Efficient Integrative Multi-SNP QTL Mapping using Deterministic Approximation of Posteriors

Xiaoquan Wen*[1] and Yeji Lee[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, USA

Abstract

With the increasing availability of functional genomic data (Consortium *et al.*, 2012, Kundaje *et al.*, 2015, Ardlie *et al.*, 2015), incorporating genomic annotations into QTL mapping has become a standard analytical procedure. However, the existing analysis methods often lack rigor and/or computational efficiency. We present a novel algorithm to perform integrative multi-SNP QTL mapping in a probabilistic hierarchical model framework that enables accurate and efficient joint enrichment analysis and the identification of multiple causal variants.

Integrating genomic annotations into QTL mapping provides at least two unique advantages: it improves the power of QTL discoveries by prioritizing functional variants, and it helps link the observed association signals with the underlying molecular mechanisms. A Bayesian hierarchical model that accounts for multiple QTL associations has recently been proposed for this purpose (Wen *et al.*, 2015). A distinctive feature of the model is its usage of prior specification to quantitatively connect the association status of a candidate SNP with its genomic annotations through a set of regression coefficients known as *enrichment parameters*. The model has been successfully applied in mapping expression quantitative trait loci (eQTLs) across multiple tissues (Ardlie *et al.*, 2015) and across multiple populations (Wen *et al.*, 2015), and it shows great advantages over the existing standard single-SNP analysis approaches.

*xwen@umich.edu

To fit the Bayesian hierarchical model, an Expectation-Maximization (EM) algorithm (Wen *et al.*, 2015) is first applied to find the maximum likelihood estimates (MLEs) of the enrichment parameters. Given the MLEs, an empirical Bayes approach is then utilized to perform the multi-SNP mapping of the QTLs. Particularly in the E-step of the EM algorithm, the marginal posterior probability of association (which we will refer to as the Posterior Inclusion Probability, or PIP, henceforth) for each candidate SNP is computed given the current estimates of the enrichment parameters. This step remains a major computational bottleneck because of its requirement of extensive explorations of multi-SNP combinations for the QTL association model (for p candidate SNPs, the model space contains 2^p different combinations). Currently, the only feasible approach for the E-step is via a Markov Chain Monte Carlo (MCMC) algorithm (Wen *et al.*, 2015). However, the repeated execution of the MCMC algorithm in every single E-step significantly slows the model fitting, and the EM algorithm becomes computationally impractical for accommodating QTL data at the genome-wide scale. Furthermore, the inherent stochastic variation in the MCMC algorithm may affect the performance and reproducibility of the EM algorithm.

Here, we present an alternative algorithm to perform deterministic approximation of posteriors (DAP) and efficiently compute PIPs for all candidate SNPs given the point estimates of the enrichment parameters. This algorithm is mainly based on two observations. First, in almost all genetic applications, the convincing QTLs discovered from the association data are highly *sparse* compared with the number of candidate SNPs. This sparseness implies that the vast majority of the posterior probability mass in the space of all possible combinations of SNPs must be concentrated in a much lower-dimensional subspace. That is, only association models containing few SNPs are likely to have non-negligible posterior probabilities. Second, noteworthy QTL SNPs, as reflected by their non-negligible PIP values, are thought to typically show modest to strong marginal association signals in either single-SNP or conditional analysis. Based on the above observations, we designed the DAP algorithm to adaptively select a small subset of noteworthy candidate QTL SNPs and thoroughly explore the low-dimensional model space composed by these SNPs. In addition, the DAP algorithm applies a combinatorial approximation to estimate the posterior probability mass from the unexplored model space. Unlike the MCMC, the DAP algorithm is highly parallelizable, and our implementation takes full advantage of this property.

The DAP algorithm is directly applicable to QTL mapping in targeted genomic regions of limited length, e.g., the mapping of *cis*-eQTLs. For genome-wide QTL mapping, we further approximate the posterior probability of each complete association model by a product of “regional” posterior probabilities, with each region corresponding to a relatively independent LD block (Berisa and Pickrell, 2015). The DAP can then be applied to each LD block independently. This approach is similar to the strategies adopted by some commonly applied procedures (Pickrell, 2014), and we provide a rigorous mathematical argument to justify the factorization in Appendix B.

Instead of adaptively selecting a subset of high-priority SNPs, the DAP algorithm can also be applied by pre-fixing the maximum model size (namely, K) while allowing the exploration of all possible SNP combinations under the restriction. We refer to this variant of the algorithm as the DAP-K algorithm. In the special case of $K = 1$ (DAP-1), the algorithm essentially assumes that at most one causal QTL exists in the region of interest. Although this very assumption has been successfully applied by many other approaches (Pickrell, 2014, Servin and Stephens, 2007, Veyrieras *et al.*, 2008, Flutre *et al.*, 2013), it has always been formulated as an explicit prior assumption and hence requires a somewhat non-natural parameterization that also complicates the maximization step when used in the EM algorithm for enrichment analysis (Appendix D). The DAP-1 algorithm has the advantage of considerably faster computation, even compared to the adaptive version of the DAP algorithm. More importantly, it can be applied using only summary statistics from single-SNP association analysis (in the form of the marginal estimate of the genetic effect and its standard error for each SNP). This feature is particularly attractive, especially when the individual-level genotype and phenotype information is difficult to access.

We perform numerical experiments to investigate the accuracy of the adaptive DAP procedure. Specifically, we select the genotype data from a random region of 15 SNPs from the GEUVADIS project (Lappalainen *et al.*, 2013) across five population groups. We randomly assign 1 to 5 causal QTLs and simulate a quantitative phenotype based on a linear model. Using the pre-fixed hyper-parameters, we perform the *exact* Bayesian computation by enumerating all 2^{15} different association models and compare the results to the outcome of the DAP algorithm. Our simulation results indicate that the DAP algorithm yields a highly accurate approximation at only a small fraction of the computational cost (Supplementary Figs. A1 and A2 and Supplementary Tables A1 and A2).

Next, we compare the performance of the DAP and MCMC algorithms in fine-mapping multiple QTLs using a simulated multiple population eQTL data set (Wen *et al.*, 2015). Specifically, we assess their abilities to correctly identify multiple genomic regions that harbor causal eQTLs. Our results (Fig. 1, Supplementary Fig. A3, Supplementary Table A3) show that the DAP algorithm presents a significant improvement in performance compared with the MCMC algorithm with a remarkable reduction in computational time (Appendix F.3) and that both methods outperform the traditional QTL mapping approaches. In addition, Figure 1 also indicates that with prolonged sampling steps, the MCMC outputs seemingly "converges" to the DAP results.

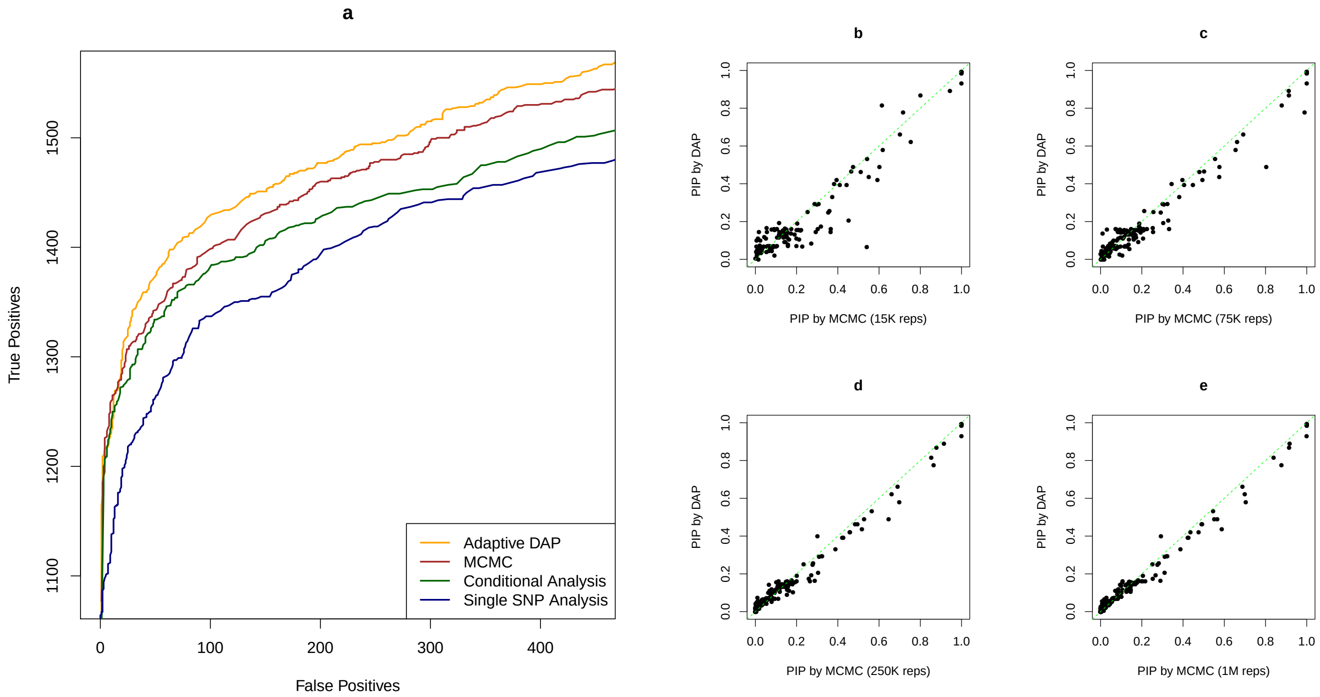

Figure 1: **(a)** Performance comparisons for multi-SNP QTL mapping. We apply different analytical approaches to a simulated data set reported in Wen *et al.* (2015) to evaluate their abilities to identify multiple independent LD blocks harboring true QTLs. The methods compared include a single-SNP analysis approach (navy blue line), a forward selection-based conditional analysis approach, the MCMC algorithm described in Wen *et al.* (2015), and the DAP algorithm. Each plotted point represents the number of true positive findings (of LD blocks) versus the false positives obtained by a given method at a specific threshold. The MCMC algorithm and the DAP algorithm are based on the Bayesian hierarchical model and clearly outperform the other two commonly applied approaches. Most importantly, the DAP algorithm presents a significant performance improvement compared with the MCMC in both accuracy and computational efficiency. **(c) - (e)** Comparison of PIP values estimated by adaptive DAP and MCMC with various running lengths. We randomly select 10 simulated data sets and run MCMC with 4 different lengths of sampling steps, ranging from 15,000 to 1 million (the results shown in panel (**a**) are based on 75,000 sampling steps for each data set). With the prolonged MCMC runs, the MCMC outcomes seemingly "converge" to the DAP results.

The integration of DAP into the EM algorithm enables the integrative analysis of large-scale QTL data sets. To investigate its performance, we simulate a modest-scale eQTL data set to mimic the genome-wide investigation of *cis*-eQTLs. We apply the proposed EM algorithm to estimate the enrichment parameter and assess the precision of the estimate. Overall, we conclude that both adaptive DAP and the DAP-1-embedded EM algorithm yield accurate estimates in our simulation setting. Although the adaptive DAP algorithm generally yields more accurate point estimates and narrower confidence intervals (Supplementary Figs. A4 and A5), the DAP-1 algorithm achieves significant savings in computational time (Appendix F.3). In comparison, the commonly applied naive enrichment analysis method severely underestimates the enrichment parameter.

Finally, we apply the new integrative QTL mapping approach to re-analyze the cross-population eQTL data set generated from the GEUVADIS project (Lappalainen *et al.*, 2013). By applying the DAP-1-embedded EM algorithm, we simultaneously examine two types of genomic annotations that are known to impact the enrichment of eQTLs: the SNP distance to the transcription start site (TSS) of the target gene and the annotations assessing the SNP's ability to disrupt transcription factor (TF) bindings. Our results (Fig. 2, Supplementary Figs. A6) indicate that SNPs that are computationally predicted to strongly disrupt TF binding are more likely to alter gene expression as eQTLs (fold change of 2.57 with a 95% CI of $[2.314, 2.855]$) compared with the SNPs that are simply located within a DNase I footprint region (fold change of 1.707 with a 95% CI of $[1.486, 1.962]$). Using the enrichment parameter estimates, we then fine-map the eQTLs of each gene using the adaptive DAP algorithm. In many cases, we observe that the quantitative annotations allow prioritizing functional SNPs that are otherwise indistinguishable because of LD. We show one such example in Fig. 2.

In summary, the DAP algorithm provides an elegant and computationally efficient tool to perform multi-SNP QTL mapping while incorporating genomic annotations and accounting for LD. It exhibits superior statistical power over the traditional single-SNP analysis-based approaches and is substantially more efficient and accurate than the MCMC-based multi-SNP analysis methods. The algorithm has been implemented in C++ and is freely available at http://github.com/xqwen/dap/.

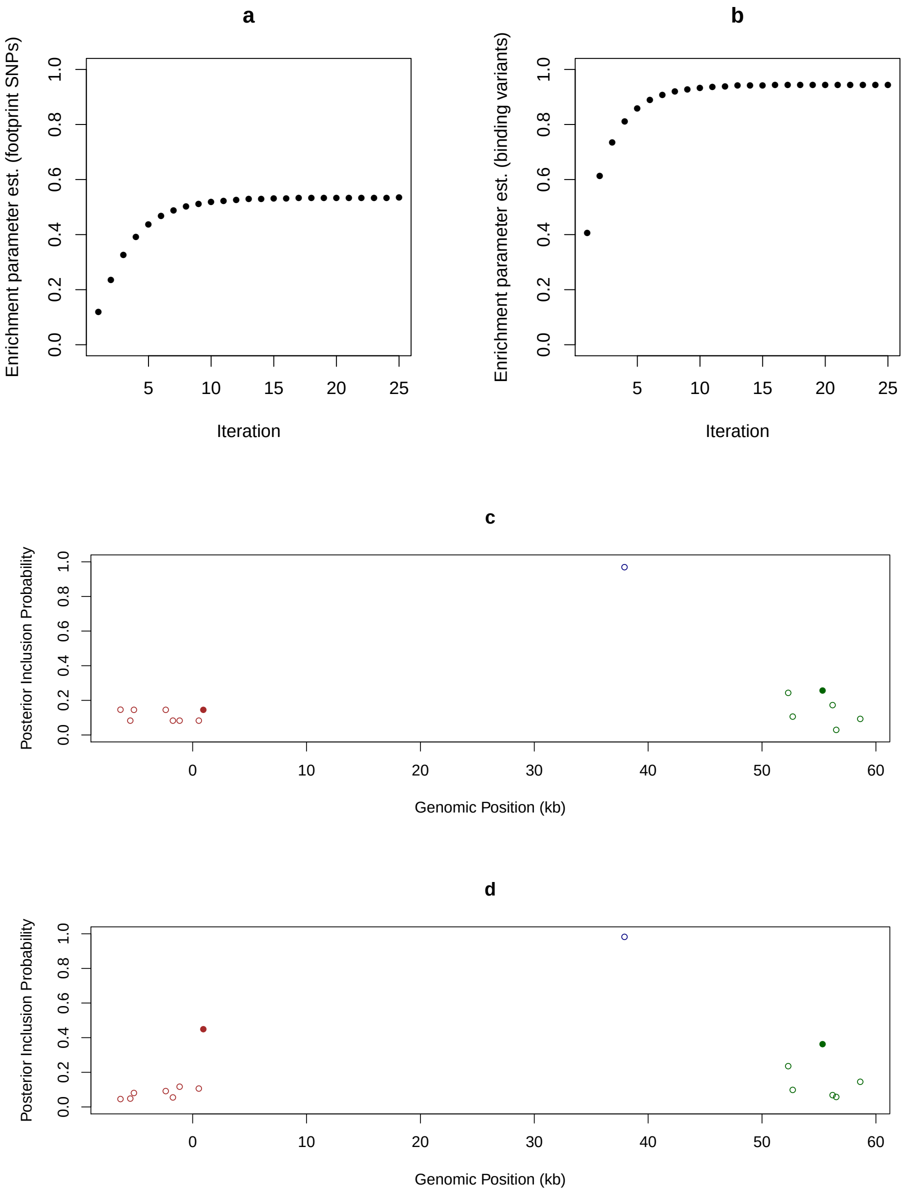

Figure 2: **(a) - (b)** Traceplots of estimates of the enrichment parameters for binding variants and footprint SNPs during the DAP-1-embedded EM iterations in analyzing the GEUVADIS data. Both estimates are stabilized after about 8 iterations. **(c) - (d)** Comparison of multi-SNP *cis*-eQTL mapping with and without incorporating functional annotations. We plot the multi-SNP QTL mapping results of gene *LY86* (Ensembl ID: ENSG00000112799) using the GEUVADIS data. Panel **(c)** shows the results assuming that all SNPs are equally likely to be associated *a priori*, i.e., no functional annotation is used. Panel **(d)** shows the results using the functional annotations with enrichment parameters estimated by the DAP-1-embedded EM algorithm. In both cases, we use the adaptive DAP algorithm to perform the multi-SNP QTL mapping and plot the SNPs with PIP > 0.02 with respect to their positions to the transcription start site. SNPs in high LD are plotted with the same color, and the filled circle indicates that a SNP is annotated as disrupting TF binding. It is clear that three independent *cis*-eQTLs exist because in both panels, the sums of the PIPs from the SNPs with the same color all $\to 1$. When incorporating functional annotation to perform integrative QTL mapping, the binding variants show much greater PIP values and are prioritized over the non-annotated SNPs in high LD.

Methods

Model and Notation. Without loss of generality, we model the associations between p SNPs and a quantitative trait of interest using the following multiple linear regression model

$$\boldsymbol{y} = \mu \mathbf{1} + \sum_{i=1}^{p} \beta_i \boldsymbol{g}_i + \boldsymbol{e},\ \ \boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \tau^{-1}\boldsymbol{I}),$$

where the vectors $\boldsymbol{y}$ and $\boldsymbol{g}_i$ denote the phenotype measurements and genotypes of SNP i from n unrelated individuals, $\boldsymbol{e}$ denotes the residual errors with variance τ^{-1}, and the parameters μ and β_i's represent the intercept and genetic effects of each SNP, respectively. For each SNP, we represent the latent association status by a binary indicator $\gamma_i := \mathbf{1}\{\beta_i \neq 0\}$. In mapping QTLs, we are interested in making joint inferences with respect to the p-vector $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_p)$. Specifically, we define the size of the association model, $||\boldsymbol{\gamma}||$, as the number of associated SNPs, i.e., $||\boldsymbol{\gamma}|| = \sum_{i=1}^{p} \gamma_i$. We assign an independent logistic prior for each SNP i to link its association status with its genomic annotations. Specifically,

$$\log \left[\frac{\Pr(\gamma_i = 1)}{\Pr(\gamma_i = 0)} \right] = \alpha_0 + \sum_{k=1}^{q} \alpha_k d_{ik},$$

where $\boldsymbol{d}_i := (d_{i1}, \ldots, d_{iq})$ denotes q genomic annotations that are specific to SNP i, and $\alpha_1, ..., \alpha_q$ are referred to as the enrichment parameters: the positive α_k value implies that as the annotation value of feature k increases, the odds of the SNP being a causal QTL increase or, equivalently, the feature k is enriched in QTLs. Finally, we denote $\boldsymbol{\alpha} := (\alpha_0, ..., \alpha_q)$ and $\boldsymbol{G} := (\boldsymbol{g}_1, ..., \boldsymbol{g}_p)$. We are interested in making inferences regarding $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

Statistical Inference. A key component of our inference procedure is the previous results (Wen, 2014) of computing the marginal likelihood of the regression model in the form of a Bayes factor, namely, $\mathrm{BF}(\boldsymbol{\gamma}) = \frac{P(\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{G})}{P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\gamma} \equiv \mathbf{0})}$. In computing this quantity, the nuisance parameters μ, τ and β_i's are all integrated out. Given the hyper-parameter $\boldsymbol{\alpha}$, it follows from the Bayes rule that

$$\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha})\, \mathrm{BF}(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'} \Pr(\boldsymbol{\gamma}' \mid \boldsymbol{\alpha})\, \mathrm{BF}(\boldsymbol{\gamma}')}. \tag{1}$$

Based on these results, an EM algorithm (Wen *et al.*, 2015) can be derived to find the MLE of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$, by treating $\boldsymbol{\gamma}$ as missing data. We provide the details of the EM algorithm in Appendix A. Briefly, in the E-step, given the current estimate of $\boldsymbol{\alpha}$, we compute the PIP for each candidate SNP by marginalizing the probabilities obtained from (1); in the M-step, we fit a logistic regression model using the current PIP value of each SNP as the response to update the estimate of $\hat{\boldsymbol{\alpha}}$. Upon convergence of the EM algorithm, we plug in $\hat{\boldsymbol{\alpha}}$ and apply (1) again to obtain the final inference result for each candidate SNP k, i.e., the PIP $\Pr(\gamma_k \mid \boldsymbol{y}, \boldsymbol{G}, \hat{\boldsymbol{\alpha}})$.

Deterministic Approximations of Posteriors. Evaluating equation (1) is critical for the EM algorithm and the final inference of PIPs. Unfortunately, exact computation is practically impossible for large numbers of SNPs because the summation in the denominator requires exploring an enormous model space. Previously, the MCMC algorithm has been proposed to circumvent this difficulty, but the computational cost is still too high to run the EM algorithm for large-scale data sets. The proposed DAP algorithm approximates the normalizing constant $C = \sum_{\boldsymbol{\gamma}'} \Pr(\boldsymbol{\gamma}' \mid \boldsymbol{\alpha}) \, \mathrm{BF}(\boldsymbol{\gamma}')$ by

$$C^* = \sum_{\boldsymbol{\gamma}' \in \Omega} \Pr(\boldsymbol{\gamma}' \mid \boldsymbol{\alpha}) \, \mathrm{BF}(\boldsymbol{\gamma}') + \epsilon, \tag{2}$$

where Ω denotes a subset of the most plausible models, and ϵ is an estimate of the approximation error $C - C^*$.

The adaptive version of DAP applies two levels of approximations. First, for larger size partitions of the model space $\{\boldsymbol{\gamma}\}$, it approximates $C_s = \sum_{||\boldsymbol{\gamma}||=s} \Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) \, \mathrm{BF}(\boldsymbol{\gamma})$ by $C_s^* = \sum_{\boldsymbol{\gamma} \in \Omega, \, ||\boldsymbol{\gamma}||=s} \Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) \, \mathrm{BF}(\boldsymbol{\gamma})$, whose computation may only involve a small subset of SNPs. (Note that for the null model ($s = 0$) and single QTL models ($s = 1$), the DAP algorithm performs exact calculations, i.e., $C_0^* = C_0$ and $C_1^* = C_1$.) For $s \geq 2$, the DAP algorithm only focuses on a subset of adaptively selected high-priority SNPs and enumerates all possible combinations within the subset. The adaptive selection of the high-priority SNPs is similar to a Bayesian version of conditional analysis (Flutre *et al.*, 2013) that naturally accounts for LD. More specifically, suppose that a "best" model with the maximum posterior probability for $||\boldsymbol{\gamma}|| = s - 1$ has been identified. The SNP selection procedure then goes through all candidate SNPs, adding a single SNP at a time to the existing best model, and evaluates their posterior probabilities of being the sole additional QTL signal (see details in Appendix C.1). Note that this procedure is similar to single-SNP analysis and is computationally trivial. The candidate SNPs whose posterior probabilities in the conditional analysis are greater than a pre-defined threshold λ (by default, $\lambda = 0.01$) are then added to the existing subset of high-priority SNPs. Finally, the DAP algorithm enumerates the updated subset of priority SNPs for all combinations of $||\boldsymbol{\gamma}|| = s$ to compute C_s^* and, in the process, records the "best" posterior model with the increased model size. At the second level of approximation, the DAP algorithm only extensively explores the relatively small model spaces. Suppose that there are truly K QTLs in p candidate SNPs. It should be clear that $\{C_s\}$ becomes a (sharply) decreasing sequence as $s > K$ and that the behavior of this decreasing sequence is mathematically predictable (Supplementary Figure A2). This behavior occurs because the marginal likelihood becomes saturated as the model size exceeds the number of true associations and because the additional prior term imposes a hefty penalty on the overall product. Utilizing this fact, we derive an approximate recursive relationship between C_s and C_{s+1} as $s \geq K$ (Appendix C.2). Based on this relationship, the stopping rule of explicit exploration is determined, and we estimate ϵ by

$$\epsilon = \sum_{s=t+1}^{p} R_s^* \quad \text{with} \quad R_{s+1}^* = \frac{p-s}{s+1}\,\omega\, R_s^* \ \text{ for } \ s = t+1, ..., p, \tag{3}$$

where t is the stopping point of extensive exploration, $\omega = \frac{1}{p}\sum_{i=1}^{p} \exp\left(\alpha_0 + \sum_{l=1}^{q} \alpha_l d_{il}\right)$ represents the average prior odds ratio across SNPs and $R_t^* = C_t^*$. This estimation essentially assumes that the marginal likelihood is completely saturated for the partitions with $s > t$, and the overall contribution to the normalizing constant from each size partition can be roughly estimated by re-calibrating the prior changes (see details in Appendix C.2).

Jointly applying the two approximation strategies significantly reduces the computational burden and yields a highly accurate approximation of the normalizing constant. It should also be noted that if the tuning parameter λ is set to 0, all SNPs are included in the analysis and the approximation from the adaptive DAP becomes almost exact; in contrast, when λ is set to relatively large values, the DAP behaves similarly to conditional analysis, and only a very few very high-probability models are explored.

The version of the DAP algorithm that pre-fixes the maximum model size at K represents a much simpler approximation procedure, i.e.,

$$C = \sum_{||\boldsymbol{\gamma}|| \leq K} \Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha})\, \mathrm{BF}(\boldsymbol{\gamma}) + \epsilon,$$

where ϵ is assumed to be negligible. For $K = 1$, the PIP of SNP i can be analytically computed by

$$\Pr(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) = \frac{\sum_{k=1}^{p} e^{\alpha_0 + \sum_{l=1}^{q} \alpha_l d_{kl}}\, \mathrm{BF}_k}{1 + \sum_{k=1}^{p} e^{\alpha_0 + \sum_{l=1}^{q} \alpha_l d_{kl}}\, \mathrm{BF}_k} \cdot \frac{e^{\sum_{l=1}^{q} \alpha_l d_{il}}\, \mathrm{BF}_i}{\sum_{k=1}^{p} e^{\sum_{l=1}^{q} \alpha_l d_{kl}}\, \mathrm{BF}_k}, \tag{4}$$

where BF_i represents the single SNP Bayes factor for SNP i and can be analytically computed using only summary-level statistics from single-SNP analysis (see details in Appendix D).

Factorization of Posteriors. In practice, we find that DAP works extremely well if the candidate SNP in $\boldsymbol{\gamma}$ spans a genomic region up to 2 Mb. For genome-wide applications, we apply an additional factorization to approximate $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})$ as

$$\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) \approx \prod_{k=1}^{L} \Pr(\boldsymbol{\gamma}_{[k]} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}), \tag{5}$$

where $\{\boldsymbol{\gamma}_{[k]} : k = 1, ..., L\}$ represents a partition of $\boldsymbol{\gamma}$ according to the LD patterns (or the relevant recombination map). This factorization is justified by the previous theoretical results (Wen and Stephens, 2010), and we provide complete details in Appendix B. Briefly, it can be shown that $\text{BF}(\boldsymbol{\gamma}) \approx \prod_{k=1}^{L} \text{BF}(\boldsymbol{\gamma}_{[k]})$. This result, along with the fact that our priors are independent across SNPs, naturally leads to the approximate factorization of the posterior probability.

Details of Simulation and Data Analysis. We provide technical details, including descriptions of the simulation schemes and analytical methods, running time benchmarks and additional analytical results, in the appendices.

Appendix A EM Algorithm for Estimating Enrichment Parameters

The EM algorithm for fitting the hierarchical model is a special case of what is described in Wen *et al.* (2015). Specifically, by treating the vector of joint association status, $\boldsymbol{\gamma}$, as the missing data, the complete data likelihood can be written as

$$P(\boldsymbol{y}, \boldsymbol{\gamma} \mid \boldsymbol{G}, \boldsymbol{\alpha}) = \Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha})\, P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\gamma}),$$

where

$$\Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{p} \left[\left(\frac{\exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)} \right)^{\gamma_i} \left(\frac{1}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)} \right)^{1-\gamma_i} \right].$$

Therefore, the complete data log-likelihood is given by

$$\log \Pr(\boldsymbol{y}, \boldsymbol{\gamma} \mid \boldsymbol{G}, \boldsymbol{\alpha}) = \sum_{i=1}^{p} \gamma_i (\boldsymbol{\alpha}'\boldsymbol{d}_i) - \sum_{i=1}^{p} \log[1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)] + \log P(\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{G}).$$

It is important to note that only the first two terms on the RHS contain $\boldsymbol{\alpha}$, the parameter of interest.

The EM algorithm is initiated by assigning $\boldsymbol{\alpha}$ to an arbitrary starting value, namely $\boldsymbol{\alpha}^{(1)}$. In the E-step of the t-th iteration, we compute

$$\begin{aligned}
\mathrm{E}\left[\log \Pr(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) \mid \boldsymbol{y}, \boldsymbol{g}, \boldsymbol{\alpha}^{(t)}\right] = & \sum_{i=1}^{p} \Pr(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{g}, \boldsymbol{\alpha}^{(t)})(\boldsymbol{\alpha}'\boldsymbol{d}_i) \\
& - \sum_{i=1}^{p} \log[1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)] \\
& + \mathrm{E}\left(\log P(\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{G}) \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}^{(t)}\right).
\end{aligned}$$

In the M-step, we find

$$\boldsymbol{\alpha}^{(t+1)} = \arg\max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{p} \Pr(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{g}, \boldsymbol{\alpha}^{(t)})(\boldsymbol{\alpha}'\boldsymbol{d}_i) - \sum_{i=1}^{p} \log[1+\exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)] \right).$$

Note that

1. the only quantities that are required from the E-step are the PIPs of all candidate SNPs.

2. the functional form of the objective function is the same as the log likelihood function of a logistic regression model with the binary response variable replaced by the corresponding PIP, $\Pr(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{g}, \boldsymbol{\alpha}^{(t)})$. Therefore, the algorithm for finding the MLEs of a logistic regression model can be directly applied in the maximization step.

Appendix B Factorization of $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})$ by LD blocks

For genome-wide QTL mapping applications, we recommend the factorization, $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) \approx \sum_{k=1}^{L} \Pr(\boldsymbol{\gamma}_{[k]} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})$, before applying the DAP in each genomic region independently. In this section, we provide the necessary mathematical justification for the proposed factorization.

It is sufficient to show that

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

Recall that $\{\boldsymbol{\gamma}_{[k]} : k = 1, 2, 3...\}$ are non-overlapping segments 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 relationship on 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; whereas if $\gamma_i = 0$, $\beta_i = 0$ with probability 1 (or 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, at the moment, we also assume the residual error variance parameter τ is known. It then follows from the result of Wen (2014) that

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

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 an 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 (2015) provide an algorithm to segment the genome into L non-overlapping loci utilizing the population parameter of recombination rate, i.e.,

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

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

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

where "$\oplus$" denotes the direct sum of the matrices. It is important to note that (B.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 (B.2) into (B.1), it follows that

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

where

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

In particular, $(\boldsymbol{W}_{[1]}, \ldots, \boldsymbol{W}_{[[L]})$ is a decomposition of the diagonal matrix $\boldsymbol{W}$ compatible to 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 an point estimate of τ (e.g., based on $\boldsymbol{y}$ and $\boldsymbol{G}_{[k]}$ for each block k) into the expression (B.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,

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

Appendix C Details of Adaptive DAP Algorithm

C.1 Adaptive Selection of Priority SNPs

Here we give a detailed account on the Bayesian conditional analysis procedure for selecting high priority SNPs in the DAP algorithm. The procedure starts with model size partition $s = 1$. Let $\boldsymbol{\gamma}^*$ denote the model with the highest posterior probability in the size partition $s - 1$, i.e.,

$$\boldsymbol{\gamma}^* = \mathrm{argmax}_{\{||\boldsymbol{\gamma}|| = s-1\}} \Pr(\boldsymbol{\gamma})\mathrm{BF}(\boldsymbol{\gamma}).$$

For each SNP i that is not included in the current best model, we compute a Bayes factor for the expanded model , $\boldsymbol{\gamma}_i^\dagger = \boldsymbol{\gamma}^* \cup \{\gamma_i = 1\}$. Assuming there is exactly one additional QTL and each candidate SNP i is equally likely to be *the* additional causal association *a priori*, the corresponding conditional posterior probability for SNP i can be computed by

$$\mathrm{PIP}_i^* = \frac{\mathrm{BF}(\boldsymbol{\gamma}_i^\dagger)/\mathrm{BF}(\boldsymbol{\gamma}^*)}{\sum_j \mathrm{BF}(\boldsymbol{\gamma}_j^\dagger)/\mathrm{BF}(\boldsymbol{\gamma}^*)} = \frac{\mathrm{BF}(\boldsymbol{\gamma}_i^\dagger)}{\sum_j \mathrm{BF}(\boldsymbol{\gamma}_j^\dagger)}. \tag{C.1}$$

The resulting quantity is a well defined posterior probability and solely determined by the relative likelihood values of the expanded models. Particular, it should be noted that (C.1) fully accounts for LD between SNPs: e.g., if two SNPs are in perfect LD, they would possess identical values which correctly reflect the uncertainty (i.e., they are indistinguishable). The procedure requires $p - s$ evaluations of Bayes factors which are computationally trivial for small s values. Given the pre-defined threshold λ, we add the SNP i into the existing set of high priority SNPs if it is not already in the set and $\text{PIP}_i^* \ge \lambda$. For $s \ge 2$, we then enumerate all s-combinations from the resulting set of priority SNPs to compute C_s^*. Also during this enumerating process, we record the new $\boldsymbol{\gamma}^*$ for the increased model size.

Intuitively, the threshold parameter λ relates to the precision of the approximate PIPs. In a way, the selection procedure roughly estimates the probability, $\Pr(\boldsymbol{\gamma}_i = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}, ||\boldsymbol{\gamma}|| = s)$, for SNP i. Note the relationship

$$\Pr(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) = \sum_{s=1}^{p} \frac{C_i}{C} \cdot \Pr(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}, ||\boldsymbol{\gamma}|| = s).$$

It can be concluded that

1. If $\Pr(\boldsymbol{\gamma}_i = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}, ||\boldsymbol{\gamma}|| = s) < \lambda$ for a given SNP at all s values, it must be the case that the overall PIP $< \lambda$.
2. The loss of precision for the PIP of SNP i due to the selection screening in a particular size partition must be $< \lambda$.

In our simulation studies, we observe that λ represents an upper bound for the average precision (measured by RMSE) of the approximate PIPs (see Tab. A1).

C.2 Stopping Rule for Explicit Exploration and Estimation of Approximation Error

When a non-associated SNP is added into an existing association model, the marginal likelihood of the model is typically non-increasing. In fact, the marginal likelihood measured by the corresponding Bayes factor usually decreases slightly due to the effect of Occam's razor built into the Bayes factor computation (Berger and Pericchi, 1996). We utilize this property to reduce the computation of DAP by eliminating the unnecessary explicit explorations of the model partitions once the sizes of the models considered exceed the number of the detectable association signals. To achieve this goal, the DAP starts the exploration with model size partition $s = 1$ for increasing s values until a stopping rule is met. The contribution from the unexplored size partitions (i.e., the approximation error) is estimated by an analytic combinatorial approximation.

To explain the stopping rule and the combinatorial approximation, we assume that there are K detectable true QTLs. In each model size partition where $s > K$, we can classify all models into $(K+1)$ mutually exclusive categories according to the number of true QTLs (0 to K) included in each association model. In the category including exactly m true QTLs, each member association model also includes $(s-m)$ non-associated SNPs, and the total number of the association models in the category is given by $\binom{p-K}{s-m}\binom{K}{m}$. We estimate the contribution to $\sum \Pr(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s)\mathrm{BF}(\boldsymbol{\gamma})$ from this particular category by the equation

$$\binom{p-K}{s-m}\binom{K}{m}\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s)\,\overline{\mathrm{BF}}_{\{m\}},$$

where $\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s)$ represents the average prior value within the category, and $\overline{\mathrm{BF}}_{\{m\}}$ is the average Bayes factor across models including m out of K detectable QTLs. The use of $\overline{\mathrm{BF}}_{\{m\}}$ is mainly based on the assumption that including non-associated SNPs in an association model does not, on average, increases the marginal likelihood/Bayes factor. Hence, we obtain

$$C_s \approx \sum_{m=0}^{K}\binom{p-K}{s-m}\binom{K}{m}\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s)\,\overline{\mathrm{BF}}_{\{m\}}$$

To relate C_{s+1} to C_s, we note that

$$\begin{aligned} C_{s+1} &\approx \sum_{m=0}^{K}\binom{p-K}{s+1-m}\binom{K}{m}\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s+1)\,\overline{\mathrm{BF}}_{\{m\}} \\ &= \sum_{m=0}^{K}\frac{p-K+m-s}{s+1-m}\binom{p-K}{s-m}\binom{K}{m}\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s+1)\,\overline{\mathrm{BF}}_{\{m\}} \\ &\le \frac{p-s}{s+1-K}\sum_{m=0}^{K}\left[\binom{p-K}{s-m}\binom{K}{m}\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s)\,\overline{\mathrm{BF}}_{\{m\}}\right]\frac{\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s+1)}{\widetilde{\Pr}(\boldsymbol{\gamma}; ||\boldsymbol{\gamma}|| = s)} \\ &\approx \frac{p-s}{s-K+1}\,\omega\,C_s \end{aligned} \tag{C.2}$$

In the last step, we approximate the quantities $\frac{\widetilde{\Pr}(\boldsymbol{\gamma};||\boldsymbol{\gamma}||=s+1)}{\widetilde{\Pr}(\boldsymbol{\gamma};||\boldsymbol{\gamma}||=s)}$ in all $K+1$ categories by the average prior odds $\omega = \frac{1}{p}\sum_{i=1}^{p}\exp\left(\alpha_0 + \sum_{l=1}^{q}\alpha_l d_{il}\right)$. Similarly, we can derive an approximate lower bound for C_{s+1} as

$$\frac{p-s-K}{s+1}\,\omega\,C_s. \tag{C.3}$$

Thus, we have shown

$$\frac{p-s}{s-K+1}\,\omega\,C_s \gtrsim C_{s+1} \gtrsim \frac{p-s-K}{s+1}\,\omega\,C_s. \tag{C.4}$$

Because K is unknown, we estimate C_{s+1} from C_s by the following approximation

$$C_{s+1} \approx \frac{p-s}{s+1}\,\omega\,C_s, \tag{C.5}$$

which does not depend on K and lies in the interval $\left(\frac{p-s-K}{s+1}\,\omega\,C_s, \frac{p-s}{s-K+1}\,\omega\,C_s\right)$. Our numerical experiment shows that the this approximation is surprisingly accurate (Fig. A2).

Our stopping rule is built upon the upper-bound specified by the inequality (C.4). Specially, the adaptive DAP stops explicit exploration at partition size $s = t$ if

$$C_t^* \le (p-t+1)\,\omega\,C_{t-1}^*. \tag{C.6}$$

The inequality essentially tests $K \ge t-1$. In addition to utilizing the combinatorial approximation, the DAP further monitors the increment of the partial sum $S_k = \sum_i^k C_i^*$. To ensure the high accuracy of the approximation, we also add an optional criteria into the stopping rule on top of (C.6), i.e.,

$$\log_{10}\left[\frac{S_t}{S_{t-1}}\right] < \kappa,\ \ \kappa > 0,$$

or equivalently,

$$\frac{C_t^*}{\sum_i^{t-1} C_i^*} < 10^\kappa - 1,$$

By default, we set $\kappa = 0.01$, which further ensures that the subsequent model size partitions have no substantial contributions to the normalizing constant. This additional criteria provides practical flexibility in running the DAP: as $\kappa \to 0$, it enforces the DAP explore all the model size partitions; whereas when κ is large, only the stopping rule (C.6) is effective.

Once the stopping rule is invoked, we estimate ϵ by

$$\epsilon = \sum_{s=t+1}^{p} R_s^*,$$

where we define $R_t^* = C_t^*$ and

$$R_{s+1}^* = \frac{p-s}{s+1}\,\omega\, R_s^*, \quad \text{for} \quad s = t, ..., p.$$

Appendix D Derivation of DAP-1 Algorithm

In this section, we give a detailed derivation for DAP-1 algorithm. It should be noted that the derivation can be generalized to DAP-K algorithm with $K > 1$.

The key assumption of the DAP-1 is that posterior probabilities of single QTL association models dominate the posterior probability space of $\{\boldsymbol{\gamma}\}$, i.e.,

$$C - \sum_{||\boldsymbol{\gamma}|| \le 1} \Pr(\boldsymbol{\gamma})\text{BF}(\boldsymbol{\gamma}) \to 0. \tag{D.1}$$

Consequently, it follows that

$$\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) \approx \begin{cases} \frac{\Pr(\boldsymbol{\gamma}|\boldsymbol{\alpha})\text{BF}(\boldsymbol{\gamma})}{\sum_{||\boldsymbol{\gamma}||\le 1}\Pr(\boldsymbol{\gamma})\text{BF}(\boldsymbol{\gamma})} & \text{if } ||\boldsymbol{\gamma}|| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

The model space of $\{\boldsymbol{\gamma} : ||\boldsymbol{\gamma}|| \le 1\}$ contains only the null model, $\boldsymbol{\gamma} = \mathbf{0}$, and all single-SNP association models. For the null model, it is clear that $\text{BF}(\boldsymbol{\gamma} = \mathbf{0}) = 1$, and we denote

$$\pi_0 := \Pr(\boldsymbol{\gamma} = 0 \mid \boldsymbol{\alpha}) = \prod_{i=1}^{p} \left(1 + \exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)\right)^{-1}.$$

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

$$\Pr(\boldsymbol{\gamma}_j^\circ \mid \boldsymbol{\alpha}) = \exp(\boldsymbol{\alpha}'\boldsymbol{d}_j) \prod_{i=1}^{p} \left(1 + \exp(\boldsymbol{\alpha}'\boldsymbol{d}_i)\right)^{-1} = \pi_0 \cdot \exp(\boldsymbol{\alpha}'\boldsymbol{d}_j),$$

and

$$\text{BF}(\boldsymbol{\gamma}_j^\circ) = \text{BF}_j,$$

and recall that BF_j denotes the Bayes factor based on the single-SNP analysis of SNP j. The computation of BF_j has been detailed by many authors (Servin and Stephens, 2007, Wakefield, 2009, Wen *et al.*, 2014). It typically requires only summary level statistics, e.g., estimated genetic effect of the target SNP and its standard error (Wakefield, 2009, Wen *et al.*, 2014), and is computationally trivial. Finally, we note that given the restrained model space, the PIP of SNP j, $\Pr(\gamma_j \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})$, coincides with $\Pr(\boldsymbol{\gamma}_j^\circ \mid \boldsymbol{\alpha})$. Given all of the above, it follows from the simple algebra that

$$\begin{aligned}\Pr(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) &= \frac{\sum_{k=1}^p e^{\alpha_0 + \sum_{l=1}^q \alpha_l d_{kl}} \,\mathrm{BF}_k}{1 + \sum_{k=1}^p e^{\alpha_0 + \sum_{l=1}^q \alpha_l d_{kl}} \,\mathrm{BF}_k} \cdot \frac{e^{\sum_{l=1}^q \alpha_l d_{il}} \,\mathrm{BF}_i}{\sum_{k=1}^p e^{\sum_{l=1}^q \alpha_l d_{kl}} \,\mathrm{BF}_k} \\ &= \big[1 - \Pr(\boldsymbol{\gamma} = \boldsymbol{0} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})\big] \cdot \frac{e^{\sum_{l=1}^q \alpha_l d_{il}} \,\mathrm{BF}_i}{\sum_{k=1}^p e^{\sum_{l=1}^q \alpha_l d_{kl}} \,\mathrm{BF}_k},\end{aligned} \tag{D.2}$$

where the first term assess the probability that the p-SNP locus contains a QTL and the second term is the conditional probability that the i-th SNP is the sole QTL. The expression (D.2) bears the great similarity to the Bayesian approaches used in Veyrieras *et al.* (2008), Flutre *et al.* (2013), Pickrell (2014), which also impose the "single QTL per locus" assumption. However, all the aforementioned approaches formulate it as a prior assumption which results in a very different parametrization. More specifically, they use a locus-level quantity, π_0, to denote the probability that a locus contains no QTLs. Conditioning on the case that the locus does contain a QTL, the prior for SNP i being the causal SNP is assigned

$$\Pr(\gamma_i = 1 \mid \boldsymbol{\gamma} \neq 0, \boldsymbol{\delta}) = \frac{e^{\sum_{l=1}^q \delta_l d_{il}}}{\sum_{k=1}^p e^{\sum_{l=1}^q \delta_l d_{kl}}}, \tag{D.3}$$

where the parameter $\boldsymbol{\delta}$ is similar to our enrichment parameter. As a result, this parametrization yields a similar expression for the PIP of SNP i,

$$\Pr(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \pi_0, \boldsymbol{\delta}) = \big[1 - \Pr(\boldsymbol{\gamma} = \boldsymbol{0} \mid \boldsymbol{y}, \boldsymbol{G}, \pi_0)\big] \cdot \frac{e^{\sum_{l=1}^q \delta_l d_{il}} \,\mathrm{BF}_i}{\sum_{k=1}^p e^{\sum_{l=1}^q \delta_l d_{kl}} \,\mathrm{BF}_k}. \tag{D.4}$$

Albeit the algebraic similarity, the parameters (π_0 and $\boldsymbol{\delta}$) in (D.4) are not so straightforwardly interpreted as $\boldsymbol{\alpha}$ in our logistic priors, partly due to the conditional nature of the prior specification (D.3). Furthermore in enrichment analysis, the M-step of the EM algorithm becomes much more involved for optimizing the objective function jointly with respect to $(\pi_0, \boldsymbol{\delta})$. In comparison, we have shown that under the parametrization of DAP-1, the maximization in the M-step is equivalent to fitting a logistic regression model for which the solutions are well-known.

Appendix E Accuracy of PIP Approximation by DAP

The DAP algorithm is designed to accurately approximate the normalizing constant C: as C^* precisely approximates C, the resulting $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})$ will be accurate, hence the marginalized PIP values. Nevertheless, as we have observed in the numerical experiments (Fig. A1) , even if the approximation of C^* is less accurate (as λ is less stringent), the approximate PIPs are still, on average, practically precise. In addition, we have also observed that DAP-1 embedded EM algorithm consistently performs well for estimating enrichment parameters in our simulation studies. In this section, we offer some theoretical discussions regarding to the precision of the PIP approximation by DAP with a focus on the performance of DAP-1 algorithm in the enrichment analysis.

First, in genetic applications, the associations are normally sparse. As a consequence, the DAP-1 assumption likely holds for the majority of the loci interrogated. Here we use the general term "locus" to refer to the genomic region where an independent DAP is applied: in genome-wide QTL mapping applications, the loci are non-overlapping LD blocks discussed in section B; in *cis*-eQTL mapping, a locus represents the cis-region of a given gene. Therefore, for most of loci, the C^* estimated by DAP-1 are indeed accurate, so are the resulting PIPs within these regions.

Second, for loci that contain more than one causal QTLs, it is clear that C^* is under-estimated by DAP-1. However, it can be argued that the approximate PIPs of the SNPs within those loci can still be, on average, accurate. To see this, we write

$$\Pr(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}) = \sum_{s=1}^{p} \frac{C_i}{C} \cdot \Pr(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}, ||\boldsymbol{\gamma}|| = s). \tag{E.1}$$

In DAP-1, we essentially estimate $\Pr(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})$ by $\Pr(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}, ||\boldsymbol{\gamma}|| = 1)$. Note that the vast majority of SNPs have overall PIPs $\to 0$, and it must be the case (in the context of genetic associations) that for such SNP k,

$$\Pr(\gamma_k = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha}, ||\boldsymbol{\gamma}|| = s) \to 0, \text{ for all } s.$$

Thus, even when $\frac{C_1}{C} \not\to 1$, the DAP-1 still provides reasonable accurate PIP estimations for the majority of SNPs that are not QTLs. The same argument can be also applied to very strong QTLs, especially the "primary" association signals whose strengths of associations measured by the single-SNP Bayes factors are orders of magnitude higher than the remaining QTLs within the same locus. Therefore, the only SNPs whose PIPs are poorly estimated by DAP-1 are those secondary QTL signals, but in most cases, it can be assured that such SNPs are in very small minority. This may explain the reason that, on average, DAP-1 is sufficient for PIP estimation in the EM algorithm. The above arguments can also be generalized to the case of the adaptive DAP algorithm where C^* is approximated less accurately.

It is worth pointing out that we would not recommend use of DAP-1 algorithm to perform the final multi-SNP QTL analysis, as it may omit strong secondary signals within a locus. In comparison, the adaptive DAP, even with relatively liberal λ, threshold generally yields satisfactory results in such setting.

Appendix F Simulation Details

F.1 Accuracy Evaluation of Adaptive DAP

In this numerical experiment, we compare the performance of the adaptive DAP algorithm to the exact Bayesian computation, in particular, we are interested in evaluating the accuracy of approximate $\Pr(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\alpha})$ and corresponding PIP values from the adaptive DAP. To be able to carry out the exact Bayesian computation with reasonable computational cost, we have to limit the number of candidate QTL SNPs in our simulation. In this experiment, we decided to set $p = 15$ and let the exact Bayesian calculation evaluates all $2^{15} = 32,768$ association models for each simulated data set.

Specifically, in each simulation, we randomly select genotypes of 15 neighboring *cis*-SNPs of a gene from the GEUVADIS data set. Keeping the population structures intact, we uniformly select 1 to 5 consistent causal QTLs and generate the phenotype measurements within each population using the linear model,

$$\boldsymbol{y} = \mu \mathbf{1} + \sum_{i=1}^{15} \beta_i \boldsymbol{g}_i + \boldsymbol{e}, \;\; \boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}), \tag{F.1}$$

where we sample $\mu \sim \text{uniform}(-2, 2)$ and set the β values of non-QTL SNPs to 0. The overall effect size of a QTL SNP is simulated by

$$\bar{\beta}_i \sim \mathrm{N}(0, 0.6^2),$$

and in each population group, we obtain

$$\beta_i \sim \mathrm{N}(\bar{\beta}_i\,,\, 0.01 \times \bar{\beta}_i^2)$$

to allow limited effect size heterogeneity across populations. Note, the details of effect size simulation is less relevant to this particular experiment, however, we consistently adopt this simulation scheme for our other simulations in this paper.

We apply both the adaptive DAP algorithm and the exact Bayesian posterior computation on a total of 1,250 simulated data sets using the same prior specification. For the DAP algorithm, we vary the threshold value in selecting high priority candidate SNPs, λ, from 0.01 to 0.05. First, we compare the true normalizing constant C with the estimated value C^* from the DAP algorithm by computing the ratio C^*/C in each simulated data set. Utilizing all SNPs of all the simulated data sets, we also calculate root-mean-square error (RMSE) to characterize the precision of PIP approximations. The result indicates that for stringent λ value, the DAP can indeed estimate the normalizing constant in very high accuracy (Tab. A1, Fig. A1), which ensures the high precision of the estimated PIPs. As the λ threshold is relaxed, the approximation of C becomes less accurate, nevertheless we observe the overall precision level of approximate PIPs is still suitable for QTL mapping applications.

Table A1: Numerical comparison between the exact calculation and the adaptive DAP algorithm at different threshold values.

λ	Mean of C^*/C	RMSE of approximate PIP
0.01	0.994	2.36×10^{-3}
0.02	0.986	5.32×10^{-3}
0.03	0.963	9.83×10^{-3}
0.04	0.921	1.40×10^{-2}
0.05	0.854	2.42×10^{-2}

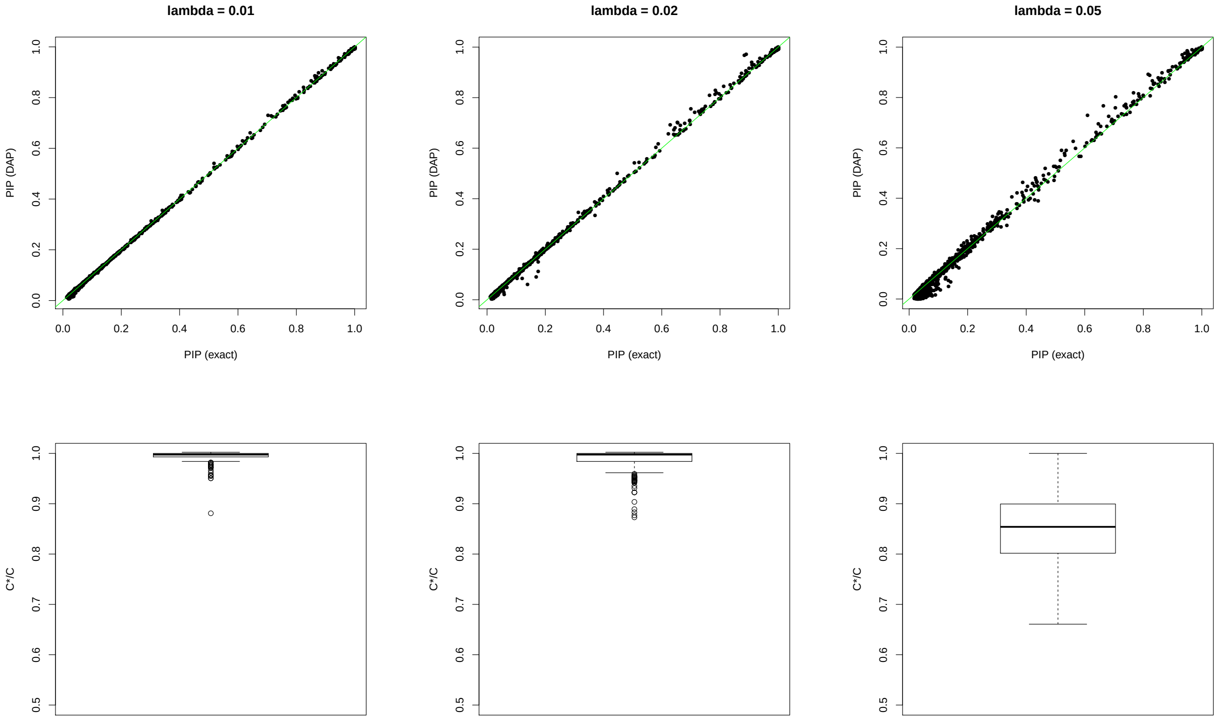

Figure A1: Assessment of accuracy of the adaptive DAP algorithm at different threshold values. In the top panel, the individual PIP approximations by the DAP are compared to the exact calculations. In the bottom panel, the distribution of C^*/C is plotted. The simulation results are obtained for threshold values $\lambda = 0.01, 0.02, 0.05$ for the DAP algorithm.

Using the simulated data set, we also benchmark the average computational time of each simulation/analysis setting and show the results in Tab. A2. All runs are performed with 10 parallel threads using the OpenMP library. For the exact calculation, the average time remains constant regardless of the number of true QTLs. The DAP algorithm represents a much reduced computational time comparing to the exact calculation. The general trend of the DAP running time is also clear (albeit few small deviations): with increasing number of true QTLs, the running time increases, and with more relaxed λ values, the running time decreases.

Table A2: Benchmark of average computational time by the DAP and exact computation. The running time is measured in second by UNIX utility program “time”. In each cell, we show the actual running time (“real” time) which is greatly reduced by parallel processing with 10 threads; in the parenthesis, the “user” time is reported, which objectively reflects the actual computational cost, i.e., this measurement is not reduced by the parallelization.

	Running Time (seconds)				
	Number of True QTLs				
Method	1	2	3	4	5
DAP ($\lambda = 0.01$)	0.097 (0.234)	0.275 (1.180)	0.733 (3.704)	1.276 (7.140)	2.527 (13.181)
DAP ($\lambda = 0.02$)	0.093 (0.268)	0.208 (0.776)	0.663 (3.128)	1.275 (6.816)	2.368 (12.965)
DAP ($\lambda = 0.03$)	0.087 (0.238)	0.133 (0.408)	0.252 (1.060)	0.844 (4.644)	1.422 (7.876)
DAP ($\lambda = 0.04$)	0.063 (0.116)	0.122 (0.312)	0.230 (0.732)	0.615 (3.064)	0.571 (2.596)
DAP ($\lambda = 0.05$)	0.050 (0.072)	0.120 (0.280)	0.139 (0.320)	0.184 (0.448)	0.180 (0.276)
Exact	19.8 (121.4)				

Finally, we use the simulated data set to examine the approximate recursive relationship derived in (C.5). Specifically, based on the results from the exact computation, we compute

$$C_s^{\#} = \frac{p-s+1}{s}\,\omega\, C_{s-1},$$

and calculate $\log_{10}\left[\frac{C_s^{\#}}{C_s}\right]$ for $s = 1, 2, ..., 14$. Fig. A2 shows the results for 4 randomly generated data sets containing $K = 1$ to 4 strong QTLs, respectively. We observe that as the model size partition is less than the size of the saturated model, expression (C.5) always severely under-estimates C_s as expected. However starting from $s = K + 1$, the estimate becomes very accurate, and the stopping rule (C.6) for halting explicit exploration works extremely well for these simulations.

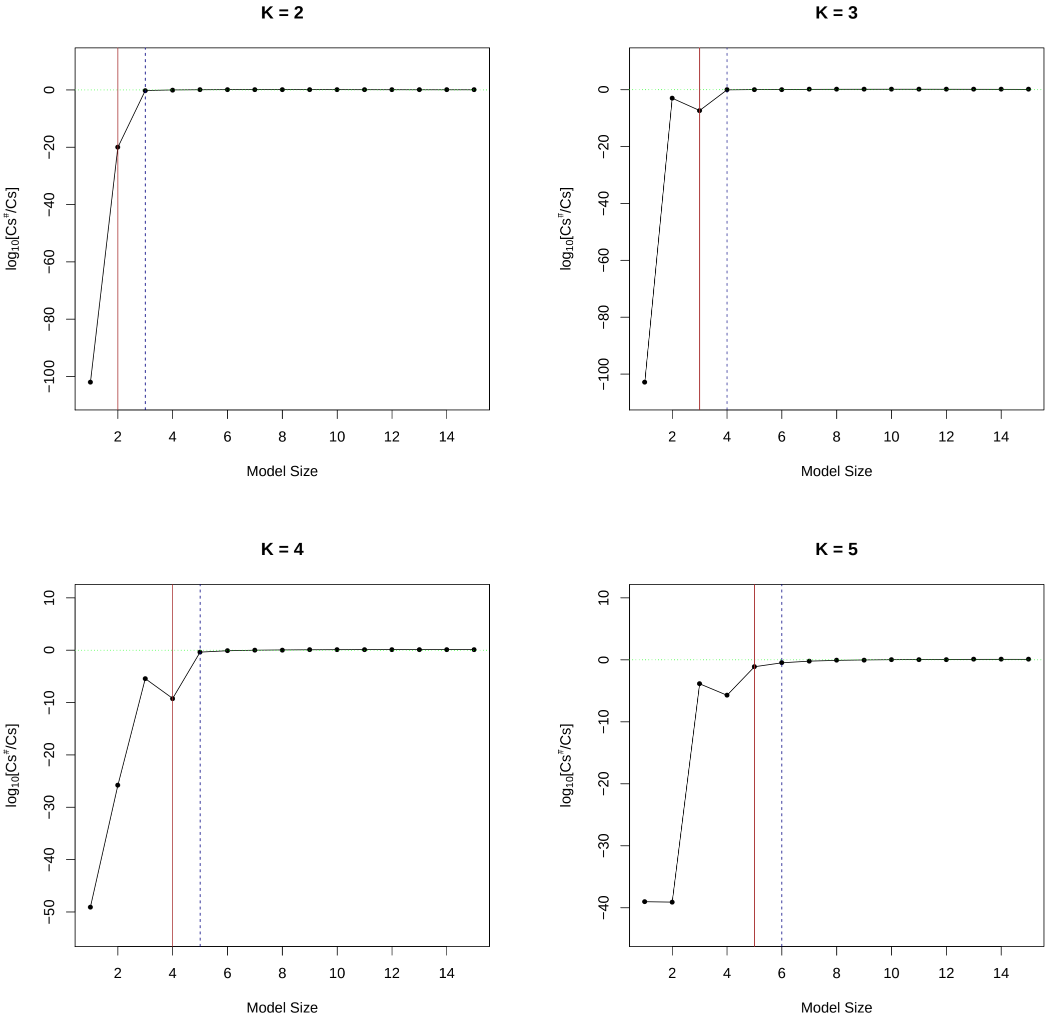

Figure A2: Examination of the recursive approximation of C_s by equation (C.5) in the simulated data sets. Each panel represents a simulated data set containing K true QTLs. The ratio of estimated value $C_s^{\#}$ (computed using the true value of C_{s-1}) over the true value C_s is plotted at log 10 scale for all model size partitions. The red vertical line indicates the size of the true association model, and the blue dotted line represents the actual stopping point that the adaptive DAP halts explicit exploration. As the model size s exceeds K, the estimation by $C_s^{\#}$ becomes very accurate in all settings.

F.2 Comparison of Adaptive DAP and MCMC

In this numerical experiment, we compare the performance of the adaptive DAP algorithm with the MCMC algorithm in multi-SNP fine mapping applications. To this end, we use the simulated data set generated in Wen *et al.* (2015) and apply the adaptive DAP algorithm using the same prior specification as the MCMC run. The details of the simulating scheme have been documented in Wen *et al.* (2015), here we only provide a brief description.

The simulation is designed to evaluate various multi-SNP analysis approaches in correctly identifying multiple independent causal QTLs using cross-population samples. We select 2,500 SNPs from the *cis*-regions of 100 random genes (i.e., 25 neighboring SNPs per region) and use their genotype data from the GEUVADIS data. As a result, the assembled genomic region consists of 100 relatively independent LD blocks with modest to high LD within each block. In each simulation, we randomly assign 1 to 4 causal QTLs and simulate a quantitative trait using the scheme described in section F.1. In total, we simulate 1,500 QTL data sets.

We applied four different types of analysis approaches to perform multi-SNP QTL mapping for each simulated data set with the goal of identifying the LD blocks that harbor the true QTLs. Those approaches are

1. a single-SNP analysis analysis which performs single-SNP meta-analysis across five population groups. A block is identified if the minimum single-SNP p-value within the block is more significant than the pre-defined threshold.

2. a conditional meta-analysis approach using the forward selection procedure. The procedure also utilizes a p-value threshold to determine if the "best" association model has been achieved. A block is identified if one of its member SNPs is included in the final "best" model.

3. an MCMC algorithm based on the hierarchical model with $\alpha_0 = \log(\frac{1}{p-1})$ and $\alpha_k = 0$ for $k > 0$. Equivalently, the model assumes each SNP is equally likely to be the causal QTL with the probability $\frac{1}{p}$, i.e., the prior expected number of causal SNP is set to 1. We compute a block level posterior inclusion probability by simply summing over the PIPs of each member SNP. A block is identified if the block-level PIP exceeds the pre-defined threshold.

4. the adaptive DAP algorithms based on the same hierarchical model as in the MCMC with λ set to 0.01 and 0.05, respectively. Same as in the MCMC analysis, we also use the block-level PIPs to identify the blocks harboring causal QTLs.

For each method, we vary the corresponding threshold for QTL calling in a wide range, from very stringent to very liberal, and record the true versus false positives at each threshold value.

We run the adaptive DAP with two different settings: $\lambda = 0.05$ and $\lambda = 0.01$, and in the main text, we present the results obtained with $\lambda = 0.01$, the comparison results between the two versions of the DAP algorithm and the MCMC are shown in Fig. A3.

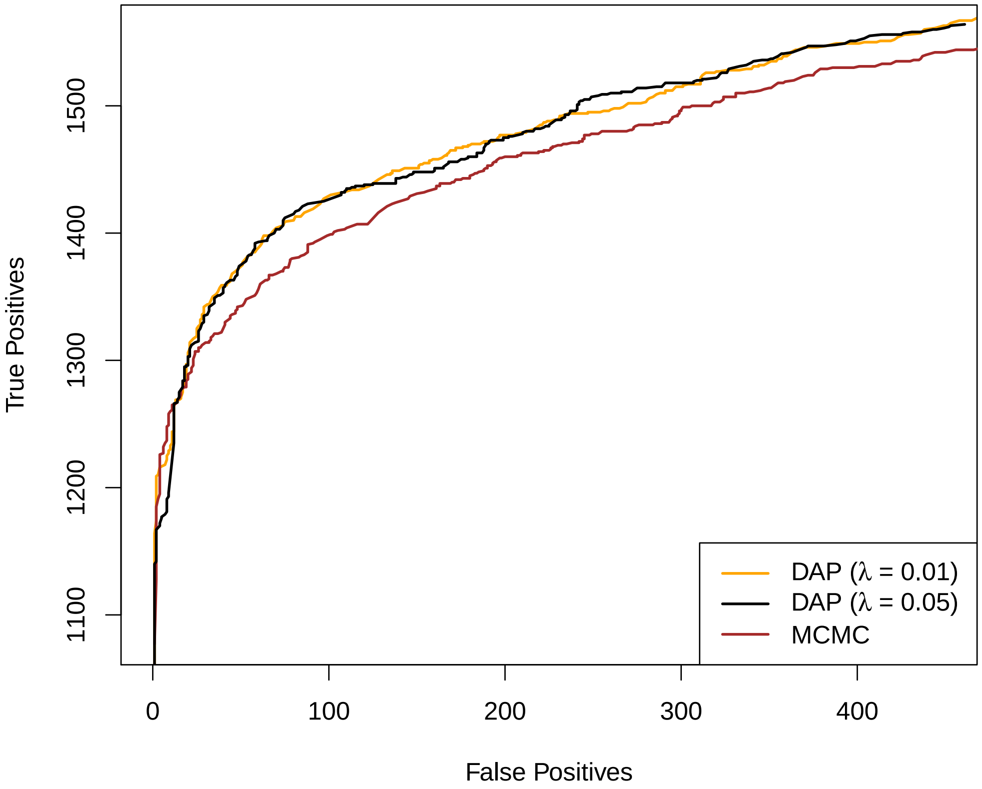

Figure A3: Additional comparisons for multi-SNP QTL mapping. Here we show the additional simulation result by running the adaptive DAP with $\lambda = 0.05$, which is most similar to the DAP outcome with the default setting ($\lambda = 0.01$), and for the most part, still outperforms the MCMC algoirthm.

It is worth emphasizing that both the MCMC and the DAP are based on the exact same Bayesian model, the difference in the outcomes reflects the performance difference of the two fitting procedures. To further investigate, we randomly select 10 data sets out of 1,500 simulations, and repeat MCMC runs with 15,000, 75,000, 250,000 and 1,000,000 sampling steps, respectively. (Note, the analysis comparing true versus false positive findings is carried out with 75,000 MCMC repeats.) We benchmark the MCMC running times and compare the resulting PIPs with the adaptive DAP ($\lambda = 0.01$) output. The results, summarized in Fig A3, Tab. A3 and Fig. 1 in the main text, indicate that adaptive DAP is more accurate and much more computationally efficient. More importantly, with the prolonged sampling steps, the MCMC results seemingly converge to the results by the DAP.

Table A3: Average running time and PIP comparison by MCMC runs with varying sampling steps. The actual running time reported from the UNIX "time" command is shown for each experiment. The DAP algorithm runs with 10 parallel threads, the average user time (i.e., approximate running time without parallelization) is 1 minute and 8.66 seconds.

	MCMC (reps)				**DAP**
	15K	75K	250K	1M	$\lambda = 0.01$
Running Time (real)	4m 2.79s	10m 28.37s	28m 50.00s	107m 46.75s	28.44s
RMSE of PIP (w.r.t DAP)	0.080	0.052	0.034	0.030	–

F.3 Evaluation of Enrichment Analysis

We perform simulation studies to evaluate the performance of DAP-embedded EM algorithm in enrichment analysis. Our simulation setting mimics the genome-wide *cis*-eQTL mapping application, however at a reduced scale. Specifically, we select a subset of 1,500 random genes from the GEUVADIS data. For each gene, 50 *cis*-SNPs are used in the simulation and we annotate 20% 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 QTLs of each gene, we then apply the scheme described in section F.1 to simulate the effect sizes of the QTLs and the expression levels across multiple population groups. We set $\alpha_1 = 0.00, 0.25, 0.50, 0.75, 1.00$, and for each α_1 value, we simulate 100 data sets. To analyze the simulated data set, we use two different implementations of the EM algorithm with the E-step approximated by the DAP-1 and the adaptive DAP with $\lambda = 0.05$, respectively. 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. In addition, we apply a naive two-stage enrichment analysis method, which first performs single-SNP association tests and classifies the association status of each SNP based on its false discovery rate (FDR, computed by Storey's q-value method) with the significance cutoff 0.01. Based on the classification result, the enrichment parameter is estimated by a 2 × 2 contingency table (which is equivalent to a simple logistic model regressing the association status on the binary annotation). We include this method because of its similarity to some commonly applied enrichment analysis strategy.

Fig. A4 plots the point estimates of α_1 ± the standard errors for each analysis method in each simulation setting across 100 simulated data sets. The estimates from the adaptive DAP and DAP-1 are seemingly unbiased. As expected, the variability of their point estimates is higher than the "best case" method because of the uncertainty in determining the true association status of each SNP. In comparison, the naive two-stage method consistently and severely underestimate the enrichment parameters, for which we offer an theoretical explanation at the end of the subsection. Although the results indicate that the adaptive DAP generate more accurate estimate in average, the performance of DAP-1 is very much comparable and completely suitable for practical applications. In addition, DAP-1 presents a great advantage in computational efficiency: the average running time for the DAP-1 embedded EM algorithm (with 10 parallel threads in E-step) is 65.05 seconds; in comparison, the adaptive DAP embedded EM runs 387.30 seconds in average (which is a combination of slightly longer iteration and longer running time per iteration). Finally, we note that both the adaptive DAP and DAP-1 under-estimate the α_0 parameter: in average, DAP-1 estimates $\hat{\alpha}_0 = -4.62$ and the adaptive DAP yields $\hat{\alpha}_0 = -4.32$ (recall, the truth is $\alpha_0 = -4.00$. This is expected, because not all QTLs are detectable from the observed association data with limited sample sizes. Therefore, the priors constructed in the final QTL analysis using the point estimates from the EM algorithm are in general *slightly conservative*, which is mostly welcomed in practice (Lappalainen *et al.*, 2013, Fairfax *et al.*, 2014).

Fig. A5 highlights the comparisons of individual estimates along with their 95% confidence intervals from 10 randomly selected simulated data sets for each setting. We omit the naive method in this comparison because of its poor performance. The plots display a similar pattern as we observed in Fig. A4 (which focuses on the variability of the point estimates). Overall, the 95% confidence intervals generated by DAP-1 and the adaptive DAP methods both display excellent coverage probability. The figure also indicates that there is considerable uncertainty in the enrichment analysis.

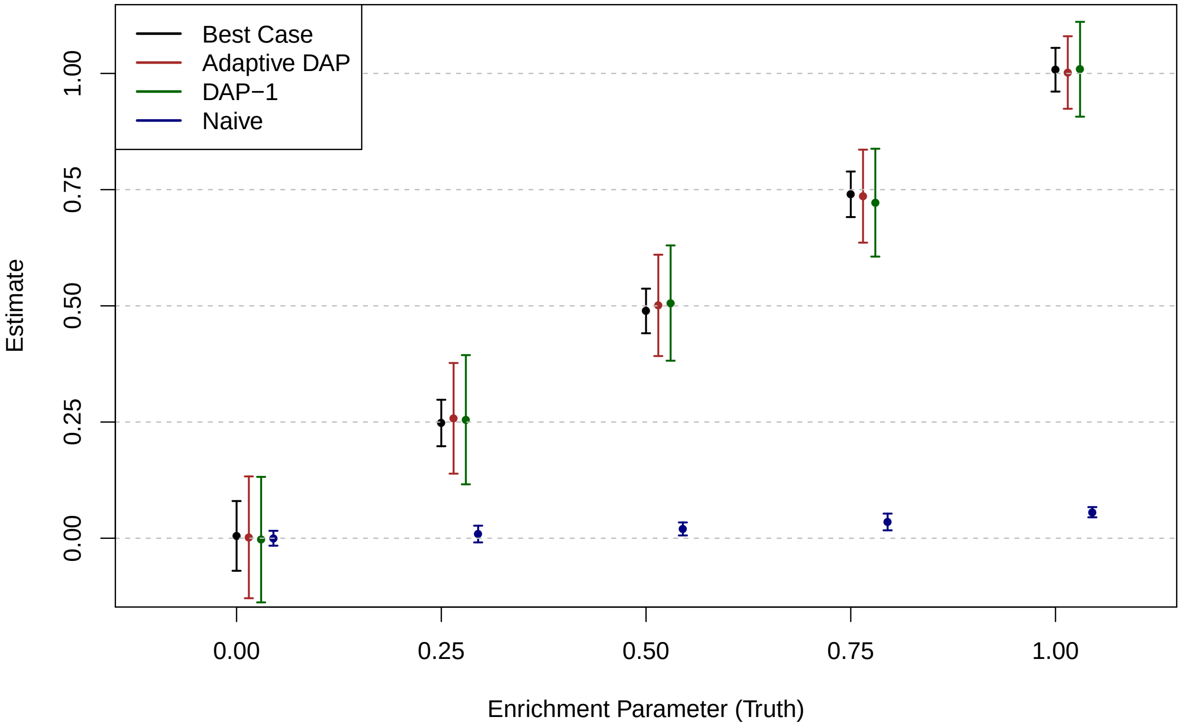

Figure A4: Point estimates of the enrichment parameter by various analysis methods in different simulation settings. The point estimate of $\alpha_1 \pm$ standard error (obtained from 100 simulated data sets) for each method is plotted for each simulation setting. The "best case" method uses the true association status and represents the optimal performance for any enrichment analysis method. Both adaptive DAP and DAP-1 methods yield unbiased estimates in all settings with the adaptive DAP being generally more accurate. The naive two-stage method consistently and severely under-estimates the enrichment parameter.

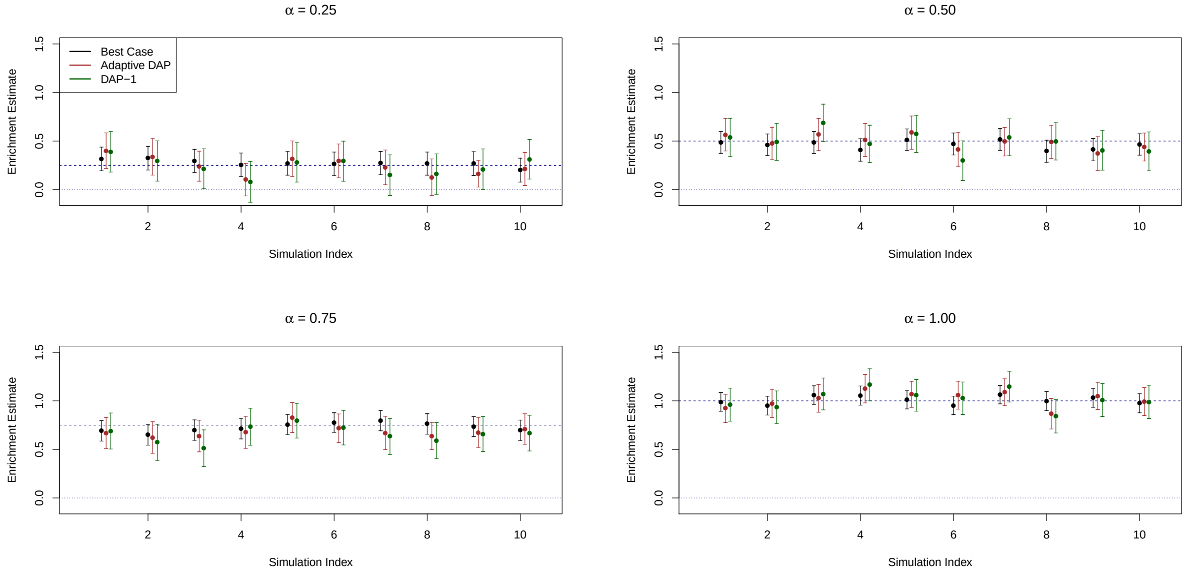

Figure A5: Comparison of individual estimates of enrichment parameter and their uncertainty quantification. Each panel represents a different simulation setting. We plot the point estimates of α_1 along with their 95% confidence intervals for each method using 10 randomly selected simulated data sets. In all settings, all methods compared ("best case", EM with adaptive DAP and EM with DAP-1) show desired coverage probability. The figure also highlights the considerable uncertainty in enrichment analysis.

Finally, we offer a brief discussion on the drawbacks of the naive two-stage analysis method. Although intuitively sensible, the hypothesis testing based classification is largely problematic. It should be noted that the accuracy of the enrichment analysis is greatly impacted by the overall errors of classifying the binary association status according to the single-SNP testing results. Unfortunately, the hypothesis testing approaches generally emphasize less on type II errors which are highly critical for overall classification errors. In fact, the common practice of hypothesis testing tends to impose very stringent controls over type I errors which can lead to an elevated level of type II errors. Furthermore, the single-SNP association based test does not account for LD, which also complicates the classification procedure. With poor classification, the estimation from the second stage becomes unwarranted. In comparison, the hierarchical model based approaches completely abandon the classification step, instead, they carry over the uncertainty of individual SNP association status (accounting for LD) into the enrichment analysis. In the theory of missing data analysis, this presents a superior statistical approach.

Appendix G Re-analysis of GEUVADIS Data

We apply the improved integrative multi-SNP QTL mapping algorithms to re-analyze the eQTL data from the GEUVADIS project. The data set contains RNA-seq data on lymphoblastoid cell line (LCL) sample from five populations: the Yoruba (YRI), CEPH (CEU), Toscani (TSI), British (GBR) and Finns (FIN). In our analysis, 420 samples who are densely genotyped in the 1000 Genomes Phase I data release are selected for eQTL mapping. The genotype and expression data are directly downloaded from the GEUVADIS project website, and we perform the additional quality control steps to remove the potential confounding factors in the RNA-seq data and re-normalize the expression levels for each gene within each population (details are described in Wen *et al.* (2015)). In total, 11,838 protein coding and lincRNA genes are included in our analysis.

Applying the DAP algorithm, we aim to perform the integrative analysis of *cis*-eQTLs and genomic annotations, including SNP distance to transcription start site (TSS) of the target gene and transcription factor (TF) binding site annotations produced by the CENTIPEDE model (Pique-Regi *et al.*, 2011). We are particularly interested in quantifying the difference in eQTL enrichment levels for the following two categories of SNPs while controlling for the SNP distance to respective TSS:

1. SNPs that are predicted to disrupt TF binding in a sequence motif, i.e., binding SNPs
2. SNPs that simply reside in a DNase I footprint region but are otherwise predicted to have weak effects on TF binding, i.e., footprint SNPs

Because the SNP distance to TSS is known to have a strong non-linear effect on enrichment level of *cis*-eQTLs Veyrieras *et al.* (2008), Wen *et al.* (2015), we group the SNPs into non-overlapping 1Kb bins and treat the belonging bin as a categorical annotation for each SNP.

The same analysis has been attempted by Wen *et al.* (2015) using the MCMC algorithm to perform E-step (EM-MCMC) in enrichment analysis. However due to the computational restraints, they were only able to run a single iteration of the EM algorithm. Although the hypothesis testing results all remain valid, i.e., both footprint and binding SNPs are indeed enriched in *cis*-eQTLs with quite distinct significance measures, the enrichment levels are likely to be under-estimated.

We run the complete DAP-1 embedded EM algorithm (EM-DAP1, available at `https://github.com/xqwen/dap/tree/master/EM_dap1`) to re-analyze the same data set. The full EM algorithm runs 25 iterations to meet our convergence criteria that requires increment of log-likelihood ≤ 0.01 between the two consecutive iterations (Fig. A6). The complete EM run takes about less than a hour on a Linux box with a single 8-core Intel Xeon 2.13GHz CPU and 96G of memory. In comparison, using the MCMC algorithm, a single round of E-step execution costs about 84 hours of computational time to fully process all 11,838 genes on the same computing system. We then perform the final round of fine mapping using the adaptive DAP algorithm with $\lambda = 0.01$.

First, we compare the estimates of the enrichment parameters by the EM-MCMC with the EM-DAP1 after a single iteration, and we find that they are remarkably similar (Table A4), which is also consistent with what we have observed in our simulation studies for the enrichment analysis. Also as expected, the final estimates from the EM-DAP1 are much greater than the estimates from a single iteration of EM. The final multi-SNP eQTL mapping results are also quantitatively different comparing to the previous results (Wen *et al.*, 2015), which is mostly attributed to the combination of the better enrichment estimates and the better multi-SNP mapping algorithm (i.e., adaptive DAP).

Table A4: Comparison of enrichment estimates by EM-DAP1 and EM-MCMC after a single iteration.

	Footprint SNPs		Binding Variants	
Method	α	95% C.I.	α	95% C.I.
EM-MCMC	0.140	$(0.039, 0.239)$	0.392	$(0.322, 0.489)$
EM-DAP1	0.119	$(-0.007, 0.245)$	0.406	$(0.303, 0.509)$

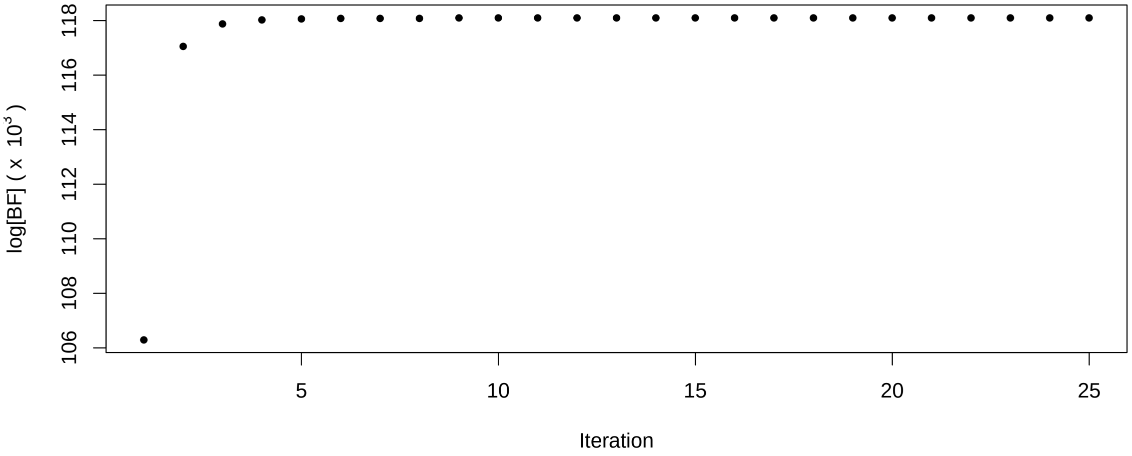

Figure A6: Traceplots of the marginal likelihood (in Bayes factor at log scale) during the DAP-1 embedded EM run in analyzing the GEUVADIS data.

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**(6235), 648–660.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**(433), 109–122.

Berisa, T. and Pickrell, J. K. (2015). Approximately independent linkage disequilibrium blocks in human populations. *bioRxiv*, page 020255.

Consortium, E. P. *et al.* (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**(7414), 57–74.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., *et al.* (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, **343**(6175), 1246949.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLOS Genetics*, **9**(5), 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**(7539), 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.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, **94**(4), 559–573.

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**(3), 447–455.

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

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**(10), e1000214.

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.

Wen, X., Stephens, M., *et al.* (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions. *The Annals of Applied Statistics*, **8**(1), 176–203.

Wen, X., Luca, F., and Pique-Regi, R. (2015). Cross-population joint analysis of eqtls: Fine mapping and functional annotation. *PLOS Genetics*, **11**(4), e1005176.