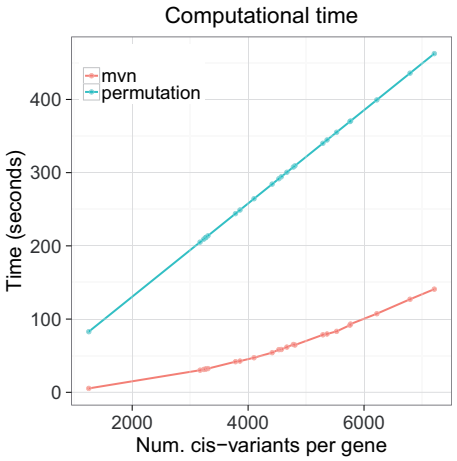

Fig. 6. Permutation and Mvn Func-eGene runtime

4 Discussion

In this article, we have introduced a new method Func-eGene that relies on the association study methods in Darnell *et al.* (2012) and Eskin (2008) and uses genomic annotations of the cis-variants to create a non-uniform prior that can detect more eGenes. We applied our method to the liver tissue dataset from the GTEx Consortium, and the results indicate that distance from TSS appears to contain enough information that is needed to find more candidate eGenes.

Our method has many layers of procedures which can be time-consuming. To reduce runtime, we introduced many ideas. We employed LR statistic which is more efficient to obtain than a *P*-value in a multi-threshold association study. We replaced the time-consuming permutation test with the use of Mvn sampling. To avoid reassessing significance thresholds at each new prior in our grid

search, we developed an approximation method which uses a subset of the tested genes. Due to these heuristics, we were able to conduct a grid search using TSS, DNase and six histone modifications as functional classes. Because our method uses grid search where the runtime increases exponentially with the number of annotations, our current method is not yet applicable for simultaneously handling a large number of annotations. One future goal is to develop a better optimization method than a grid search.

Lastly, we address the fact that in the alternative hypothesis, the true effects of the variants on the gene expression are unknown in practice. It has been demonstrated in previous studies that different choices of these effect sizes do not greatly change the outcome (Darnell *et al.*, 2012; Eskin, 2008). Another option is to consider some continuous prior density on these true effects and then integrate over their valid domain (Benner *et al.*, 2016; Hormozdiari *et al.*, 2014, 2015). This idea is another future research plan.

Acknowledgment

We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

Funding

D.D., F.H. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589 and 1331176, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. D.D. is supported by the Genomic Analysis Training Program grant T32HG002536. B.H. is supported by a grant from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea (2016-717) and a grant from the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (HI14C1731). E. E. is supported in part by the NIH BD2K award, U54EB020403. J.Z and J.E. are supported by National Institute of Health grants R01ES024995 and U01HG007912, NSF CAREER Award no. 1254200, and an Alfred P. Sloan Fellowship (J.E.).

Conflict of Interest: none declared.

References

Benner,C. *et al.* (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, btw018.

Brem,R.B. and Kruglyak,L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA*, **102**, 1572–1577.

Bryois,J. *et al.* (2014) Cis and trans effects of human genomic variants on gene expression. *PLoS Genetics*, **10**, e1004461.

Darnell,G. *et al.* (2012) Incorporating prior information into association studies. *Bioinformatics*, **28**, i147–i153.

Davis,J.R. *et al.* (2016) An efficient multiple-testing adjustment for eQTL studies that accounts for linkage disequilibrium between variants. *Am. J. Hum. Genet.*, **98**, 216–224.

Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.*, **33**, 364–376.

Eskin,E. (2008) Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Gen. Res.*, **18**, 653–660.

Gilad,Y. *et al.* (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.

Gusev,A. *et al.* (2014) Regulatory variants explain much more heritability than coding variants across 11 common diseases. *BioRxiv*, 004309.

Hormozdiari,F. *et al.* (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.

Hormozdiari,F. *et al.* (2015) Identification of causal genes for complex traits. *Bioinformatics*, **31**, i206–i213.

Rubin,D. *et al.* (2006) A method to increase the power of multiple testing procedures through sample splitting. *Stat. App. Genet. Mol. Biol.*, **5**.

Stegle,O. *et al.* (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, **7**, 500–507.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.*, **100**, 9440–9445.

Sul,J.H. *et al.* (2015) Accurate and fast multiple-testing correction in eQTL studies. *Am. J. Hum. Genet.*, **96**, 857–868.

The GTEx Consortium (2015) The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**, 648–660.

The Roadmap Epigenomics Mapping Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

van de Geijn,B. *et al.* (2015) WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Met.*, **12**, 1061–1063.