Supplementary Materials for “Molecular QTL Discovery Incorporating Genomic Annotations using Bayesian False Discovery Rate Control”

Xiaoquan Wen

Department of Biostatistics, University of Michigan

A Model Details and Bayes Factor Evaluation

In this section, we provide the full details of the prior specification for the linear model (1) in the main text.

In general, we consider the effect sizes of the regression coefficients at the scale of signal-to-noise ratios with respect to τ_l. We assign a normal prior to the intercept term μ_l, i.e.,

$$\mu_l \sqrt{\tau_l} \sim \mathrm{N}(0, \psi^2).$$

When performing inference, we let $\psi \to \infty$ and essentially assume a flat prior for the intercept. This prior provides no shrinkage effect, and allows the intercept term absorbing much of variation in $\boldsymbol{y}_l$. Conditional on $\gamma_{l_i} = 1$, we also assign a normal prior to the genetic effect β_{l_i}. More specifically, we assume

$$\beta_{l_i} \sqrt{\tau_l} \mid \gamma_{l_i} = 1 \, \sim \, \mathrm{N}(0, \phi^2).$$

In particular, the prior genetic heritability explained by the single SNP association of l_i can be computed by

$$\frac{2 f_{l_i} (1 - f_{l_i}) \phi^2}{1 + 2 f_{l_i} (1 - f_{l_i}) \phi^2},$$

where f_{l_i} represents the allele frequency of the SNP l_i. Finally, we assign an inverse gamma prior to τ, i.e.,

$$\tau \sim \Gamma(\kappa/2, \lambda/2),$$

and in inference, we consider the limiting form $\kappa, \lambda \to 0$.

Given ϕ and taking limits with respect to the hyper-parameters ψ, κ and λ. Many authors have shown that the Bayes factor of the linear model $\mathrm{BF}(\boldsymbol{\gamma}_l, \phi)$ can be analytically computed (Servin and Stephens, 2007) or approximated (Wakefield, 2009, Wen, 2014). Although a single ϕ value would allow a wide range of effect sizes (from $-\infty$ to ∞), we use a mixture normal distribution to mimic the long-tailed effect size distribution observed in practice by considering L different ϕ values in a grid. Finally, we evaluate the Bayes factor with respect to $\boldsymbol{\gamma}_l$ by averaging over L different ϕ_k values, i.e.,

$$\mathrm{BF}(\boldsymbol{\gamma}_l) = \sum_{k=1}^{L} \pi_k \mathrm{BF}(\boldsymbol{\gamma}_l, \phi_k),$$

where π_k denote the weight on grid value ϕ_k. Typically, we select a grid of ϕ values at the heritability scale informed by the practical observations in relevant genetic association analysis.

B Bayesian FDR Control and Connection to Frequentist Approaches

In the context of QTL discovery, the multiple testing problem can be framed as a binary decision problem with respect to $\boldsymbol{\gamma}_l$ for $l = 1, ..., L$. We define a binary indicator, Z_l, for each locus l and set $Z_l = 1$ if $\boldsymbol{\gamma}_l \neq \mathbf{0}$ and 0 otherwise. Further, we denote the collection of the observed phenotype-genotype data by $\boldsymbol{Y}$. Let the function $\delta_l(\boldsymbol{Y})$ denote a decision (0 or 1) on Z_l based on the observed data, and define the total discoveries by $D := \sum_{l=1}^{L} \delta_l$. Following the formulation of Müller *et al.* (2006), the False Discovery Proportion (FDP), which is also a random variable, can be defined as the proportion of false discoveries among total discoveries, i.e.,

$$\mathrm{FDP} := \frac{\sum_{l=1}^{L} \delta_l (1 - Z_l)}{D \vee 1}. \tag{B.1}$$

Recall,

$$u_l := \Pr(Z_l = 0 \mid \boldsymbol{Y}) = 1 - \mathrm{E}(Z_l \mid \boldsymbol{Y}); \tag{B.2}$$

thus, the Bayesian False Discovery Rate is naturally defined as

$$\mathrm{BFDR} := \mathrm{E}(\mathrm{FDP} \mid \boldsymbol{Y}) = \frac{\sum_{l=1}^{L} \delta_l \, u_l}{D \vee 1}, \tag{B.3}$$

where the conditional expectation is taken with respect to $\boldsymbol{Z} := (Z_1, \ldots, Z_L)$. Moreover, the frequentist control of the False Discovery Rate focuses on the quantity

$$\mathrm{FDR} := \mathrm{E}(\mathrm{FDP}) = \mathrm{E}\left[\mathrm{E}(\mathrm{FDP} \mid \boldsymbol{Y})\right], \tag{B.4}$$

where the additional expectation is taken with respect to $\boldsymbol{Y}$ over (hypothetically) repeated experiments. It is important to note that controlling the Bayesian FDR is a *sufficient but not necessary* condition to control the frequentist FDR; thus the Bayesian FDR control is more stringent in theory.

As demonstrated by Newton *et al.* (2004) and Müller *et al.* (2006), the Bayesian FDR control is based on the following natural decision rule

$$\delta_l^*(t) = I\left(u_l < t\right). \tag{B.5}$$

For a pre-defined FDR level α, the threshold t_α in (B.5) is determined by

$$t_\alpha = \arg\min_t \left(\frac{\sum_{l=1}^{L} \delta_l^*(t)\, u_l}{D(t) \vee 1} \leq \alpha \right), \tag{B.6}$$

i.e., the rejection set determined by the threshold t_α is the largest such set with the averaged false discovery probability $\leq \alpha$. In practice, we use the following simple algorithm proposed by (Newton *et al.*, 2004) to determine t_α:

1. *sort* u_l*'s in ascending order: i.e.,* $u_{(1)} \leq u_{(2)} \leq ... \leq u_{(L)}$.

2. *start from* $m = 1$ *and compute the partial mean using the sorted sequence of* $\{u_{(l)}\}$ *for* $s_m = \sum_{l=1}^{m} u_{(l)}/m$

3. *stop if* $s_m > \alpha$

4. $t_\alpha = u_{(m-1)}$ *if* $m > 1$*, and 0 otherwise; and reject the hypotheses corresponding to* $u_{(1)}, ..., u_{(m-1)}$.

C EM Algorithm for Estimating Enrichment Parameters

C.1 EM Algorithm Details

In this section, we outline the EM algorithm to estimate the enrichment parameter $\boldsymbol{\alpha}$. We denote $\boldsymbol{\Gamma} := \{\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_L\}$, $\boldsymbol{D} := \{\boldsymbol{D}_1, ..., \boldsymbol{D}_L\}$ and $\boldsymbol{G} = \{\boldsymbol{G}_1, ..., \boldsymbol{G}_L\}$. By treating $\boldsymbol{\Gamma}$ as missing data, we obtain the complete data likelihood by the following

$$\begin{aligned} P(\boldsymbol{Y}, \boldsymbol{\Gamma} \mid \boldsymbol{G}, \boldsymbol{D}, \boldsymbol{\alpha}) &= \Pr(\boldsymbol{\Gamma} \mid \boldsymbol{D}, \boldsymbol{\alpha}) P(\boldsymbol{Y} \mid \boldsymbol{G}, \boldsymbol{\Gamma}) \\ &= \prod_{l=1}^{L} \Pr(\boldsymbol{\gamma}_l \mid \boldsymbol{D}_l, \boldsymbol{\alpha}) \prod_{l=1}^{L} P(\boldsymbol{y}_l \mid \boldsymbol{G}_l, \boldsymbol{\gamma}_l), \end{aligned} \tag{C.1}$$

where the factorization is based on the conditional independence relationships induced by the hierarchical model. We further re-write the prior probability $\Pr(\boldsymbol{\gamma}_l \mid \boldsymbol{D}_l, \boldsymbol{\alpha})$ using the logistic model,

$$\Pr(\boldsymbol{\gamma}_l \mid \boldsymbol{D}_l, \boldsymbol{\alpha}) = \prod_{i=1}^{p_l} \left[\left(\frac{\exp(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i})}{1 + \exp(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i})} \right)^{\gamma_{l_i}} \left(\frac{1}{1 + \exp(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i})} \right)^{1-\gamma_{l_i}} \right].$$

Therefore, the complete data log-likelihood is given by

$$\log P(\boldsymbol{Y}, \boldsymbol{\Gamma} \mid \boldsymbol{G}, \boldsymbol{D}, \boldsymbol{\alpha}) = \sum_{l=1}^{L} \sum_{i=1}^{p_l} \gamma_{l_i} (\boldsymbol{\alpha}' \boldsymbol{d}_{l_i}) - \sum_{l=1}^{L} \sum_{i=1}^{p_l} \log[1 + \exp(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i})] + \sum_{l=1}^{L} \log P(\boldsymbol{y}_l \mid \boldsymbol{G}_l, \boldsymbol{\gamma}_l).$$

The EM algorithm is initiated at an arbitrary starting point $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(1)}$. In the E-step of the t-th iteration, we compute

$$\begin{aligned}
\mathrm{E}\left[\log P(\boldsymbol{Y}, \boldsymbol{\Gamma} \mid \boldsymbol{G}, \boldsymbol{D}, \boldsymbol{\alpha}) \mid \boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{D}, \boldsymbol{\alpha}^{(t)}\right] = & \sum_{l=1}^{L} \sum_{i=1}^{p_l} \Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i}) \\
& - \sum_{l=1}^{L} \sum_{i=1}^{p_l} \log[1 + \exp(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i})] \\
& + \mathrm{E}\left(\log P(\boldsymbol{y}_l \mid \boldsymbol{\gamma}_l, \boldsymbol{G}_l) \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{\alpha}^{(t)}\right).
\end{aligned}$$

That is, we evaluate the posterior inclusion probability $\Pr(\boldsymbol{\gamma}_{l_i} \mid \boldsymbol{y}_l, \boldsymbol{\alpha}^{(t)})$ for each candidate SNP and for all loci. In the M-step, we find

$$\boldsymbol{\alpha}^{(t+1)} = \arg\max_{\boldsymbol{\alpha}} \left(\sum_{l=1}^{L} \sum_{i=1}^{p_l} \Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i}) - \sum_{l=1}^{L} \sum_{i=1}^{p_l} \log[1 + \exp(\boldsymbol{\alpha}' \boldsymbol{d}_{l_i})] \right).$$

Note that the objective function coincides with the log likelihood function of a logistic regression model with the binary response variable, γ_{l_i}, replaced by its corresponding posterior expectation, $\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})$. For this well-known concave function, the optimization can be achieved by numerical solutions such as Newton-Raphson and iterative re-weighted least square (IRLS) algorithms. Our implementation in TORUS uses the Newton-Raphson algorithm.

An alternative strategy for the M-step in practice is to take advantage of the nature of the posterior probability $\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})$ and adopt a generalized EM algorithm. Specifically, instead of attempting maximizing the likelihood function in the M-step, we find a new set of parameter values, $\boldsymbol{\alpha}^{(t+1)}$ to simply increase the likelihood. To achieve this, we aim to improve the curve fitting for the mean model

$$g\left(\mathrm{E}(\gamma_{l_i})\right) = \boldsymbol{\alpha}' \boldsymbol{d}_{l_i},$$

by a least squares algorithm, where $g(\cdot)$ denote the logistic link function. In particular, we use the posterior expectation, $\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})$, in place of $\mathrm{E}(\gamma_{l_i})$ in each M-step. As Bayesian posterior probabilities, the PIPs are never *exactly* 0 or 1 (although they can approaching 0 or 1 very closely). Mathematically, the solution of the least squares algorithm is the root of the following linear system,

$$\begin{aligned}
& \sum_{l=1}^{L} \sum_{i=1}^{p_l} \left(\mathrm{logit}[\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})] - \boldsymbol{\alpha}' \boldsymbol{d}_{l_i} \right) = 0, \\
& \sum_{l=1}^{L} \sum_{i=1}^{p_l} \left(\mathrm{logit}[\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})] - \boldsymbol{\alpha}' \boldsymbol{d}_{l_i} \right) \cdot d_{l_i 1} = 0, \\
& \qquad \vdots \\
& \sum_{l=1}^{L} \sum_{i=1}^{p_l} \left(\mathrm{logit}[\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)})] - \boldsymbol{\alpha}' \boldsymbol{d}_{l_i} \right) \cdot d_{l_i m} = 0.
\end{aligned} \tag{C.2}$$

In comparison, the solution that maximizes the log-likelihood function is the root of the following non-linear system,

$$\begin{aligned}
&\sum_{l=1}^{L}\sum_{i=1}^{p_l}\left(\Pr(\gamma_{l_i}=1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)}) - \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})}\right) = 0, \\
&\sum_{l=1}^{L}\sum_{i=1}^{p_l}\left(\Pr(\gamma_{l_i}=1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)}) - \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})}\right)\cdot d_{l_i 1} = 0, \\
&\vdots \\
&\sum_{l=1}^{L}\sum_{i=1}^{p_l}\left(\Pr(\gamma_{l_i}=1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}^{(t)}) - \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})}\right)\cdot d_{l_i m} = 0.
\end{aligned} \tag{C.3}$$

In practice given the same PIP input, we find the solutions by the two algorithms for a single M-step are typically close, and importantly the solution by the least squares algorithm does increase the log-likelihood. As a result, both maximization strategies lead to numerically almost identical final result for $\hat{\boldsymbol{\alpha}}$. (Although exceptions exist in the case that the binary responses are extremely imbalance.) The main benefit of the least square algorithm is that it provides an analytic tool for us to understand the effect of approximations of PIPs on the enrichment parameter estimate.

C.2 Approximation of Posterior Inclusion Probability

The computational difficulty of the EM algorithm mainly lies in evaluating the PIPs in the E-step. The situation is the same as computing u_l in Bayesian FDR control where the exact computation is intractable. To ease computation, we apply the same deterministic approximation technique. The key assumption is again that posterior probabilities of single QTN association models dominate the posterior probability space of $\{\boldsymbol{\gamma}\}$ for locus l, i.e.,

$$\frac{\sum_{||\boldsymbol{\gamma}||\le 1}\Pr(\boldsymbol{\gamma}_l=\boldsymbol{\gamma} \mid \boldsymbol{D}_l, \boldsymbol{\alpha})\mathrm{BF}(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'}\Pr(\boldsymbol{\gamma}_l=\boldsymbol{\gamma}' \mid \boldsymbol{D}_l, \boldsymbol{\alpha})\mathrm{BF}(\boldsymbol{\gamma}')} \to 1. \tag{C.4}$$

Note that the model space of $\{\boldsymbol{\gamma} : ||\boldsymbol{\gamma}|| \le 1\}$ contains only the null model, $\boldsymbol{\gamma}_l = \boldsymbol{0}$, and all single-SNP association models. We denote

$$\pi_{l,0} := \Pr(\boldsymbol{\gamma}_l = 0 \mid \boldsymbol{D}_l, \boldsymbol{\alpha}) = \prod_{i=1}^{p_l}\left(1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})\right)^{-1}.$$

Let $\boldsymbol{\gamma}_j^{\circ}$ denote the single-SNP association model where the j-th SNP is the assumed QTN. Clearly,

$$\Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}_j^{\circ} \mid \boldsymbol{D}_l, \boldsymbol{\alpha}) = \exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_j})\prod_{i=1}^{p_l}\left(1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_i})\right)^{-1} = \pi_{l,0}\cdot\exp(\boldsymbol{\alpha}'\boldsymbol{d}_{l_j}),$$

and

$$\mathrm{BF}(\boldsymbol{\gamma}_j^\circ) = \mathrm{BF}_{l_j},$$

and recall that BF_{l_j} denotes the Bayes factor based on the single-SNP analysis of SNP j at locus l. Finally, we note that given the restrained model space, the PIP of SNP j, $\Pr(\gamma_{l_j} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha})$, coincides with the posterior model probability, $\Pr(\boldsymbol{\gamma}_l = \boldsymbol{\gamma}_j^\circ \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha})$. Given all of the above, it follows from the simple algebra that

$$\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha}) \approx \frac{e^{\boldsymbol{\alpha}'\boldsymbol{d}_{l_i}}\,\mathrm{BF}_{l_i}}{1 + \sum_{k=1}^{p_l} e^{\boldsymbol{\alpha}'\boldsymbol{d}_{l_k}}\,\mathrm{BF}_{l_k}}, \tag{C.5}$$

which can be analytically evaluated given $\boldsymbol{\alpha}$.

C.3 Justification for Claim 1

We provide necessary arguments for Claim 1 by considering a scenario that the sample size in QTL mapping is reasonably large. Under such condition, the behavior of the proposed approximation (C.5) is better understood. Without loss of generality, we first assume all candidate SNPs are uncorrelated, i.e., they are in linkage equilibrium (and we will relax this assumption at the end of the section). As a consequence, it follows that the Bayes factor for single SNP association model $\mathrm{BF}_{l_i} \gg 1$ if the SNP l_i is a QTN, and $\mathrm{BF}_{l_i} \to 0$, otherwise.

First, we qualitatively compare the approximate PIP (C.5) with the corresponding value from the exact Bayesian computation. Note that if a locus contains at most a single QTN, the approximation (C.5) is accurate based on the arguments provided in the main text, i.e., the difference between the approximation and the exact calculation is negligible. Here, we mainly focus on the situation where a candidate locus contains multiple QTNs. In such a locus l, if the SNP l_i is *not* associated with the trait of interest, because $\mathrm{BF}_{l_i} \to 0$ given our sample size assumption, the approximation yields that $\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{\alpha}) \to 0$ for any reasonable value of $\boldsymbol{\alpha}$, and this result should be consistent with the exact Bayesian calculation under the same assumption. If the SNP l_i is genuinely associated, we expect that the $\Pr(\gamma_{l_i} = 1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{\alpha}) \to 1$, as the sample size increases. This is no longer the case in the approximation (C.5): although individually, it is expected that $\mathrm{BF}_{l_i} \gg 1$ for any QTN in the locus, the PIP evaluated by (C.5) is determined by the relative magnitude of the Bayes factor (which is a function of the genetic effect size, see Wakefield (2009) for details). Consequently, they are no longer guaranteed to all converge to 1. For example, consider two independent QTNs l_i and l_j with different effect sizes with $\beta_i > \beta_j$, the single SNP Bayes factors follow the relationship with sufficiently large samples,

$$\mathrm{BF}_{l_i} \gg \mathrm{BF}_{l_j} \gg 1.$$

For any reasonable $\boldsymbol{\alpha}$ values, the approximation leads to the PIP of l_i close to 1, whereas the PIP of l_j is clearly < 1. In a different situation that $\beta_i \approx \beta_j$, we expect the approximate PIPs for l_i and l_j are both close to 0.5. In conclusion, the approximate PIPs of QTNs are shrunk towards 0 if a locus harbors multiple association signals.

We then proceed to consider the impact of potentially shrunken PIP values on the empirical Bayes estimates of the prior probabilities. Note that in the EM algorithm, the M-step of the t-th iteration directly provides the prior probabilities for the E-step of the $(t+1)$-th iteration, i.e., for the SNP l_i, the quantity is

$$\frac{\exp[\boldsymbol{\alpha}^{(t+1)\prime}\boldsymbol{d}_{l_i}]}{1+\exp[\boldsymbol{\alpha}^{(t+1)\prime}\boldsymbol{d}_{l_i}]},$$

which also can be considered as the fitted value for l_i in the logistic regression model solved in the t-th M-step. Importantly, these fitted values can be viewed as unbiased estimates of their input PIPs (or unbiased predicted values in a logistic regression). This justified by the first equation in (C.3). Now consider two coupled EM runs starting with the same initial value $\boldsymbol{\alpha}^{(1)}$. The first EM run (EM-exact) uses (hypothetically) exact PIP values from each E-step and is expected to yield the desired unbiased MLE for the enrichment parameter $\boldsymbol{\alpha}$; and the second EM run (EM-appx) applies approximation (C.5) that shrinks PIPs toward 0. Given the shrinkage effect on PIP by the approximation, the priors for EM-appx in next iteration of the E-step are also shrunk toward 0 comparing to EM-exact, especially for the multiple causal SNPs co-existing in a locus. As a consequence, we expect that the approximate PIP values are further biased toward 0 in the next immediate E-step for those SNPs. In general, the joint effects of more conservative prior input and the approximation (C.5) lead to more conservative prior estimates in all future iterations. Let the coupled EM runs end when both runs achieves the pre-defined convergence criteria. It is easy to conclude that EM-appx yields the prior estimates biased toward 0 in comparison to EM-exact. Therefore, the Claim 1 is justified.

In addition, using the GEM algorithm with the least squares described in the section B.2, we explore the impact of the PIP approximation on the individual enrichment parameter estimates. To this end, we further assume that conditional on SNPs being QTNs, their genetic effect sizes are no longer relevant to their genomic annotations. Furthermore, we assume a more explicit approximate relationship between the approximate PIP by (C.4), $\widehat{\text{PIP}}_{l_i}(\boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha})$, for SNP l_i with its true value at the logistic scale by

$$\text{logit}[\widehat{\text{PIP}}_{l_i}(\boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha})] \approx \text{logit}\left[\Pr(\gamma_{l_i}=1 \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{D}_l, \boldsymbol{\alpha})\right] - \Delta_{l_i}(\boldsymbol{\beta}_l), \text{ and } \Delta_{l_i}(\boldsymbol{\beta}_l) \geq 0. \quad \text{(C.6)}$$

More specifically, (C.6) assumes that the effect of reasonable $\boldsymbol{\alpha}$ values on the shrinkage effect is negligible, which generally holds in practice for reasonably large sample size. By plugging (C.6) into the GEM procedure, it should be clear that in each M-step with the least squares implementation, the shrinkage effect by the PIP approximation is orthogonal to the annotation $\boldsymbol{D}_l$ and absorbed only by the intercept term, α_0, in the enrichment parameter which is under-estimated; whereas $\alpha_1, ..., \alpha_m$ remain roughly intact. As a consequence, by considering two coupled GEM runs with the exact and the approximate PIPs, we conclude that the approximation (C.5) leads to a downward biased estimate of α_0, while the estimates of $\alpha_1, ..., \alpha_m$ remain unbiased.

Finally, we address the issue of LD on the PIP calculation. In brief, the presence of LD may result in that γ_{l_i} becomes non-identifiable in association analysis (Guan and Stephens, 2011). Consider two perfectly correlated SNPs, however only one of the two is the actual QTN. Based on genetic association data alone, the two SNPs are indistinguishable. Consequently, each is supposed to be assessed PIP = 0.50 asymptotically. The interpretation is that although we are almost certain one of the two SNPs is the QTN (the sum of the PIPs → 1), there is no information to distinguish the two. In general, in the presence of LD, instead of expecting that single SNP PIP values converge to 1, we should expect the sum of individual PIPs of a cluster of SNPs in LD converge to 1 in the exact Bayesian computation. Furthermore, the distributive pattern of PIPs within a LD cluster is determined by the degree of genotype correlation with the causal SNP and the relevant genomic annotations. Therefore, the approximation (C.5) should shrink the PIPs of all member SNPs within a cluster close to 0, and the other arguments and our main conclusions remain unchanged.

C.4 Evaluation of Enrichment Parameter Estimate by Simulation

We perform numerical simulations to validate our findings in the previous section. Our simulation setting mimics the application of genome-wide *cis*-eQTL mapping, however at a reduced scale. Specifically, we select a subset of 5,000 random genes from the GEUVADIS data. For each gene, 50 *cis*-SNPs are used in the simulation and we annotate 30% of the SNPs with a binary feature. For each SNP, the association status is determined by a Bernoulli trail with the success (i.e. associated) probability given by

$$p = \frac{\exp(-4 + \alpha_1 d)}{1 + \exp(-4 + \alpha_1 d)},$$

where d is the SNP specific binary annotation value, and α_1 is the true enrichment parameter. Given the true QTNs of each gene, we then apply the scheme described in section E to simulate the effect sizes of the QTNs and the expression levels for the 343 European individuals. We set $\alpha_1 = -0.25, 0.00, 0.25, 0.50, 0.75, 1.00$, and for each α_1 value, we simulate 100 data sets. We use the proposed EM algorithm with approximation (C.5) to analyze the simulated data sets. For comparison, we also estimate α_1 using a logistic regression with the true association status as the outcome variable and the annotations as the predictor. This analysis represents a theoretical best case scenario, and its results should be regarded as the optimal bound for the analyses that infer the latent association status from the genotype-phenotype data. The results (Supplementary Fig. 1) indicate that the EM algorithm based on our posterior approximation scheme consistently yields unbiased estimates for α_1. The decrease of estimation efficiency from the theoretical optimal estimator, represented by the difference of the standard errors, is not large. However, we do find the baseline parameter α_0 is consistently under-estimated, as predicted in the previous section.

D Derivation of Approximation (11)

For a given genomic locus, our goal is to show that

$$\Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) \,\mathrm{BF}(\boldsymbol{\gamma}) \approx \prod_{k=1}^{L} \Pr(\boldsymbol{\gamma}_{[k]} \mid \boldsymbol{\alpha}) \cdot \prod_{k=1}^{L} \mathrm{BF}(\boldsymbol{\gamma}_{[k]}).$$

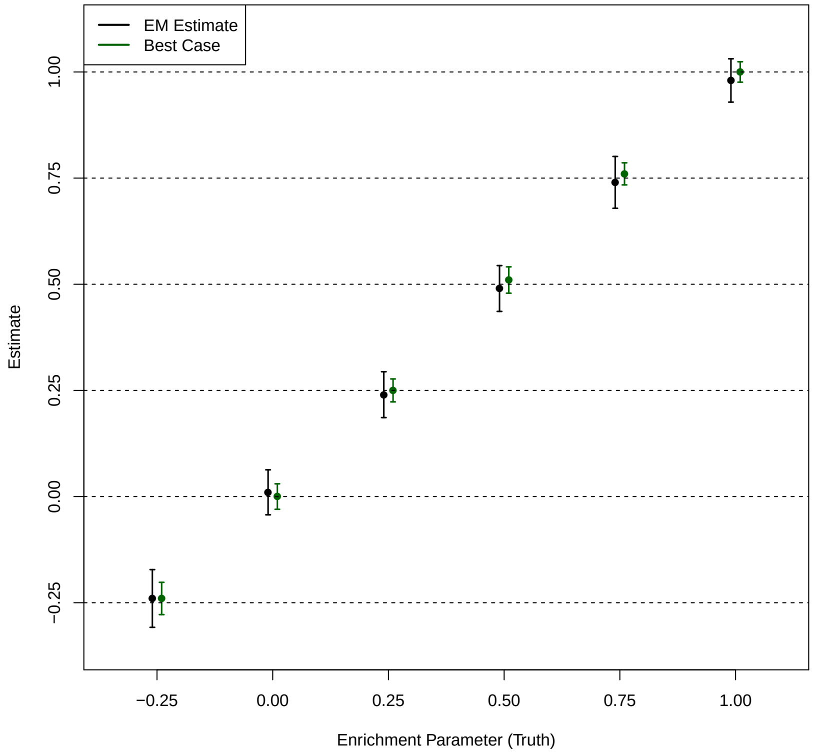

Figure 1: Point estimates of the enrichment parameter in simulations. The point estimate of α_1 ± standard error (obtained from 100 simulated data sets) for each method is plotted for each simulation setting. The "best case" method uses the true association status and represents the optimal performance for any enrichment analysis method. The estimate based on the EM algorithm using the posterior approximation yields unbiased estimate but with larger variation than the optimal method, which is fully expected.

Recall that $\{\boldsymbol{\gamma}_{[k]} : k = 1, 2, 3...\}$ form a partition of the vector $\boldsymbol{\gamma}$. Because the prior probabilities are assumed to be independent across SNPs, it follows trivially that $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) = \prod_{k=1}^{L} \Pr(\boldsymbol{\gamma}_{[k]} \mid \boldsymbol{\alpha})$.

To show that

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

we note the result from Wen (2014),

$$\mathrm{BF}(\boldsymbol{\gamma}) = \int P(\boldsymbol{\beta} \mid \boldsymbol{\gamma})\, \mathrm{BF}(\boldsymbol{\beta})\, d\boldsymbol{\beta},$$

where the probability $P(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ defines the prior effect size given association status $\boldsymbol{\gamma}$. Furthermore, note the independent relationship of the prior effect sizes across SNPs,

$$P(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{p} P(\beta_i \mid \gamma_i).$$

If $\gamma_i = 1$, β_i is assigned a normal prior (the derivation can be trivally extended to the case $\beta_i \mid \gamma_i = 1$ is assigned a mixture normal prior), whereas if $\gamma_i = 0$, $\beta_i = 0$ with probability 1 (or is represented by a degenerated normal distribution, $\beta_i \sim \mathrm{N}(0, 0)$). Equivalently, we write

$$\boldsymbol{\beta} \mid \boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}),$$

where $\boldsymbol{W}$ is a diagonal prior variance-covariance matrix, and for $\boldsymbol{\gamma} \neq \mathbf{1}$, $\boldsymbol{W}$ is singular.

Without loss of generality, we assume that both the phenotype vector $\boldsymbol{y}$ and the genotype vectors $\boldsymbol{g}_1, ..., \boldsymbol{g}_p$ are centered, i.e., the intercept term in the association model is exactly 0. Furthermore, we also assume that the residual error variance parameter τ is known. It then follows from the result of Wen (2014) that

$$\mathrm{BF}(\boldsymbol{\beta}; \boldsymbol{W}) = |\boldsymbol{I} + \tau \boldsymbol{G}'\boldsymbol{G}\boldsymbol{W}|^{-\frac{1}{2}} \cdot \exp\left(\frac{1}{2}\boldsymbol{y}'\boldsymbol{G}\left[\boldsymbol{W}(\boldsymbol{I} + \tau\boldsymbol{G}'\boldsymbol{G}\boldsymbol{W})^{-1}\right]\boldsymbol{G}'\boldsymbol{y}\right). \tag{D.1}$$

This expression provides the theoretical basis for the factorization. In particular, the $p \times p$ sample covariance matrix $\frac{1}{n}\boldsymbol{G}'\boldsymbol{G}$ is a well-known estimate of $\mathrm{Var}(\boldsymbol{G})$. In other words, $\boldsymbol{G}'\boldsymbol{G}$ can be viewed as a noisy observation of $n\mathrm{Var}(\boldsymbol{G})$. Using population genetic theory, Wen and Stephens (2010) show that $\mathrm{Var}(\boldsymbol{G})$ is extremely banded. Based on this result, Berisa and Pickrell (2016) recently provided an algorithm to segment the genome into L non-overlapping loci utilizing the population parameter of the recombination rate, i.e.,

$$\boldsymbol{G} = (\boldsymbol{G}_{[1]}, \ldots, \boldsymbol{G}_{[L]}),$$

and we approximate $\boldsymbol{G}'\boldsymbol{G}$ by a block diagonal matrix

$$\widehat{\boldsymbol{G}'\boldsymbol{G}} = \boldsymbol{G}'_{[1]}\boldsymbol{G}_{[1]} \oplus \cdots \oplus \boldsymbol{G}'_{[L]}\boldsymbol{G}_{[L]}, \tag{D.2}$$

where "$\oplus$" denotes the direct sum of the matrices. It is important to note that (D.2) should be viewed as a de-noised version of $\boldsymbol{G}'\boldsymbol{G}$ with non-zero entries outside the LD blocks shrunk to exactly 0. By plugging (D.2) into (D.1), it follows that

$$\mathrm{BF}(\boldsymbol{\beta}; \boldsymbol{W}) = \prod_{k=1}^{L} \mathrm{BF}_{[k]}, \tag{D.3}$$

where

$$\mathrm{BF}_{[k]} = |\boldsymbol{I} + \tau \boldsymbol{G}'_{[k]}\boldsymbol{G}_{[k]}\boldsymbol{W}_{[k]}|^{-\frac{1}{2}} \cdot \exp\left(\frac{1}{2}\boldsymbol{y}'\boldsymbol{G}_{[k]}\left[\boldsymbol{W}_{[k]}(\boldsymbol{I} + \tau\boldsymbol{G}'_{[k]}\boldsymbol{G}_{[k]}\boldsymbol{W}_{[k]})^{-1}\right]\boldsymbol{G}'_{[k]}\boldsymbol{y}\right). \tag{D.4}$$

In particular, $(\boldsymbol{W}_{[1]}, \ldots, \boldsymbol{W}_{[[L]})$ is a decomposition of the diagonal matrix $\boldsymbol{W}$ compatible with the decomposition of $\boldsymbol{G}$.

Finally, following Wen (2014), we integrate out the residual error variance parameter τ for each $\mathrm{BF}_{[k]}$ by applying the Laplace approximation. This step results in plugging in a point estimate of τ (e.g., based on $\boldsymbol{y}$ and $\boldsymbol{G}_{[k]}$ for each block k) into the expression (D.4). Taken together, we have shown that

$$\mathrm{BF}(\boldsymbol{\gamma}) \approx \prod_{k=1}^{L} \int P(\boldsymbol{\beta}_{[k]} \mid \boldsymbol{\gamma}_{[k]})\, \mathrm{BF}_{[k]}\, d\boldsymbol{\beta}_{[k]},$$

and consequently,

$$\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) \approx \prod_{k=1}^{L} \Pr(\boldsymbol{\gamma}_{[k]} \mid \boldsymbol{y}_l, \boldsymbol{G}_l, \boldsymbol{\alpha}).$$

E Simulation Details

In this section, we provide the details of the scheme for simulating genetic effects of casual QTNs and the individual-level quantitative traits.

As described in the main text, we perform Bernoulli trials for each of the candidate SNPs and determine its association status with the target (expression) quantitative trait. For each causal QTN, we then draw its genetic effect from a Normal distribution, i.e.,

$$\beta_{l_i} \mid \boldsymbol{\gamma}_{l_i} = 1 \sim \mathrm{N}(0, 0.6^2).$$

The individual-level expression levels for locus l are then simulated according to the linear model

$$\boldsymbol{y}_l = \sum_{i=1}^{p_l} \beta_{l_i} \boldsymbol{\gamma}_{l_i} \boldsymbol{g}_{l_i} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \mathrm{N}(0, \boldsymbol{I}).$$

In our scheme, the genetic association by a causal QTN explains 0.7% (for QTN with minor allele frequency 1%) to 15% (for QTN with minor allele frequency 50%) of the heritability, which is quite realistic.

In the analysis, we compute the single SNP Bayes factor using the analytic formula by Wakefield (2009). More specifically, we assume the prior genetic effect of an QTN is drawn from a mixture of Normal distribution, i.e.,

$$\beta_{l_i} \mid \boldsymbol{\gamma}_{l_i} = 1 \sim \sum_i \pi_i \mathrm{N}(0, \phi_i^2).$$

And we use a grid of ϕ values, $\{\phi : 0.1, 0.2, 0.4, 0.8\}$, and set $\pi_1 = \cdots = \pi_4 = 0.25$. We also apply this setting for analyzing the eQTL data from the GTEx project.

F Binning of *cis*-SNPs by DTSS

Here we describe the binning scheme to annotate the SNPs in the *cis* region of a target gene. We place each SNP into 21 unequally spaced bins according to its DTSS (Table 1). The the bins are smaller and denser as close to TSS, which helps capture the rapid decay of QTL signals. This binning scheme leads to estimation of 21 enrichment parameters in the enrichment analysis.

Table 1: Binning scheme for *cis*-SNPs

Bin	Range (kb)	Size (kb)
-10	< -500	500
-9	$[-500, -250)$	250
-8	$[-250, -100)$	150
-7	$[-100, -50)$	50
-6	$[-50, -25)$	25
-5	$[-25, -10)$	15
-4	$[-10, -5)$	5
-3	$[-5, -2.5)$	2.5
-2	$[-2.5, -1)$	1.5
-1	$[-1, -0.5)$	0.5
0	$[-0.5, 0.5)$	1
1	$[0.5, 1)$	0.5
2	$[1, 2.5)$	1.5
3	$[2.5, 5)$	2.5
4	$[5, 10)$	5
5	$[10, 25)$	15
6	$[25, 50)$	25
7	$[50, 100)$	50
8	$[100, 250)$	150
9	$[250, 500)$	250
10	> 500	500

References

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**(2), 283–285.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, pages 1780–1815.

Müller, P., Parmigiani, G., and Rice, K. (2006). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8*, volume 0, pages 349–370. Oxford University Press.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**(2), 155–76.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLOS Genetics*, **3**(7), e114.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic epidemiology*, **33**(1), 79–86.

Wen, X. (2014). Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics*, **70**(1), 73–83.

Wen, X. and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, **4**(3), 1158.