Submitted to the Annals of Applied Statistics

MOLECULAR QTL DISCOVERY INCORPORATING GENOMIC ANNOTATIONS USING BAYESIAN FALSE DISCOVERY RATE CONTROL

By Xiaoquan Wen

University of Michigan

Mapping molecular QTLs has emerged as an important tool for understanding the genetic basis of cell functions. With the increasing availability of functional genomic data, it is natural to incorporate genomic annotations into QTL discovery. Discovering molecular QTLs is typically framed as a multiple hypothesis testing problem and solved using false discovery rate (FDR) control procedures. Currently, most existing statistical approaches rely on obtaining p-values for each candidate locus through permutation-based schemes, which are not only inconvenient for incorporating highly informative genomic annotations but also computationally inefficient. In this paper, we discuss a novel statistical approach for integrative QTL discovery based on the theoretical framework of Bayesian FDR control. We use a Bayesian hierarchical model to naturally integrate genomic annotations into molecular QTL mapping and propose an empirical Bayes-based computational procedure to approximate the necessary posterior probabilities to achieve high computational efficiency. Through theoretical arguments and simulation studies, we demonstrate that the proposed approach rigorously controls the desired type I error rate and greatly improves the power of QTL discovery when incorporating informative annotations. Finally, we demonstrate our approach by analyzing the expression-genotype data from 44 human tissues generated by the GTEx project. By integrating the simple annotation of SNP distance to transcription start sites, we discover more genes that harbor expression-associated SNPs in all 44 tissues, with an average increase of 1,485 genes per tissue.

1. Introduction. With the advancements in sequencing technology, mapping quantitative trait loci (QTLs) with cellular phenotypes has emerged as a powerful tool for understanding the genetic basis of cell functions. Recent QTL mapping studies using RNA-seq, ChIP-seq, DNaseI-seq, ATAC-seq and DNA methylation data have revealed that an abundance of genetic variants are associated with various cellular phenotypes (Ardlie et al., 2015; Banovich et al., 2014; Degner et al., 2012; Ding et al., 2014; McVicker et al., 2013). Subsequently, the discovery of molecular QTLs has provided valuable

Keywords and phrases: Molecular QTL; Genomic Annotations; Bayesian FDR Control; QTL Mapping.

insights for understanding the molecular mechanisms of complex diseases, as demonstrated by Neto et al. (2013).

Compared to traditional QTL mapping, a distinctive feature of molecular QTL analysis is that tens of thousands of molecular phenotypes are simultaneously measured using high-throughput sequencing technology (e.g., genome-wide gene expression profiling by RNA-seq) in addition to high-density genotyping. Consequently, it prevents direct applications of well-established statistical approaches designed for traditional QTL mapping (Churchill and Doerge, 1994; Doerge and Churchill, 1996; Neto et al., 2012) and imposes a new type of statistical challenge. In this paper, we use the term QTL to refer to the genomic regions that harbor trait-associated causal variants, and following Veyrieras et al. (2008), we refer to the actual causal variants as quantitative trait nucleotides (QTNs). In practice, the statistical analysis of molecular QTLs typically consists of two stages: the primary goal of the first stage is to screen a large number of candidate loci and identify QTLs, and we refer to this process as *QTL discovery*; in the second stage, a fine-mapping analysis is performed to determine the potential QTNs in each discovered QTL. Our primary focus in this paper is the statistical analysis for QTL discovery. In addition to providing a list of candidate QTLs for fine-mapping analysis, QTL discovery is also uniquely important for identifying relevant genes and biological pathways for network and pathway analysis. In mapping *cis* expression quantitative trait loci, the candidate genomic regions are generally formed by the coding and the neighboring regulatory regions of each target gene, and in this context, QTL discovery is also known as eGene discovery (Ardlie et al., 2015; Lappalainen et al., 2013; Sul et al., 2015).

QTL discovery is framed as a multiple hypothesis testing problem, for which the null hypothesis asserts no QTN within each locus of interest. The standard (frequentist) approach has been established in applications of *cis*-eQTL mapping and can be straightforwardly generalized to other molecular QTL analyses (Ardlie et al., 2015; Flutre et al., 2013; Lappalainen et al., 2013). The standard approach, which represents the current state-of-the-art in QTL discovery, first performs single SNP association testing for all phenotype-SNP pairs. For each locus, the minimum p-value from all member SNPs is then regarded as the locus-level test statistic and is subsequently converted into a locus-level p-value for the hypothesis testing. Because the null distribution of the locus-level test statistic (i.e., the minimum p-value from single SNP testing) is generally unknown due to complicated patterns of linkage disequilibrium (LD), extensive permutations are required to obtain the locus-level p-values. Subsequently, false discovery rate (FDR) control procedures, e.g., Benjamini-Hochberg (Benjamini and Hochberg, 1995) and Storey's q-value procedures (Storey, 2003), are applied to account for multiple testing of tens of thousands of loci at the genome-wide scale.

Although the frequentist procedure is statistically justified and widely applied, there is no principled way to flexibly incorporate valuable genomic annotations. With the increasing availability of functional genomic data (ENCODE Project Consortium et al., 2012; Kundaje et al., 2015; Pique-Regi et al., 2011), the scientific community has accumulated substantial knowledge on the functional roles of individual genetic variants. Incorporating such prior knowledge into QTL discovery and prioritizing the genomic loci that harbor well-annotated functional variants intuitively improve the statistical power for QTL discovery. Similarly, results from existing QTL analyses can also provide valuable insights into the distributive patterns of causal QTNs. For example, almost all of the analyses in *cis*-eQTL mapping confirm that associated casual SNPs tend to cluster around transcription start sites (TSS) and that the abundance of signals rapidly decreases away from TSS (Ardlie et al., 2015; Lappalainen et al., 2013; Wen, Luca and Pique-Regi, 2015). In light of this strong pattern, it appears natural to use SNP distance to TSS (DTSS) as an annotation and up-weight the SNPs close to TSS in eQTL analysis *a priori* rather than treating every *cis*-SNP equally. However, to the best of our knowledge, a principled approach that can effectively incorporate known genomic annotations into the analysis of QTL discovery does not exist.

In this paper, we propose a Bayesian hierarchical model to integrate SNP-level annotations into QTL mapping. Bayesian approaches have been shown to be effective and flexible in dealing with complex settings in traditional QTL mapping (Ball, 2001; Breitling et al., 2008; Sillanpää and Arjas, 1999; Stephens and Fisch, 1998; Yi et al., 2005). Here, we propose an efficient computational framework, named TORUS, to perform Bayesian multiple hypothesis testing for molecular QTL discovery and study its statistical properties. Through simulations, we demonstrate the superiority of the proposed approach over the state-of-the-art, gold-standard approach in terms of both computational efficiency and power of QTL discovery. Finally, we demonstrate our approach by analyzing eQTL data from 44 human tissues that were recently released by the NIH GTEx project.

2. Methods.

In this section, we describe our Bayesian hierarchical model for integrative molecular QTL discovery and present the computational strategies to perform the highly efficient Bayesian FDR control procedure. We implement the proposed statistical methods in the software package TORUS (QTL Discovery utilizing Genomic Annotations), which is freely available from `https://github.com/xqwen/torus`.

2.1. *Bayesian Hierarchical Model for QTL Discovery.* We consider a general problem of QTL mapping at the genome-wide scale. In particular, we assume that there are L genomic loci (in many cases, the loci are naturally formed by including coding and regulatory regions of the target genes), each of which contains p_l SNPs for $l = 1, ..., L$. Without loss of generality, we consider a sample of n unrelated individuals, and for each locus l, we model the locus-specific quantitative trait measurement $\boldsymbol{y}_l$ of the sample using the following linear regression model,

$$(1) \qquad \boldsymbol{y}_l = \mu_l \mathbf{1} + \sum_{i=1}^{p_l} \beta_{l_i} \boldsymbol{g}_{l_i} + \boldsymbol{e}_l, \ \boldsymbol{e}_l \sim \mathrm{N}(\mathbf{0}, \tau_l^{-1} \boldsymbol{I}),$$

where n-vectors $\boldsymbol{g}_{l_i}, \boldsymbol{e}$ represent the sample genotypes at genetic variant l_i and the residual errors, respectively. (At present, we assume that $\boldsymbol{y}_l$ is a univariate quantitative trait, and we relax this assumption and extend the framework to multivariate quantitative traits in section 2.5.) Furthermore, we denote $\boldsymbol{G}_l := (\boldsymbol{g}_{l_1}, ..., \boldsymbol{g}_{l_{p_l}})$. Note that the model allows multiple SNP associations within a given locus. The regression coefficients μ_l and β_{l_i} represent the intercept and the genetic effect of each genetic variant, respectively, and τ_l^{-1} denotes the residual error variance. Following Wen (2014), we further denote the latent binary association status of each genetic variant l_i by $\gamma_{l_i} := \mathbf{1}(\beta_{l_i} \neq 0)$ (i.e., γ_{l_i} is dichotomized from the corresponding β_{l_i}), and $\boldsymbol{\gamma}_l := (\gamma_{l_1}, ..., \gamma_{l_{p_l}})$.

Our prior specifications for the parameters μ, β_{l_i} and τ are mostly standard and follow directly from Marin and Robert (2007), and we leave the details to Appendix A of the Supplementary Materials. Most importantly, we use the prior specification for $\boldsymbol{\gamma}_l$ to incorporate variant-level genomic annotation information. Specifically, we assume that the γ_{l_i}s are *a priori* independent and that

$$(2) \qquad \log \left[\frac{\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{\eta}, \boldsymbol{d}_{l_i})}{\Pr(\gamma_{l_i} = 0 \mid \boldsymbol{\eta}, \boldsymbol{d}_{l_i})} \right] = \eta_0 + \sum_{j=1}^{m} \eta_j d_{l_i j},$$

where we use $\boldsymbol{d}_{l_i} := (d_{l_i 1}, ..., d_{l_i m})$ to denote the variant-specific genomic features, and the hyper-parameter $\boldsymbol{\eta} := (\eta_0, ..., \eta_m)$, which characterizes the enrichment level of each genomic feature in trait-associated genetic variants, is referred to as the enrichment parameter. Finally, we use $\boldsymbol{D}_l := (\boldsymbol{d}_{l_1}, ..., \boldsymbol{d}_{l_p})$ to denote the collection of the annotation data at locus l. Note that in the special case where no genomic annotation is used in the analysis, the prior model (2) contains a single parameter η_0 that quantifies the prevalence of trait-associated genetic variants among all candidate SNPs. We refer to this special case of the model as the baseline model.

In molecular QTL mapping, it is sometimes desirable to explicitly account for the relatedness of the samples and/or the polygenic effects on the quantitative trait of interest. To this end, we simply modify the regression model (1) by adding a random effect term, $\boldsymbol{u}_l$, i.e.,

$$\begin{gathered}\boldsymbol{y}_l = \mu_l \mathbf{1} + \sum_{i=1}^{p_l} \beta_{l_i} \boldsymbol{g}_{l_i} + \boldsymbol{u}_l + \boldsymbol{e}_l, \\ \boldsymbol{u}_l \sim \mathrm{N}(\mathbf{0}, \theta_l \tau_l^{-1} \boldsymbol{A}) \text{ and } \boldsymbol{e}_l \sim \mathrm{N}(\mathbf{0}, \tau_l^{-1} \boldsymbol{I}).\end{gathered} \tag{3}$$

In particular, $\boldsymbol{A}$ represents the known kinship (correlation) matrix, which can be efficiently pre-estimated from the available genotype data, and the unknown variance component parameter θ_l characterizes the magnitude of the random effect. Here, we simply note that our main results based on the regression model (1) can be straightforwardly extended to the model (3) by applying the recent results on Bayesian model comparison, i.e., the analytic computation of Bayes factors, in linear mixed models (Wen, 2015).

2.2. *Multiple Hypothesis Testing and Bayesian FDR Control .* The problem of QTL discovery is framed as a hypothesis testing problem. Specifically, we identify locus l as a QTL if the null hypothesis asserting no trait-associated genetic variant within the locus, i.e.,

$$H_0 : \boldsymbol{\gamma}_l = \mathbf{0},$$

is rejected. The problem of multiple hypothesis testing arises because we perform simultaneous testing for tens of thousands of loci across the genome when mapping molecular QTLs.

To take full advantage of our hierarchical model, we adopt a Bayesian FDR control strategy (Müller et al., 2004; Newton et al., 2004). Briefly, the Bayesian FDR control requires computing the posterior probability $u_l := \Pr(\boldsymbol{\gamma}_l = 0 \mid \boldsymbol{y}_l, \boldsymbol{G}_l)$ to summarize the evidence for (or against) the null hypothesis for each locus l. The null hypothesis is intuitively rejected if the corresponding u_l is smaller than the pre-defined threshold. Based on a pre-defined FDR control level α, a straightforward algorithm (Newton et al., 2004) can be applied to determine the induced rejection threshold t_α such that

$$t_\alpha = \arg\max_t \left(\frac{\sum_{u_l \le t} u_l}{[\sum_l \mathbf{1}(u_l \le t) \vee 1]} \le \alpha \right), \tag{4}$$

where the expression $\sum_l \mathbf{1}(u_l \le t)$ in the denominator represents the total number of rejections at threshold t, and the expression $\sum_{u_l \le t} u_l$ in the numerator represents the expected false rejections at threshold t. The Bayesian FDR control procedure is naturally connected to its frequentist counterpart (Appendix B of the Supplementary Materials), with the primary difference being that the Bayesian FDR is conditional on the observed data in hand whereas the frequentist procedure computes FDR over hypothetically repeated experiments. Furthermore, Müller et al. (2004) proved that the Bayesian procedure is theoretically optimal in the sense that it minimizes the corresponding false non-discovery rate (FNR, a measure of type II error).

2.3. *Approximate Computation of Posterior Probability.* By computing u_l based on the proposed Bayesian hierarchical model, the Bayesian FDR control procedure naturally allows genomic annotations to be leveraged in QTL discovery. However, the exact evaluation of u_l requires integrating out all enrichment parameters and exploring an enormous space of all possible association models representing different values of $\boldsymbol{\gamma}_l$. This evaluation becomes a computationally daunting challenge even for a single locus, let alone the genome-wide application for tens of thousands of loci. To overcome the computational difficulty, we propose applying two levels of approximation.

A critical intermediate step in evaluating u_l is to compute the probability $\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{\eta})$ for a given value of the enrichment parameter $\boldsymbol{\eta}$. Specifically, we evaluate this quantity by

$$\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\eta}) = \frac{\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{D}_l, \boldsymbol{\eta})}{\sum_{\boldsymbol{\gamma}'} \Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}' \mid \boldsymbol{D}_l, \boldsymbol{\eta}) \, \mathrm{BF}(\boldsymbol{\gamma}')}, \tag{5}$$

where $\mathrm{BF}(\boldsymbol{\gamma})$ denotes the Bayes factor

$$\mathrm{BF}(\boldsymbol{\gamma}) := \frac{\Pr(\boldsymbol{y}_l \mid \boldsymbol{G}_l, \boldsymbol{\gamma}_l = \boldsymbol{\gamma})}{\Pr(\boldsymbol{y}_l \mid \boldsymbol{G}_l, \boldsymbol{\gamma}_l \equiv \mathbf{0})}$$

and represents the marginal likelihood for $\boldsymbol{\gamma}_l = \boldsymbol{\gamma}$ (by definition, $\mathrm{BF}(\mathbf{0}) = 1$). Although the calculation of $\mathrm{BF}(\boldsymbol{\gamma})$ for any given $\boldsymbol{\gamma}$ can be achieved analytically for a wide range of linear model systems (Wen, 2014), it is practically unfeasible to enumerate all possible $\boldsymbol{\gamma}$ values when the number of SNPs within a locus is large (for p SNPs in a locus, there are a total of 2^p $\boldsymbol{\gamma}$ values to enumerate). Here, we propose using the approximation

$$\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\eta}) \approx \frac{\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{D}_l, \boldsymbol{\eta})}{\sum_{||\boldsymbol{\gamma}'|| \le K} \Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}' \mid \boldsymbol{D}_l, \boldsymbol{\eta}) \, \mathrm{BF}(\boldsymbol{\gamma}')}, \tag{6}$$

where $||\boldsymbol{\gamma}'||$ denotes the number of non-zero indicators in vector $\boldsymbol{\gamma}'$ (i.e., the 0-norm of the $\boldsymbol{\gamma}'$ vector), and the subset $\{\boldsymbol{\gamma}' : ||\boldsymbol{\gamma}'|| \le K\}$ consists of only the models with no more than K associated SNPs and hence represents a (much) reduced model space. We summarize the property of the proposed approximation (6) in the following lemma.

LEMMA 1 *The approximation (6) represents a conservative upper bound for the posterior probability* $\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\eta})$.

The proof is trivial by noting the inequality

$$\frac{\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{D}_l, \boldsymbol{\eta})}{\sum_{||\boldsymbol{\gamma}'|| \le K} \Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}' \mid \boldsymbol{D}_l, \boldsymbol{\eta}) \, \mathrm{BF}(\boldsymbol{\gamma}')} \ge \frac{\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{D}_l, \boldsymbol{\eta})}{\sum_{\boldsymbol{\gamma}'} \Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}' \mid \boldsymbol{D}_l, \boldsymbol{\eta}) \, \mathrm{BF}(\boldsymbol{\gamma}')}.$$

Lemma 1 is critically important for ensuring that applying the approximation (6) does not inflate type I errors for any given $\boldsymbol{\eta}$ value. Furthermore, we note that the approximation is accurate if

$$\frac{\sum_{||\boldsymbol{\gamma}'|| > K} \Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}' \mid \boldsymbol{D}_l, \boldsymbol{\eta}) \, \mathrm{BF}(\boldsymbol{\gamma}')}{\sum_{||\boldsymbol{\gamma}'||} \Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}' \mid \boldsymbol{D}_l, \boldsymbol{\eta}) \, \mathrm{BF}(\boldsymbol{\gamma}')} \to 0, \tag{7}$$

i.e., the posterior probability mass is concentrated on the space of association models containing few QTNs ($||\boldsymbol{\gamma}'|| \le K$). To justify this approximation and select approximate K in practice, we note the following observation:

OBSERVATION 1 *For the vast majority of genomic loci in molecular QTL mapping, often only one QTN can be detected with high probability.*

This observation has been made in many applications of molecular QTL mapping, e.g., Wen, Luca and Pique-Regi (2015) report that in a cross-population eQTL analysis, there are less than 7% of 11,838 integrated protein-coding and linc-RNA genes that show evidence of harboring more than a single QTN, which is likely a consequence of the combination of the true underlying biology, the noise level of the current experimental technology in measuring molecular phenotype, and the limitation of sample sizes in current practice of molecular QTL mapping. Based on this observation, we simply set $K = 1$ for the approximation (6) to achieve the best computational efficiency. Note that the approximation becomes the most accurate when there is at most one causal SNP within a locus. In this special case, the approximation has an analytic form, i.e.,

$$\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\eta}) \approx \frac{1}{1 + \sum_{k=1}^{p} e^{\boldsymbol{\eta}' \boldsymbol{d}_{l_k}} \, \mathrm{BF}_{l_k}}, \tag{8}$$

where BF_{l_k} denotes the Bayes factor for the k-th SNP from the single SNP association model and can be computed analytically from the corresponding single SNP testing statistics (Servin and Stephens, 2007; Wakefield, 2009). One of the most attractive advantages is that the full analysis now requires only summary-level statistics rather than full individual-level genotype-phenotype data that can be difficult to obtain due to privacy issues.

Importantly, note that the state-of-the-art frequentist approach based on the locus-level test statistic of the minimum p-value from single-variate testing also implicitly assumes $K = 1$ for the alternative model: as shown in De la Cruz et al. (2010), the minimum p-value exhibits the best power if and only if the locus of interest contains exactly one QTN.

Our second approximation involves using an empirical Bayes approach to evaluate u_l by

$$\hat{u}_l := \Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \hat{\boldsymbol{\eta}}), \tag{9}$$

where $\hat{\boldsymbol{\eta}}$ denotes the maximum likelihood estimate of the enrichment parameter. This approach essentially replaces the integration of $\boldsymbol{\eta}$ in the full Bayesian procedure by a more computationally efficient optimization procedure. To determine the MLE of $\boldsymbol{\eta}$, we derive an EM algorithm (details are given in Appendix C.1 of the Supplementary Materials) by treating $\boldsymbol{\gamma}_l$s as missing data and naturally pooling information across all candidate loci. Briefly, in the E-step, we compute the posterior inclusion probability (PIP) of each SNP in each locus given the current estimate of $\boldsymbol{\eta}$, and in the M-step, we solve a classic convex optimization problem that is equivalent to fitting a logistic regression model using the PIPs from the E-step as the response variable and annotations as predictors. In the context of molecular QTL mapping, tens of thousands of simultaneous molecular phenotype measurements can be regarded as a large size of approximately independent samples for the inference of the enrichment parameter. (The justification for the assumption of approximate independence is given in the Discussion section.) Consequently, the influence of any prior distribution on $\boldsymbol{\eta}$ is likely diminished, and its posterior distribution is expected to be highly peaked under such a setting that is close to the ideal asymptotic setup. Therefore, we expect that the empirical Bayes approach and the full Bayesian inference behave similar with respect to the inference of $\boldsymbol{\eta}$, as demonstrated in Wen (2011).

The computational difficulty of the EM algorithm lies in the evaluation of PIPs in the E-step, whose exact computation is intractable. The Monte Carlo EM (MCEM) algorithm (Levine and Casella, 2001) is a possible solution designed for such a scenario in the literature, in which the PIPs can be obtained by sampling from an MCMC algorithm in each E-step. However, running the extensive MCMC algorithm in each E-step is computationally expensive and is not ideal for dealing with the molecular QTL data at the genome-wide scale.

Motivated by Observation 1, we again apply the approximation strategy that is similar to (8) to compute the PIP for each SNP i at locus l, i.e.,

$$\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\eta}) \approx \frac{e^{\boldsymbol{\eta}'\boldsymbol{d}_{l_i}}\,\mathrm{BF}_{l_i}}{1 + \sum_{k=1}^{p} e^{\boldsymbol{\eta}'\boldsymbol{d}_{l_k}}\,\mathrm{BF}_{l_k}}, \tag{10}$$

which enables the highly efficient implementation of the EM algorithm. We provide the detailed derivation of the approximation (10) in Appendix C.2 of the Supplementary Materials. The approximation (10) essentially focuses on only a small sub-space of all possible alternative models that contain at most a single QTN. Intuitively, it yields accurate PIPs if locus l harbors either no or exactly one QTN.

Based on Observation 1, we fully expect that the EM algorithm utilizing the approximation (10) in the E-step yields an accurate point estimate for the enrichment parameter $\boldsymbol{\eta}$ in practical settings. Nevertheless, we also study the behavior of the proposed EM algorithm under a more general setting without any restriction on the number of genomic loci harboring multiple QTNs. We summarize our conclusion in the following claim.

CLAIM 1 *For the genomic loci that harbor multiple QTNs, the approximation (10) shrinks the posterior inclusion probabilities of the causal SNPs toward 0. Consequently, the SNP-level priors based on the EM estimates are biased toward 0 compared to the corresponding empirical Bayes estimates based on the exact inference.*

We provide theoretical justifications for Claim 1 in Appendix C.3 of the Supplementary Materials.

In addition, assuming that the genetic effects of QTNs (i.e., the β_{l_i}s) are irrelevant to the genomic annotations, we further show that by applying the approximation (10), the EM estimates of the enrichment parameters, $\eta_1, ..., \eta_m$, are unbiased, whereas the estimate of the intercept term, η_0, is downward biased (Appendix C.3 of the Supplementary Materials). We also perform numerical experiments to validate this finding (Appendix C.4 and Figure 1 of the Supplementary Materials).

Taken together, our overall procedure approximates u_l for Bayesian FDR control by

$$\hat{u}_l = \frac{1}{1 + \sum_{k=1}^{p} e^{\hat{\boldsymbol{\eta}}' \boldsymbol{d}_{l_k}} \, \mathrm{BF}_{l_k}}, \tag{11}$$

which represents a conservative approximation of u_l. The approximation is justified by the results of Lemma 1 and Claim 1, and it ensures that the desired FDR level is rigorously controlled.

Although the approximation strategy apparently sacrifices power, particularly for the loci harboring multiple QTNs, we expect good overall power in molecular QTL discovery based on the prevailing empirical evidence that the vast majority of the candidate loci contain no more than a single QTN (i.e., Observation 1). Furthermore, we note that the explicit assumption of $K = 1$ is not more conservative than the implicit assumption made by the state-of-the-art minimum p-value-based approach. Moreover, the approximation enables highly efficient computation for extremely large volumes of molecular QTL data. More importantly, it provides a principled approach to evaluate and incorporate genomic annotations into the discovery procedure, which can greatly boost the statistical power. Nevertheless, we provide a natural extension in section 2.4 to relax the assumption of $K = 1$ per locus.

2.4. *Extension to Allow Multiple QTNs per Locus.* In this section, we outline a practical strategy to allow multiple QTNs per locus in our computational procedure. The key idea is to segment each candidate genomic locus into roughly independent linkage disequilibrium (LD) blocks and instead assume $K \leq 1$ per LD block. Within each LD block, the genetic variants are typically highly correlated. Consequently, even if multiple independent association signals co-exist among tightly linked genetic variants, they are practically not identifiable. It is therefore reasonable to assume at most a single QTN per LD block.

Suppose that locus l is partitioned into M disjoint LD blocks, and we denote the partition by $\{\boldsymbol{\gamma}_{l,[k]} : k = 1, 2, ..., M\}$. We show (in Appendix D of the Supplementary Materials) that the posterior probability, $\Pr(\boldsymbol{\gamma}_l \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\eta})$, can be approximated by

$$\Pr(\boldsymbol{\gamma}_l \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\eta}) \approx \prod_{k=1}^{M} \Pr(\boldsymbol{\gamma}_{l,[k]} \mid \boldsymbol{y}_l, \boldsymbol{G}_{l,[k]}, \boldsymbol{D}_{l,[k]}, \boldsymbol{\eta}). \tag{12}$$

Briefly, this is because our priors are independent across SNPs, and in addition, it can be shown that

$$\mathrm{BF}(\boldsymbol{\gamma}_l) \approx \prod_{k=1}^{M} \mathrm{BF}(\boldsymbol{\gamma}_{l,[k]}).$$

based on the analytic result from Wen (2014).

In the EM algorithm for estimating $\hat{\boldsymbol{\eta}}$, the approximation (8) allows directly adopting the proposed computational strategy by treating each LD block as the unit for analysis. To compute the approximate false discovery probability for locus l, we again apply (12) by noting

$$\Pr(\boldsymbol{\gamma}_l = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \hat{\boldsymbol{\eta}}) \approx \prod_{k=1}^{M} \Pr(\boldsymbol{\gamma}_{l,[k]} = \mathbf{0} \mid \boldsymbol{y}_l, \boldsymbol{G}_{l,[k]}, \boldsymbol{D}_{l,[k]}, \hat{\boldsymbol{\eta}}) \tag{13}$$

A working receipt for segmenting genomic regions into LD blocks has been proposed and made available by Berisa and Pickrell (2016). For the European population, their algorithm segments the human autosomal genome into 1,700 roughly independent LD blocks with an average size of 1.6 Mb. The size of the LD block is indeed very similar to the size of a single candidate genomic locus interrogated in typical molecular QTL mapping. For this reason, we do not further segment the candidate loci in the applications discussed in this paper.

2.5. *Extension to Multivariate Quantitative Traits.* Thus far, our description of the proposed method has focused on univariate quantitative traits, e.g., gene expressions and DNA methylation measurements. Our framework can be straightforwardly extended to applications in which the quantitative trait is measured by multivariate variables, e.g., in the case of using ATAC-seq data to quantify chromatin accessibility. To accommodate multivariate quantitative trait data, we simply replace the model (1) by a multivariate linear regression model, which naturally accounts for the correlations between multiple components of the trait. In the example of ATAC-seq data, the response variable for each individual at each locus can simply be described by a row vector with each entry representing the sequencing read counts from a pre-defined genomic window. To perform the Bayesian FDR control, it only requires adjusting the single SNP association Bayes factor according to the modified multivariate linear regression model, and such results are available in the literature (Wen, 2014).

2.6. *Extension to QTL Data Composing Multiple Heterogeneous Groups.* Molecular QTL data collected from multiple heterogeneous sources have become increasingly available (Ardlie et al., 2015; Barreiro et al., 2012; Maranville et al., 2011; Wen, Luca and Pique-Regi, 2015). Joint analysis of QTL data across multiple heterogeneous groups not only improves the power of identifying consistent QTL signals across groups (Flutre et al., 2013; Wen, Luca and Pique-Regi, 2015) but also helps to correctly map group-specific

QTL signals (Barreiro et al., 2012; Flutre et al., 2013; Maranville et al., 2011). Utilizing the previous statistical results from computing Bayes factors from heterogeneous subgroups (Wen and Stephens, 2014), the proposed approach can be straightforwardly applied in those scenarios for QTL discovery while integrating genomic annotations.

3. Simulation Study.

We perform a series of realistic simulation studies to demonstrate the power, robustness and computational efficiency of the proposed statistical procedure in molecular QTL discovery.

3.1. *QTL Discovery without Annotation.*

In the first simulation, we generate genome-wide eQTL data sets assuming no influence from any genomic feature. Our goal is to evaluate the performance of the proposed Bayesian procedure under the baseline scenario and to compare it with the commonly applied standard permutation-based approach.

We select 11,761 protein coding and linc-RNA genes from the GEUVADIS project (Lappalainen et al., 2013) and the genotype data from 343 European individuals. For each gene, we randomly select 50 *cis*-SNPs with a minor allele frequency of ≥ 0.05. With probability $1 - \pi_0$, we randomly assign 1 to 3 eQTNs. Given the eQTNs for each gene, we simulate the expression levels using a multiple regression model (Appendix E of the Supplementary Materials). We generate 20 data sets for each π_0 value and vary the value of π_0 from 0.1 to 0.9.

Without annotation information, the Bayesian hierarchical model is reduced to a simple form with a single parameter in the logistic prior, which assumes *a priori* each candidate SNP independently and equally likely to be the causal eQTL. For comparison, we analyze the simulated data sets using the software package eGENE-MVN (Sul et al., 2015). This package implements the gold-standard QTL discovery method that uses the minimum single SNP association p-value in a locus as the test statistic; however, it finds the corresponding locus-level p-value in a considerably more efficient manner. After obtaining the locus-level p-values, we perform FDR control and identify QTLs using Storey's q-value method.

The simulation results (Table 1) indicate that both TORUS and the gold-standard approach control FDR at the desired level and that their powers are very similar across all π_0 levels. As discussed in section 2.3, because both methods, explicitly or implicitly, assume $K = 1$ for the alternative scenario, it is expected that they are both overly conservative when π_0 is low (in such scenarios, the proposed Bayesian approach appears to be more conservative than the frequentist approach); when π_0 is large, the "one causal SNP per locus" assumption becomes closer to the truth, and the realized false discovery rates achieve the desired control level for both methods (in such cases, the frequentist approach appears to be more conservative than the proposed Bayesian approach).

In addition, to examine the robustness of the proposed approach, we reanalyze the simulated data but include the annotation of SNP distance to TSS (details described in simulation study II). As expected, the enrichment analysis indicates little impact of the annotation to the eQTLs in the simulated data set (due to our simulation scheme), and the results for eQTL discovery remain virtually identical.

Most importantly, our computational time benchmark shows that the proposed Bayesian method is considerably more efficient: to analyze a single simulated data set on a Linux box with an 8-core Intel Xeon 2.13 GHz CPU, the average running time for the Bayesian method is approximately 2 minutes 25 seconds (with 12 parallel processing threads); in comparison, eGENE-MVN requires approximately 3 hours and 45 minutes (also with 12 parallel threads) for the same computational task.

Table 1

Comparison of TORUS and the gold-standard minimum p-value-based QTL discovery procedure without genomic annotations using simulated eQTL data. The realized FDR and power are computed by averaging over the analysis results of 20 simulated QTL data sets at the genome-wide scale. Both methods control the desired FDR level at 5% in all settings. The standard method achieves slightly higher power when the proportion of candidate loci being QTLs (i.e., $1 - \pi_0$) is high, whereas the proposed Bayesian procedure yields slightly better power when the proportion is low. In all cases, however, the powers are comparable. In addition, we perform the proposed approach using SNP distance to TSS as an annotation, and the results remain virtually identical, as expected.

	TORUS without annotation		Minimum p-value Method	
π_0	FDR	Power	FDR	Power
0.10	0.010	0.842	0.024	0.864
0.33	0.028	0.807	0.038	0.810
0.50	0.038	0.801	0.040	0.789
0.67	0.045	0.767	0.047	0.743
0.90	0.049	0.739	0.049	0.701

3.2. *QTL Discovery with Annotation.* Our second simulation study attempts to mimic a commonly observed phenomenon in *cis*-eQTL mapping: eQTL signals tend to cluster around transcription start sites of the corresponding target genes and rapidly decrease away from TSS. We use the same set of 11,761 genes from the GUEVADIS project but include all SNPs within a 1 Mb radius from the TSS of each gene. On average, there are 5,856 SNPs per gene (median of 5,892). We do not impose any restrictions on the minor allele frequencies of the *cis*-SNPs and take all the genotypes directly from the GUEVADIS project. During the simulation, causal eQTL SNPs are randomly assigned by a probability computed from a continuous function of SNP distance to TSS (DTSS, measured in kb and denoted by d), i.e.,

$$p(d) = \mu e^{-\lambda\,|d|}, \tag{14}$$

where λ controls the rate of decay in the expected number of causal eQTL SNPs away from TSS and μ determines the overall expected number of *cis*-eQTLs. We experiment with two different λ values, $\lambda = 0.02$ and $\lambda = 0.1$, corresponding to relatively modest and fast rates of decay, respectively. We then set the μ values to keep the overall expected number of causal eQTL SNPs comparable across the schemes. Note that our simulation function (14) is not compatible with the functional form of our logistic prior (2).

For each simulation setting, we generate 20 data sets and analyze each data set with and without incorporating DTSS information. We do not run the minimum p-value-based approach on these considerably larger data sets due to the high computational cost. Nevertheless, we fully expect its performance to be similar to that of the proposed Bayesian approach ignoring DTSS information based on our evaluation in the first simulation study. When utilizing DTSS information, we follow the approaches used in Degner et al. (2012) and Veyrieras et al. (2008) to dissect the genome into variable sizes of distance bins. In general, we use smaller sized bins in the close vicinity of TSS and larger sized bins away from TSS. The details on the binning of SNPs are presented in Appendix F of the Supplementary Materials.

Our results indicate that by estimating the enrichment parameters, the Bayesian approach effectively characterizes the impact of the DTSS on the eQTL abundance. The estimation of the eQTL signal decay rates with respect to TSS is quite accurate (Figure 1), although our estimation model is very different than the data generative model. We also find that the baseline prevalence (which corresponds to the parameter μ in (14)) is slightly underestimated, which results in the estimates of the SNP-level priors and the FDR control being overly conservative. This result is likely because of the combination of our approximation strategy and the relatively small sample size. Utilizing the highly informative quantitative priors substantially improves the power of eQTL discovery (Table 2). For the modest decay rate, incorporating DTSS in eQTL discovery results in a 15% (or 7 percentage point) power gain, whereas in the fast decay case, we find that there is a 25% (or 10 percentage point) increase in power, which results in correctly discovering ~ 1000 more eQTLs/eGenes on average.

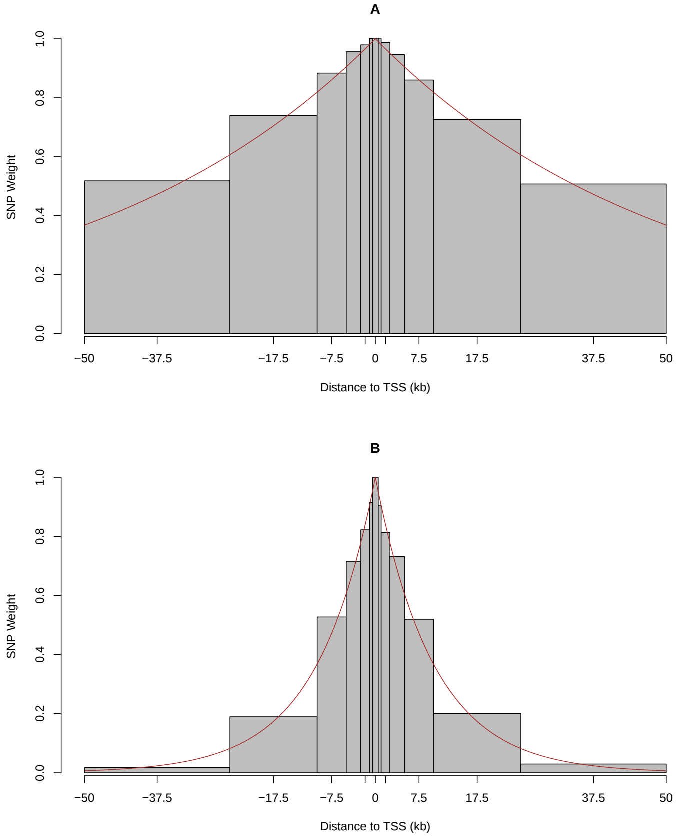

Fig 1. *TORUS estimates of eQTL signal decay rates with respect to DTSS in simulations. Panels A and B plot the estimates by the EM algorithm for the modest and fast decay rates, respectively. Each bar in the plot represents a distance bin. To determine the height of the bar, we compute the prior association probability of a SNP located in the corresponding distance bin by plugging in the MLEs (averaged over 20 simulated data sets) using equation (2). We then normalize the resulting probabilities with respect to the center bin such that the center bar always has a weight of 1. For visualization purposes, we choose to highlight the 100 kb region centered around TSS. The red lines in both panels denote the true decay rate according to the generating functions. It is clear that the enrichment estimates from the EM algorithm capture the overall patterns of the decay effect quite accurately.*

4. Real Data Application: Analysis of GTEx eQTL Data. We analyze the eQTL data sets from the GTEx project (release version 6), which consist of genotype and expression phenotype data from 44 human tissues. The sample sizes in this data release vary from 70 (uterus) to 361 (muscle skeletal). The genotype and expression data have been subjected to the

TABLE 2

Comparison of QTL discovery with and without incorporating genomic annotations using simulated eQTL data. We simulate the eQTL data sets such that the majority of QTN signals decay according to the function $p(d)$. The annotation model uses the SNP distance to TSS as annotations, whereas the baseline model does not. For both the modest and rapid decay functions, we observe a substantial power gain by incorporating relevant annotations into the QTL discovery.

	Baseline Model		Annotation Model	
Decay Function	FDR	Power	FDR	Power
$p(d) = 0.005\, e^{-0.02\,\lvert d\rvert}$	0.006	0.468	0.009	0.538
$p(d) = 0.02\, e^{-0.1\,\lvert d\rvert}$	0.010	0.406	0.009	0.509

standard quality control protocols performed by the GTEx consortium. We download the summary-level statistics, $\hat{\beta}, \text{se}(\hat{\beta})$, for each gene-SNP pair computed by the software package MatrixEQTL (Shabalin, 2012) directly from the GTEx portal. The GTEx portal also provides a list of eQTLs/eGenes for each tissue obtained by the gold-standard minimum p-value approach using permutation and Storey's q-value procedure.

We first run the proposed Bayesian method at the baseline without using any annotations to identify eQTLs at the FDR 0.05 level, and the result is shown in Figure 2A. Compared with the permutation result, it displays a pattern that is very similar to what we observed in the first simulation study: at the baseline level, the Bayesian method appears to be optimal when the detectable QTL signals are overly low, whereas when the detectable signals are high, it performs slightly worse than the gold-standard approach.

We then include the SNP DTSS annotations into the hierarchical model and re-analyze the data using the proposed QTL discovery procedure. We find that the eQTL discovery is uniformly improved: in each single tissue, incorporating DTSS yields more eQTLs than using either the baseline model or the gold-standard permutation approach. On average, we discover 1,475 more eQTLs/eGenes per tissue compared with the gold-standard approach when incorporating DTSS information in the hierarchical model. Most importantly, we find great concordance between the eQTLs/eGenes discovered: on average, 93% of the eQTLs discovered by the gold-standard permutation procedure are also identified by TORUS.

Computationally, both analyses by TORUS complete within 1 hour of running time for a single tissue. On a distributed computing cluster, the full analysis for all 44 tissues takes less than 12 hours.

5. Discussion. In this article, we have introduced a powerful statistical approach for discovering molecular QTLs using high-throughput sequencing

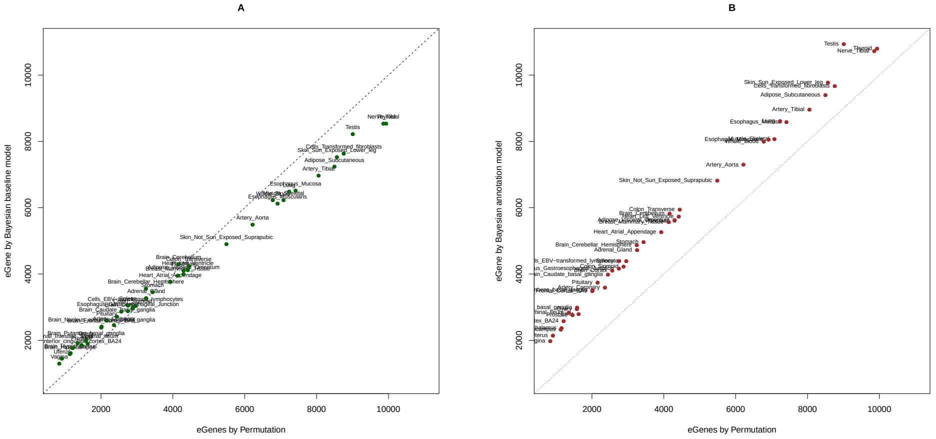

Fig 2. *eQTL discovery from GTEx data by TORUS. We plot the number of eQTLs discovered by TORUS versus the minimum p-value approach in each tissue. Each point represents a single tissue. Panels A and B present the TORUS results in the baseline and by incorporating DTSS annotations, respectively. The pattern observed in panel A is very similar to what we observed in the first simulation study. With the incorporation of DTSS annotations, TORUS discovers more eQTLs in all tissues.*

data and dense genotype data. Through a combination of theoretical derivations, simulation studies and real applications, we have demonstrated that i) our proposed novel approach rigorously controls pre-defined false discovery rates in QTL discovery; ii) by naturally integrating highly informative genomic annotation, the proposed approach consistently exhibits superior power compared with the current gold-standard approaches; and iii) our implementation of the proposed statistical methods exhibits superb computational efficiency and is several hundreds times faster than the standard approach by avoiding extensive permutations.

The proposed Bayesian hierarchical model naturally integrates genomic annotations into QTL mapping in an elegant probabilistic framework. Under this framework, Bayesian FDR control becomes a natural choice for solving the arising multiple hypothesis testing problem. In comparison, it is less straightforward to employ a p-value-based approach while taking full advantage of informative information from genomic annotations, if not impossible.

One of our main statistical contributions in this paper is the proposed analytic approximation framework for efficient evaluations of posterior probabilities required by the Bayesian FDR control procedure. Traditionally, the complexity of the computation has been a major obstacle that prevents the use of Bayesian FDR control in large-scale genomic applications. As we have shown in the paper, the exact Bayesian inference is intractable, and the sampling-based numerical solutions, e.g., MCMC or Monte Carlo EM algorithms, documented in the computational statistics literature also do not scale up to the genome-wide applications of molecular QTL mapping. Our solution of using analytic approximations strikes a desired balance between the statistical and computational efficiency: the proposed algorithm scales well with genome-wide molecular QTL data and rigorously controls the desired level of FDR; most importantly, in practice, we observe that incorporating relevant genomic information through the proposed procedure significantly outweighs the price for loss of statistical efficiency in pursing a conservative approximation for u_l.

The approximate independence among molecular phenotypes is a critical assumption that we have made for multiple testing in molecular QTL discovery as well as for the enrichment parameter estimation in the EM algorithm. In reality, directly measured molecular phenotypes (e.g., expression levels of different genes) are often interacting through biological pathways or networks and are hence correlated. Similarly, batch effects can also artificially introduce a dependence structure among simultaneously measured molecular phenotypes. In both cases, the correlations can be effectively accounted for by techniques of latent factor modeling (Carvalho et al., 2012; Leek and Storey, 2007). Primarily aimed to control for undesired batch effects, it is now a standard practice that molecular phenotypes are pre-processed by latent factor controlling procedures such as factor analysis and surrogate variable analysis. As a by-product, such methods can lead to a significant reduction in the (global) correlations among the processed/transformed molecular phenotypes. It then becomes reasonable to assume the approximate independence of the (processed) molecular phenotypes used in the QTL mapping. In our simulations and examples shown in this paper, we use the software package PEER (Stegle et al., 2012), an implementation of the factor analysis model, to regress out latent factors and observed controlled variables and take the residuals for the subsequent *cis*-eQTL mapping. Finally, we acknowledge that small clusters of correlated phenotypes may still be present even after the pre-processing procedure. In the EM algorithm, the existence of such a local correlation structure essentially reduces the *independent* instances of the sharing of the enrichment parameters. However, given the overall large number of molecular phenotypes in a typical genome-wide experiment and the lack of a global correlation structure, we expect that the estimation efficiency for the enrichment parameters is still sufficiently high.

The SNP distance to TSS is probably the most convenient genomic annotation. Nevertheless, we have demonstrated that the proper use of DTSS helps to resolve a long-standing dilemma in *cis*-eQTL mapping: the choice of the *cis*-region length. It is well known that most *cis*-eQTL signals are clustered around TSS and become sporadic away from it. This finding appears to suggest that one should focus on a relatively narrow *cis* region (e.g., $\sim$ 100 kb) to reduce the multiple testing burden and discover more eGenes. However, such an approach will inevitably miss some distant yet strong signals, and the accumulative loss of signals across all genes can be severe. In our proposed approach, we select a rather large *cis* region and use the enrichment analysis to assess a prior weight of each SNP by their DTSS. Consequently, neighboring SNPs of TSS are up-weighted, and distant SNPs are relatively down-weighted. This weighting naturally solves the dilemma: the focus is on close-by SNPs, but strong distant signals can still overcome the prior weighting penalty and be uncovered.

In practice, we also note that the accuracy of the annotations can have a profound impact on QTL discovery. Within our proposed inference framework, the inaccuracy of the annotations intuitively leads to some undesired shrinkage of the enrichment estimates for corresponding features toward 0. Consequently, the poor quality of the annotations diminishes the benefit of the integrative QTL mapping (although there should be no inflation of type I errors). In our future work, we aim to extend our current framework to account for potential imprecision/uncertainties in genomic annotations.

Finally, we want to emphasize that, in our view, QTL discovery is not the endpoint of the molecular QTL analysis. Rather, it primarily serves as a screening procedure to prioritize a subset of candidate loci that are highly likely to harbor causal trait-associated variants – a strategy that is well demonstrated and widely applied in genome-wide association studies (GWAS). It is natural to follow up with a fine-mapping analysis on the identified molecular QTLs. The fine-mapping analysis is typically framed as a (Bayesian) variable selection problem (Wen, 2014) and serves two purposes: first, it narrows the candidates of causal SNPs (e.g., by constructing a credible set for each causal SNP); second, it identifies potentially multiple independent association signals. Our QTL discovery approach naturally connects to the downstream (Bayesian) multi-SNP fine-mapping analysis by supplying the necessary SNP-level priors for fine-mapping analysis based on the prior model (2) and the point estimate of the enrichment parameter.

6. Supplementary Material and Software Distribution. The GTEx summary-level statistics can be downloaded from the GTEx portal (`http://www.gtexportal.org/home/`). The simulation scripts and the software package TORUS (including source code) can be downloaded from `https://github.com/xqwen/torus/`

7. Acknowledgment. We thank Seungguen Lee for helpful discussions and the area editor and the two anonymous referees for their valuable comments. This work is supported by NIH grants HG007022 (PI G. Abecasis) and MH101825 (PI M. Stephens).

References.

Ardlie, K. G., Deluca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M. et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** 648–660.

Ball, R. D. (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics* **159** 1351–1364.

Banovich, N. E., Lan, X., McVicker, G., Van de Geijn, B., Degner, J. F., Blischak, J. D., Roux, J., Pritchard, J. K. and Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLOS Genetics* **10** e1004663.

Barreiro, L. B., Tailleux, L., Pai, A. A., Gicquel, B., Marioni, J. C. and Gilad, Y. (2012). Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proceedings of the National Academy of Sciences of the United States of America* **109** 1204–1209.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **57** 289–300.

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32** 283–285.

Breitling, R., Li, Y., Tesson, B. M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L. V., De Haan, G., Su, A. I. et al. (2008). Genetical genomics: spotlight on QTL hotspots. *PLoS Genet* **4** e1000232.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2012). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association.*

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138** 963–971.

ENCODE Project Consortium et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

De la Cruz, O., Wen, X., Ke, B., Song, M. and Nicolae, D. L. (2010). Gene, region and pathway level analyses in whole-genome studies. *Genetic epidemiology* **34** 222–231.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E. et al. (2012). DNase [thinsp] I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482** 390–394.

Ding, Z., Ni, Y., Timmer, S. W., Lee, B.-K., Battenhouse, A., Louzada, S., Yang, F., Dunham, I., Crawford, G. E., Lieb, J. D. et al. (2014). Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLOS Genetics* **10** e1004798.

Doerge, R. W. and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142** 285–294.

Flutre, T., Wen, X., Pritchard, J. and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLOS Genetics* **9** e1003486.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J. et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330.

Lappalainen, T., Sammeth, M., Friedländer, M. R., T Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G. et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501** 506–511.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3** e161.

Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics* **10** 422–439.

Maranville, J. C., Luca, F., Richards, A. L., Wen, X., Witonsky, D. B., Baxter, S., Stephens, M., Di Rienzo, A. and Gibson, G. (2011). Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLOS Genetics* **7** e1002162.

Marin, J.-M. and Robert, C. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Science & Business Media.

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y. and Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science* **342** 747–749.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays. *Journal of the American Statistical Association* **99** 990–1001.

Neto, E. C., Keller, M. P., Broman, A. F., Attie, A. D., Jansen, R. C., Broman, K. W. and Yandell, B. S. (2012). Quantile-based permutation thresholds for quantitative trait loci hotspots. *Genetics* **191** 1355–1365.

Neto, E. C., Broman, A. T., Keller, M. P., Attie, A. D., Zhang, B., Zhu, J. and Yandell, B. S. (2013). Modeling causality for pairs of phenotypes in system genetics. *Genetics* **193** 1003–1013.

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–76.

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y. and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research* **21** 447–455.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLOS Genetics* **3** e114.

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

Sillanpää, M. J. and Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151** 1605–1619.

Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **7** 500–507.

Stephens, D. and Fisch, R. (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* 1334–1347.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **31** 2013–2035.

Sul, J. H., Raj, T., de Jong, S., de Bakker, P. I., Raychaudhuri, S., Ophoff, R. A., Stranger, B. E., Eskin, E. and Han, B. (2015). Accurate and Fast Multiple-Testing Correction in eQTL Studies. *The American Journal of Human Genetics* **96**.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M. and Pritchard, J. K. (2008). High-resolution mapping of expression-

QTLs yields insight into human gene regulation. *PLOS Genetics* **4** e1000214.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genetic epidemiology* **33** 79–86.

Wen, X. (2011). Bayesian analysis of genetic association data, accounting for heterogeneity PhD thesis, The University of Chicago.

Wen, X. (2014). Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics* **70** 73–83.

Wen, X. (2015). Bayesian model comparison in genetic association analysis: linear mixed modeling and SNP set testing. *Biostatistics* kxv009.

Wen, X., Luca, F. and Pique-Regi, R. (2015). Cross-population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLOS Genetics* **11** e1005176.

Wen, X. and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions. *The Annals of Applied Statistics* **8** 176–203.

Yi, N., Yandell, B. S., Churchill, G. A., Allison, D. B., Eisen, E. J. and Pomp, D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170** 1333–1344.

Department of Biostatistics
University of Michigan
1415 Washington Heights
Ann Arbor, Michigan 48109
USA
E-mail: xwen@umich.edu