Computational Approaches for Analyzing High-Throughput Genomic Data

by

Yeji Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Xiaoquan Wen, Chair
Professor Michael Boehnke
Professor Stephen Parker
Professor Laura Scott

Yeji Lee

yejilee@umich.edu

ORCID iD: 0000-0002-9011-8613

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my advisor, Professor Xiaoquan William Wen, who have been supportive of all my researches and academic life. This dissertation would not have been possible without his guidance and inspiration, I am truly indebted to him.

I am grateful to all of my dissertation committee members. Professor Michael Boehnke, professor Laura Scott and professor Stephen Parker have provided me professional and personal guidance on my communication skills and the dissertation.

I would also like to thank every members in Center for Statistical Genetics in University of Michigan. Their passionate attitudes and valuable opinions on my projects have been inspired me in so many ways.

Last but not least, I would like to express my gratitude to my parents and younger sister, who have been endlessly supported me with their warm love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

With the improvement of high-throughput technologies, association studies related to molecular phenotypes have become increasingly significant. Associated genetic variants found from studies based on high-throughput omics experiments provide valuable information to help understand biological mechanisms behind complex traits. While analyses using high-throughput data can play a crucial role to study complex traits, many analytical challenges remain unresolved.

This dissertation primarily focuses on two outstanding issues in genetic association analysis of high-throughput sequence data. First, when incorporating functional annotations into multi-SNP association analyses and the number of candidate SNPs increases, computational burden increases. Second, there is a need to identify reproducible signals between studies. Measuring reproducibility between assays in high-throughput experiments and association results between studies is crucial to assess the quality of the overall procedures and the association evidence.

In Chapter 2, we propose an algorithm to incorporate functional annotations into Bayesian multi-SNP analysis based on a probabilistic hierarchical model. The proposed algorithm, name as deterministic approximation of posteriors (DAP), shows superior accuracy and computational efficiency over the existing methods, including Markov Chain Monte Carlo (MCMC) algorithms to fit a sparse Bayesian variable selection model.

In Chapter 3, we propose a probabilistic quantification of association evidence,

accounting for linkage disequilibrium (LD). By identifying a set of SNPs in LD and representing a single association signal, we are able to construct credible sets and perform appropriate false discovery rate (FDR) control in Bayesian multi-SNP association analysis. We also derive a set of sufficient summary statistics that lead to equivalent inference results as using individual-level data.

In Chapter 4, we propose a set of computational methods to measure reproducibility among high-throughput sequencing experiments. In particular, we propose a statistical approach to take advantage of the fact that a strong and genuine signal is expected to show the same directional effects in multiple studies.We design a novel Bayesian hierarchical model and estimate the posterior probability of each testing unit (e,g, SNP) being reproducible under a proposed set of prior probabilities. We also propose visualization tools and quantification measures tool to assess the overall reproducibility among multiple experiments.

In three chapters of the dissertation, we discuss several issues in studies utilizing high-throughput data and propose computational methods to deal with these issues.

CHAPTER I

Introduction

With the improvement of high-throughput technologies, association study related to molecular phenotype has become significant recently. Recently many studies have been successfully discovered quantitative trait loci (QTL) that are associated with the regulation of gene expression, histone modification, DNA methylation. These findings, along with genomic variants that are revealed from genomewide association studies (GWAS), are expected to provide valuable evidence to understand underlying mechanisms behind complex traits.

There have been several issues in association studies utilizing high-throughput data. One of the issues is related to the integrative analysis functional annotations and multi-SNP analysis.

Thanks to large-scale studies that utilize high-through data, functional annotations on regulatory variants become feasible. This also make association studies incorporating functional annotations feasible. Most studies considering functional annotations have been commonly performed at single-SNP analyses. However, single-SNP association analysis often fails to identify multiple signals co-exist in small genomic regions. Incorporating annotations into multi-SNP analysis studies can enables to find multiple independent signals exist in small genomic region.

It can also improve the power to identify QTL and provided valuable information on understanding molecular mechanisms. One of Major challenges in incorporating annotations into multi-SNP analyses is that as number of tested SNP increases, computational burden increases rapidly.

Another issue of utilizing high-throughput data is measuring reproducibility between high-throughput assays and association studies. To ensure the quality of data processing, it is crucial to quantify the degree of concordance between replicate assay. Also, some QTLs identified from a single association study often fails to be discovered in other studies, since their signals are not strong enough to overcome study-specific variations. Therefore, classifying eQTLs that can be reproducible between studies can provide valuable information on finding strong causal variants. While several methods have been proposed to measure reproducibility, they are based on rank-transformed information and ignore the directional information of estimated effects.

In this dissertation, we propose statistical methods to deal with the two issues discussed. In Chapter 2, we propose an algorithm to incorporate functional annotations into Bayesian multi-SNP analysis based on a probabilistic hierarchical model. This algorithm is called deterministic approximation of posteriors (DAP). Compared with exiting methods, DAP shows the improvement in accuracy and computational efficiency in fitting a sparse Bayesian variable selection model. We also apply DAP to *cis*-eQTL results of GTEx study.

In Chapter 3, we propose a probabilistic quantification of association evidence which accounts for linkage disequilibrium (LD). By identifying a set of SNPs in LD that represent a single association signal, the construction of credible sets and performing appropriate false discovery rate (FDR) in Bayesian multi-SNP association analysis become feasible. We also derive a set of sufficient summary statistics that

leads to the equivalent inference results as using individual-level data.

In Chapter 4, we propose computational methods to measure reproducibility for high-throughput experiments. First, we propose a visualization tool to assess the degree of concordance, and a measure to quantify it. Second, we propose a Bayesian hierarchical model to estimate the posterior probability of each testing unit being reproducible. This model assumes that a strong and genuine signal is expected to show the same directional effects across studies.

CHAPTER II

Effcient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors

2.1 Introduction

Association analysis has become a powerful tool for identifying genetic variants that impact complex traits at both the organismal and molecular levels: in the past decade, genome-wide association studies (GWAS) have successfully identified a rich catalog of genetic variants that are linked to many human diseases. Most recently, molecular QTL mapping has revealed an abundance of quantitative trait loci (QTLs) for cellular phenotypes such as gene expression *Lappalainen et al.* (2013), *Ardlie et al.* (2015), chromatin accessibility *Degner et al.* (2012), histone modifications *McVicker et al.* (2013) and DNA methylation *Banovich et al.* (2014). Nevertheless, the causal molecular pathways from genetic variants to complex phenotypes remain poorly understood *Albert and Kruglyak* (2015). This is mainly because a good proportion of identified trait-associated variants are located in the non-coding regions of the genome, and our knowledge of the functional roles of non-coding variants is generally lacking. With the recent advancements in high-throughput experimental technologies, functional annotations for regulatory variants have become increasingly available *ENCODE Project Consortium* (2012), *Kundaje et al.* (2015), *Ardlie et al.* (2015). As a consequence, it is now feasible to perform association analysis incorpo-

rating functional genomic annotations. The integrative analysis strategy presents two obvious advantages: first, it improves the power of association analysis by prioritizing functional variants; second, it helps to reveal the underlying molecular mechanisms that lead to the observed associations.

In the past, integrative analysis was typically performed by searching for overlaps between putative association signals and SNP annotations. This analysis strategy implicitly assumes that a SNP with specific genomic annotations is likely causal. To justify the results from the post-hoc overlapping analysis, quantitatively validating this implicit assumption from the observed association data, which essentially requires estimating the enrichment levels of the annotations in the association signals, is critical. This point becomes particularly crucial when multiple types of annotations are used, and a rigorous quantitative enrichment analysis should help to determine which annotations are relevant and how much we should weigh each annotation. The availability of functional annotations also enables high-resolution multi-SNP genetic association analysis. From both GWAS and molecular QTL mapping studies, it is increasingly evident that multiple independent association signals can co-exist in a relatively small genomic region. Multi-SNP fine-mapping analysis has now become a standard procedure to tease out potential multiple association signals. It is only natural that genomic annotations are integrated into this process.

Recently, a few computational approaches for integrative enrichment and association analysis have been proposed and successfully demonstrated in molecular QTL mapping *Veyrieras et al.* (2008), *Gaffney et al.* (2012) and GWAS *Pickrell* (2014), *Kichaev et al.* (2014). However, these existing approaches make simplifying assumptions for either enrichment analysis *Kichaev et al.* (2014) or multi-SNP fine-mapping analysis *Veyrieras et al.* (2008), *Pickrell* (2014). Therefore, the power of integrative

analysis has not been maximized and can be further improved. In addition, computational efficiency has always been a hurdle in terms of applying a probabilistic integrative analysis approaches to genetic data at the genome-wide scale.

In this chapter, we propose a probabilistic hierarchical model that is generalized from our recent work *Wen et al.* (2015b) to describe multi-SNP genetic associations while accounting for functional genomic annotations. Based on this model, we consider analyzing genetic association data in two settings: traditional GWAS and molecular *cis*-QTL mapping studies. Note that a distinct feature of molecular QTL mapping is that tens of thousands (or hundreds of thousands) of molecular phenotypes (e.g., gene expression, DNA methylation) are simultaneously measured and analyzed, which imposes some unique statistical challenges. In addition, the candidate genomic region for each molecular phenotype is typically defined in the proximity of relevant genomic landmarks of the corresponding molecular phenotypes (e.g., transcription start site of a target gene for expression phenotypes) and much smaller in length (usually spanning 1 to 2 Mb) comparing to GWAS. We outline a 3-stage inference procedure to sequentially perform enrichment analysis, QTL discovery and multi-SNP fine-mapping. One of our main contributions is a computationally efficient algorithm for Bayesian multi-SNP association analysis. This fast fitting algorithm, named Deterministic Approximation of Posteriors (DAP), facilitates the proposed rigorous integrative inference procedure. Compared to the alternative fitting algorithm, i.e., the Markov Chain Monte Carlo (MCMC) algorithm, we show that the DAP is several hundreds times faster and more accurate for genetic association analysis. Taking full advantage of the DAP algorithm, we lay out the analytic strategies for analyzing genetic association data from GWAS and molecular *cis*-QTL mapping studies, and we demonstrate the proposed procedures through a series of

simulation studies and real data applications.

2.2 Methods

2.2.1 Model and Notation

First, we consider a generic setting of association analysis of a single quantitative trait and p SNPs both measured for n unrelated individuals. We model the genotype-phenotype association using a multiple linear regression model,

$$\vec{\boldsymbol{y}} = \mu\mathbf{1} + \sum_{i=1}^{p} \beta_i \vec{\boldsymbol{g}}_i + \vec{\boldsymbol{e}},\ \vec{\boldsymbol{e}} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}). \tag{2.1}$$

For each SNP i, we denote its binary association status, γ_i, by dichotomizing its corresponding genetic effect β_i, i.e, $\gamma_i = 1$ if $\beta_i \neq 0$ and 0 otherwise. In particular, we refer to the causal SNPs for which $\gamma_i = 1$ as the quantitative trait nucleotides (QTNs) *Veyrieras et al.* (2008). Our primary interest for association analysis is the inference of $\vec{\boldsymbol{\gamma}} := (\gamma_1, ..., \gamma_p)$. To integrate genomic annotation into the association analysis, we assume that having certain genomic features will increase (or decrease) the odds that a particular SNP is a QTN. Equivalently, certain genomic features are enriched (or depleted) in QTNs. We quantitatively represent this assumption using an *a priori* independent logistic model for each γ_i, i.e.,

$$\log\left[\frac{\Pr(\gamma_i = 1)}{\Pr(\gamma_i = 0)}\right] = \alpha_0 + \sum_{k=1}^{q} \alpha_k d_{ik}, \tag{2.2}$$

where $\vec{\boldsymbol{d}}_i := (d_{i1}, \ldots, d_{iq})$ denotes q genomic annotations that are specific to SNP i at a particular locus and $\alpha_1, ..., \alpha_q$ are referred to as the *enrichment parameters*. Note that the annotations can be either categorical or continuous in this framework. We assume that the phenotype data, $\vec{\boldsymbol{y}}$, the genotype data, $\boldsymbol{G} := (\vec{\boldsymbol{g}}_1, ..., \vec{\boldsymbol{g}}_p)$, and the annotation data, $\boldsymbol{D} := (\vec{\boldsymbol{d}}_1, ..., \vec{\boldsymbol{d}}_p)$, are observed, while the enrichment parameters, $\vec{\boldsymbol{\alpha}} := (\alpha_0, \alpha_1, ..., \alpha_q)$, are unknown.

For molecular QTL mapping, tens of thousands of phenotypes are simultaneously measured, and we denote the collection of all measured phenotypes by $\boldsymbol{\mathcal{Y}} := (\vec{\boldsymbol{y}}_1, ..., \vec{\boldsymbol{y}}_L)$. For each phenotype, a small genomic region, typically spanning 1 to 2 Mb and on average containing a few thousands of SNPs, is pre-defined as the candidate locus in the proximity of relevant genomic landmarks of the corresponding molecular phenotypes, and we denote the union of the SNP genotypes from all candidate loci by $\boldsymbol{\mathcal{G}} := (\boldsymbol{G}_1, ..., \boldsymbol{G}_L)$. Similarly, we use $\boldsymbol{\mathcal{D}} := (\boldsymbol{D}_1, ..., \boldsymbol{D}_L)$ and $\boldsymbol{\Gamma} := (\vec{\boldsymbol{\gamma}}_1, ..., \vec{\boldsymbol{\gamma}}_L)$ to denote the collections of annotations and latent association status, respectively.

In GWAS, there is usually only one phenotype of interest, which can be viewed as a special case of molecular QTL mapping. Nevertheless, it is important to note that the candidate region for GWAS spans the whole genome.

2.2.2 Inference Procedure

We propose an inference procedure consisting of three inter-related stages to fit the proposed hierarchical model. Sequentially, these stages are as follows:

1. estimating the enrichment parameter $\vec{\boldsymbol{\alpha}}$ using the full data $\mathcal{Y}, \mathcal{G}$ and $\mathcal{D}$ for enrichment analysis
2. screening candidate loci for QTL discovery
3. performing multi-SNP fine-mapping for the high-priority loci identified in stage 2

The maximum likelihood estimate (MLE) of $\vec{\boldsymbol{\alpha}}$ can be obtained by the EM algorithm proposed in our recent work *Wen et al.* (2015b). Briefly, the EM algorithm treats $\boldsymbol{\Gamma}$ as missing data and pools information across all available loci. In the E-step, the posterior inclusion probability (PIP) for each SNP i at each locus l (namely,

$\Pr(\gamma_{l_i} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}^{(t)}))$ is computed given the current estimate of $\vec{\boldsymbol{\alpha}}$; in the M-step, a logistic regression model is fit by plugging in the PIPs as the response variables and SNP annotations as predictors. The estimate of $\vec{\boldsymbol{\alpha}}$ is subsequently updated by the corresponding fitted regression coefficients.

Given the MLE of the enrichment parameter, $\hat{\vec{\boldsymbol{\alpha}}}$, we then attempt to identify genomic loci that are likely to harbor causal QTNs. This is achieved by testing the null hypothesis, $H_0 : \vec{\boldsymbol{\gamma}}_l = \mathbf{0}$, for each candidate locus l using a Bayesian false discovery rate (FDR) control procedure. Specifically, the null hypothesis is rejected if the locus-level posterior probability $\Pr(\vec{\boldsymbol{\gamma}}_l = \mathbf{0} \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \hat{\vec{\boldsymbol{\alpha}}})$ is smaller than the pre-defined threshold determined by the observed data and desired FDR control level *Wen* (2016). At the end of this stage, we gather a list of potential QTLs for fine-mapping.

Finally, we perform multi-SNP fine-mapping analysis for the identified QTLs. In particular, we compute the posterior distribution for each locus l, namely, $\Pr(\vec{\boldsymbol{\gamma}}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \hat{\vec{\boldsymbol{\alpha}}})$, to i) identify potentially multiple independent association signals within locus l and ii) assess the importance of each SNP by computing its PIP, i.e., $\Pr(\gamma_{l_i} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}^{(t)})$. A credible set of potential causal SNPs for each independent signal can then be constructed from the resulting PIPs in a manner similar to previously proposed methods *Maller et al.* (2012), *Wen et al.* (2015b). This Bayesian approach for multi-SNP analysis has been known to present some unique advantages over the traditional conditional analysis approach. For example, it fully accounts for patterns of linkage disequilibrium (LD) and shows superior power in discovering independent association signals *Guan and Stephens* (2011), *Wen et al.* (2015b).

This 3-stage procedure represents a coherent empirical Bayes strategy to fit the proposed hierarchical model for inference. In all three stages, the computational

difficulty lies in the efficient evaluation of the posterior probability $\Pr(\vec{\gamma}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}})$. We propose an algorithm to tackle this problem in the following sections. The software package implementing the computational approaches (in C++ programming language) is freely available (Web Resources).

2.2.3 Deterministic Approximation of Posteriors

The computation of the target posterior probability $\Pr(\vec{\gamma}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}})$ is conceptually straightforward by applying the Bayes theorem, i.e.,

$$\Pr(\vec{\gamma}_l = \vec{\gamma} \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}) = \frac{\Pr(\vec{\gamma} \mid \vec{\boldsymbol{\alpha}})\,\mathrm{BF}(\vec{\gamma})}{\sum_{\vec{\gamma}'} \Pr(\vec{\gamma}' \mid \vec{\boldsymbol{\alpha}})\,\mathrm{BF}(\vec{\gamma}')}, \tag{2.3}$$

where the Bayes factor

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

represents the marginal likelihood function of $\vec{\gamma}_l$ evaluated at $\vec{\gamma}$. Based on (2.3), the PIP of each candidate SNP can be subsequently marginalized from $\Pr(\vec{\gamma}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}})$.

For any given $\vec{\gamma}$ value, both the Bayes factor (whose computation involves integrating out the the nuisance parameters μ, β and σ^2) and the prior probability can be analytically evaluated *Servin and Stephens* (2007), *Wen* (2014). The difficulty lies in evaluating the normalizing constant

$$C := \sum_{\vec{\gamma}} \Pr(\vec{\gamma}_l = \vec{\gamma} \mid \vec{\boldsymbol{\alpha}})\,\mathrm{BF}(\vec{\gamma}).$$

For a locus consisting of p candidate SNPs, the exact computation requires enumerating all 2^p possible $\vec{\gamma}$ values; hence, it is intractable even for modest p. Previously, the only feasible solution was to employ a Markov Chain Monte Carlo (MCMC) algorithm *Guan and Stephens* (2011), *Wilson et al.* (2010), *Wen et al.* (2015b). However, the MCMC algorithm is computationally too costly in our grand scheme for integrative genetic association analysis: the evaluation of $\Pr(\vec{\gamma}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}})$ for every locus is

required for each E-step in the EM algorithm for enrichment analysis. Furthermore, the inherent stochastic variation in the MCMC algorithm may affect the performance and reproducibility of the overall analysis.

Here, we present an alternative algorithm to perform deterministic approximation of posteriors (DAP) for each locus and efficiently compute PIPs for all candidate SNPs. This algorithm is mainly motivated by two observations in genetic association analysis. First, in almost all genetic applications, the number of convincing QTLs (i.e., those have relatively large effect sizes) discovered from the association data are typically small compared with the number of candidate SNPs within a candidate locus (typically 1 to 2 Mb). In molecular QTL mapping, this observation is also supported by many recent experimental work *Patwardhan et al.* (2009), *Findlay et al.* (2014), *Savic et al.* (2013). It 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 a few SNPs are likely to have non-negligible posterior probabilities within a locus. 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 design 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 within each candidate locus. 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. More specifically,

the proposed DAP algorithm approximates the normalizing constant C by

$$C^* = \sum_{\vec{\boldsymbol{\gamma}}' \in \Omega} \Pr(\vec{\boldsymbol{\gamma}}_l = \vec{\boldsymbol{\gamma}}' \mid \vec{\boldsymbol{\alpha}}) \, \mathrm{BF}(\vec{\boldsymbol{\gamma}}') + \epsilon, \tag{2.4}$$

where Ω denotes a subset of the selected most plausible models to be explored explicitly and ϵ is an estimate of the approximation error $C - \sum_{\vec{\boldsymbol{\gamma}}' \in \Omega} \Pr(\vec{\boldsymbol{\gamma}}_l = \vec{\boldsymbol{\gamma}}' \mid \vec{\boldsymbol{\alpha}}) \, \mathrm{BF}(\vec{\boldsymbol{\gamma}}')$. The key to the DAP algorithm is the construction of the set Ω: it is desirable that models in Ω capture the vast majority of the posterior probability mass; on the other hand, Ω should be compact enough for efficient exploration. In this chapter, we propose two different approaches to construct Ω. In both cases, we define the size of the association model, $||\vec{\boldsymbol{\gamma}}_l||$, as the number of assumed QTNs (also known as the 0-norm of the vector $\vec{\boldsymbol{\gamma}}_l$), i.e., $||\vec{\boldsymbol{\gamma}}_l|| = \sum_{i=1}^{p} \gamma_{l_i}$, and partition the complete model space of $\{\vec{\boldsymbol{\gamma}}_l\}$ by the size of association models, i.e., $\{\vec{\boldsymbol{\gamma}}_l\} = \{||\vec{\boldsymbol{\gamma}}_l|| = 0\} \cup \{||\vec{\boldsymbol{\gamma}}_l|| = 1\} \cup \cdots \cup \{||\vec{\boldsymbol{\gamma}}_l|| = p\}$.

Adaptive DAP Algorithm

The first approach, named adaptive DAP, includes the null model and all the single SNP association models in the candidate set Ω. For a larger size of candidate models, it approximates $C_s := \sum_{||\vec{\boldsymbol{\gamma}}||=s} \Pr(\vec{\boldsymbol{\gamma}} \mid \vec{\boldsymbol{\alpha}}) \, \mathrm{BF}(\vec{\boldsymbol{\gamma}})$ by a corresponding estimate $C_s^* = \sum_{\vec{\boldsymbol{\gamma}} \in \Omega_s} \Pr(\vec{\boldsymbol{\gamma}} \mid \vec{\boldsymbol{\alpha}}) \, \mathrm{BF}(\vec{\boldsymbol{\gamma}})$, where Ω_s consists of a subset of association models with size s but is constructed only from a set of adaptively selected high-priority SNPs. 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 $||\vec{\boldsymbol{\gamma}}|| = s - 1$ 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 QTN (see the details

in the Appendix A.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, which is a valid probability measure (by default, we set $\lambda = 0.01$), 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 $||\vec{\gamma}|| = s$ to compute C_s^* and, in the process, records the "best" posterior model with the increased model size.

Additionally, the adaptive DAP only extensively explores the model partitions with relatively small sizes. 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 (Appendix A.2). 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 A.2). Based on this relationship, the stopping rule for explicit exploration is determined, and we estimate ϵ by

$$\epsilon = \sum_{s=t+1}^{p} R_s^* \text{ with } R_{s+1}^* = \frac{p-s}{s+1}\,\omega\, R_s^* \text{ for } s = t+1, ..., p, \tag{2.5}$$

where t is the stopping point of the extensive exploration, $R_t^* = C_t^*$, and $\omega = \frac{1}{p}\sum_{i=1}^{p} \exp\left(\alpha_0 + \sum_{l=1}^{q} \alpha_l d_{il}\right)$ represents the average prior odds ratio across SNPs. 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 the details in Appendix A.2). To ensure a high accuracy for the approximation, we also build in an *optional* criterion on top of the stopping rule

by monitoring the convergence of the partial sum $S_k = \sum_i^k C_i^*$ and enforcing the exploration until

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

or, equivalently $\frac{C_t^*}{\sum_i^{t-1} C_i^*} < 10^\kappa - 1$. By default, we set $\kappa = 0.01$. This additional criterion only makes a difference for the partitions whose model sizes barely exceed the estimated size of the saturated models: instead of using the combinatorial estimate of the corresponding C_s^*, it enforces additional DAP explorations for more accurate evaluations.

Finally, it should be recognized that the built-in tuning parameters (λ, κ) enable great flexibility to run the adaptive DAP. As both $\lambda \to 0$ and $\kappa \to 0$, the adaptive DAP enumerates *all* models and becomes an *exact* calculation with no loss of precision, whereas when λ is very large, the behavior of the DAP algorithm becomes very similar to the commonly applied step-wise conditional analysis that has very high computational efficiency. In practice, we attempt to strike a good balance between the precision and efficiency.

DAP-K Algorithm

Instead of adaptively selecting a subset of high-priority SNPs from all the model size partitions, 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 utilized 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 A.3). The DAP-1 algorithm provides the advantage of considerably faster computation, even when compared with 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 provide the derivation and other technical details for the DAP-K algorithm in the Appendix A.3.

Applying DAP in Inference

We use both variants of the DAP algorithms in our inference procedure. Specifically, we propose applying the DAP-1 algorithm in the EM algorithm for enrichment analysis and the adaptive DAP for multi-SNP fine-mapping at the last stage.

The performance of the enrichment analysis mostly relies on the *average* accuracy of the PIP estimates. We show, both theoretically (Appendix E) and numerically (Figure 2.4), that the DAP-1 algorithm provides on average precise estimates suitable for enrichment analysis. Most importantly, the DAP-1 algorithm exhibits the best computational efficiency among the appropriate alternatives (e.g., adaptive DAP, MCMC).

For the multi-SNP analysis in the final fine-mapping stage, we strongly recommend applying the adaptive DAP algorithm. Although the DAP-1 algorithm only yields inferior results for a small proportion of the loci that harbor multiple QTNs, we argue that identifying multiple independent association signals from those loci is of particular importance for the overall analysis. To achieve better accuracy for all

loci, the adaptive DAP seems a logical choice for multi-SNP fine-mapping analysis.

2.2.4 Application to GWAS

In practice, the DAP works well for small genomic regions harboring a handful of QTNs. This is typically the case in molecular QTL mapping, where candidate loci usually span no more than 2 Mb. When there are more QTNs (e.g., > 5) in a locus, the adaptive DAP exploration with high precision may become time consuming because the size of the candidate set Ω grows exponentially fast with the increasing number of independent signals. Nevertheless, in applications of GWAS, we essentially consider a single locus that spans the whole genome, and for a single trait, the number of independent association signals may range from hundreds to thousands.

To apply the DAP to GWAS (or molecular QTL mapping with considerably larger candidate loci), we propose an additional approximation that factorizes $\Pr(\vec{\boldsymbol{\gamma}}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}})$ (where locus l spans a much larger genomic region) into

$$\Pr(\vec{\boldsymbol{\gamma}}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}) \approx \prod_{k=1}^{K} \Pr(\vec{\boldsymbol{\gamma}}_{[k]} \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}), \tag{2.6}$$

where $\{\vec{\boldsymbol{\gamma}}_{[k]} : k = 1, ..., K\}$ represents a partition of $\vec{\boldsymbol{\gamma}}_l$ by sets of non-overlapping LD blocks. This factorization is based on previous theoretical results *Wen and Stephens* (2010), *Wen* (2014). Recently, Berisa and Pickrell *Berisa and Pickrell* (2016) provided a working recipe to segment the full genome based on the population-specific LD structures. Based on these results, we provide mathematical arguments to justify the factorization (Appendix A.4). Briefly, applying the analytic approximation of the Bayes factors *Wen* (2014), it can be shown that

$$\mathrm{BF}(\vec{\boldsymbol{\gamma}}) \approx \prod_{k=1}^{K} \mathrm{BF}(\vec{\boldsymbol{\gamma}}_{[k]}).$$

This result, along with the fact that our priors are independent across SNPs, naturally leads to the approximate factorization of the posterior probability. As an important consequence, the factorization (2.6) suggests that the DAP can be applied to each LD block independently.

2.3 Results

First, we perform a series of simulation studies to examine the accuracy and efficiency of the proposed DAP algorithms in our inference procedure. We then apply the proposed approach to analyze two large-scale eQTL data sets.

2.3.1 Simulation Studies

Enrichment Analysis with DAP

The integration of DAP into the EM algorithm enables the efficient estimation of enrichment parameters using large-scale QTL data sets. To investigate the performance of the enrichment analysis, we simulate a modest-scale eQTL data set to mimic the genome-wide investigation of *cis*-eQTLs. Specifically in each simulation, we select a subset of 1,500 random genes from the GEUVADIS data *Lappalainen et al.* (2013). For each gene, the real genotypes of 50 *cis*-SNPs from 343 European individuals are used in the simulation. We annotate 20% of the SNPs with a binary feature. For each SNP, we determine its binary association status by performing a Bernoulli trial with the success rate $p = \frac{\exp(-4+\alpha_1 d)}{1+\exp(-4+\alpha_1 d)}$. Given the QTNs, we then simulate the expression levels according to a multiple linear regression model with residual error variance set to 1. More specifically, the genetic effect of each QTN is drawn from an independent normal distribution $\mathrm{N}(0, 0.6^2)$. As a result, the simulated data sets resemble the practically observed *cis*-eQTL data (Figure 2.1). We vary the α_1 values from 0.00 to 1.00, and we generate 100 data sets for each α_1 value.

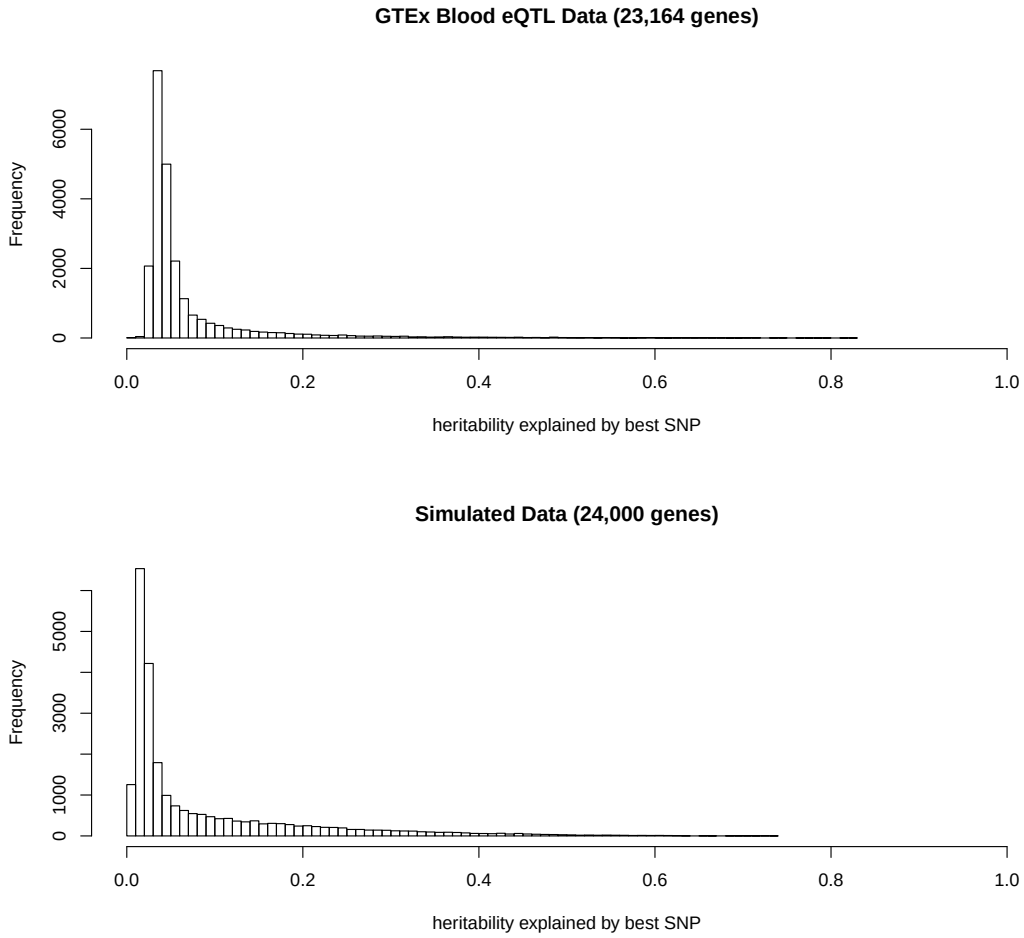

Figure 2.1: Comparison of simulated data set with the actual GTEx whole blood *cis*-eQTL data. For each gene in each data set, we find the best associated SNP based on single-SNP testing, and compute the heritability explained by the best SNP using a simple linear regression model. The histograms show the distribution of the heritability across all genes. The similarity of the two histograms indicates that the simulated data sets closely resemble the real observed *cis*-eQTL data.

We analyze the simulated data sets using two different implementations of the EM algorithm with the E-step approximated by the DAP-1 and the adaptive DAP. For evaluation, we also estimate α_1 by fitting a logistic regression model using the true association status of each SNP. This analysis represents a theoretical best-case scenario, and its results should be regarded as the bound of the most optimal outcome from any analysis that infers the latent association status ($\mathbf{\Gamma}$) from observed data.

Figure 2.2 shows that the estimates from the adaptive DAP and DAP-1 are both

seemingly unbiased. As expected, the variability of the point estimates from both DAP implementations is higher than that from the best-case method because of the uncertainty in determining the true association status of each SNP. The estimates of the 95% confidence intervals from the individual simulations also confirm this finding (Figure 2.3). Although the adaptive DAP seemingly generates more accurate estimates on average, we conclude that the numerical performance of DAP-1 is very comparable. Importantly, DAP-1 provides superior computational efficiency: the average running time for the DAP-1-embedded EM algorithm (with 10 parallel threads in the E-step) is 65.05 seconds; in comparison, the adaptive DAP-embedded EM runs for 387.30 seconds on average (which is a combination of slightly longer iterations and longer running times per iteration).

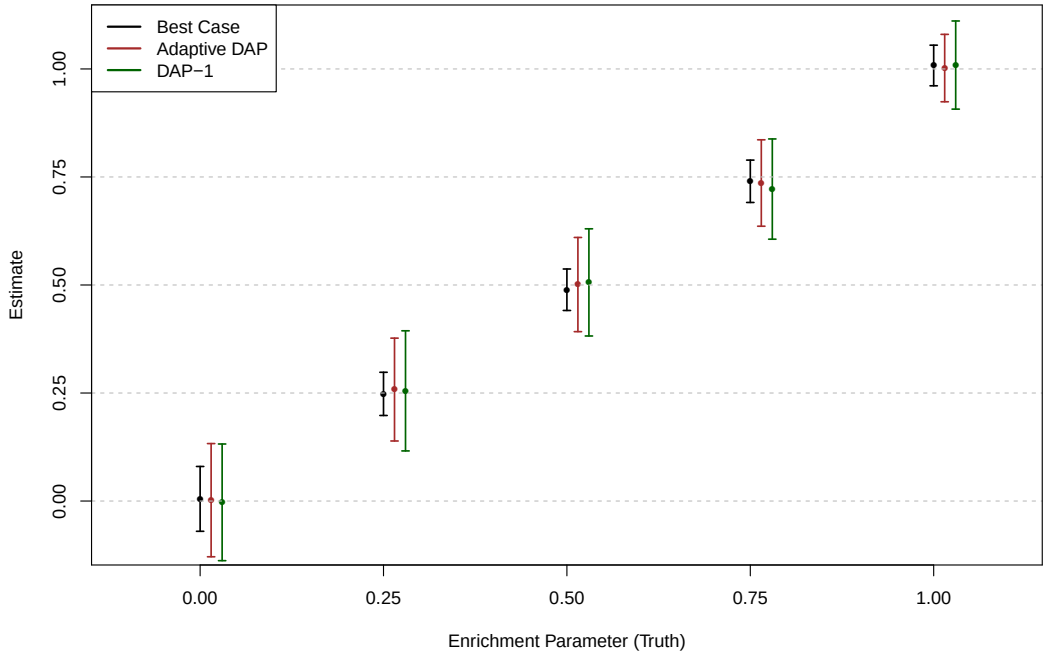

Figure 2.2: Point estimates of the enrichment parameter produced using various analysis methods in different simulation settings. The point estimate of the $\alpha_1 \pm$ standard error (obtained from 100 simulated data sets) for each method is plotted for each simulation setting. The "best case" method uses the true association status and represents the optimal performance for any enrichment analysis method. Both the adaptive DAP and DAP-1 methods yield unbiased estimates in all settings, although the adaptive DAP-embedded EM algorithm generates slightly smaller standard errors.

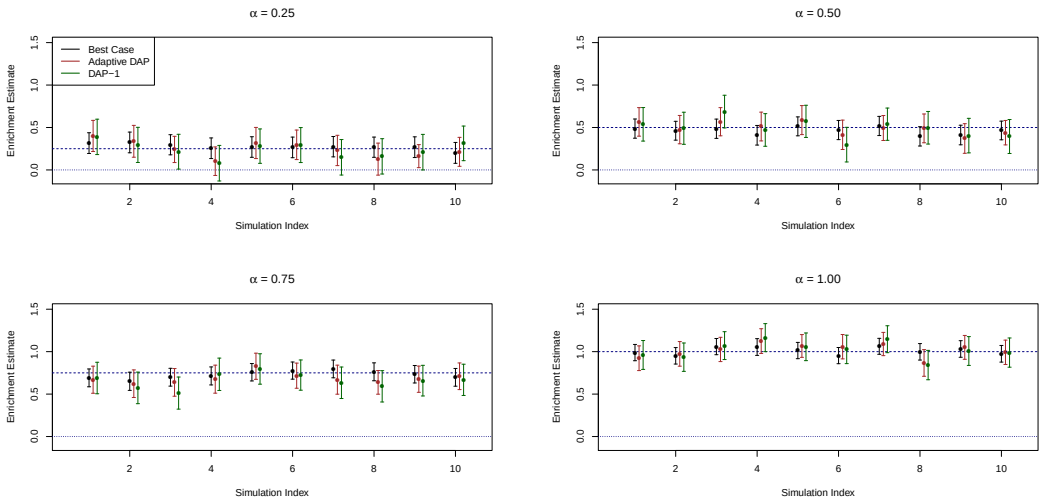

Figure 2.3: Comparison of individual estimates of the enrichment parameter and their uncertainty quantification. Each panel represents a different simulation setting. We plot the point estimates of α_1 along with their 95% confidence intervals for each method using 10 randomly selected simulated data sets. In all settings, all the methods compared ("best case", EM with adaptive DAP and EM with DAP-1) show the desired coverage probability. The figure also highlights the considerable uncertainty in enrichment analysis.

Finally, we note that both the adaptive DAP and DAP-1 algorithms underestimate the α_0 parameter: on average, DAP-1 estimates $\hat{\alpha}_0 = -4.62$, and the adaptive DAP yields $\hat{\alpha}_0 = -4.32$ (recall that the truth is $\alpha_0 = -4.00$). This is fully expected, largely because of the limitation of the statistical power in detecting weak association signals. The practical consequence is that the empirical Bayes priors constructed for the final stage of multi-SNP fine mapping analysis are slightly conservative. However, we argue that the conservative priors generally lead to reduced false discoveries and may be welcomed in practice for fine-mapping analysis.

Accuracy of the Adaptive DAP Algorithm

In the second numerical experiment, we compare the performance of the adaptive DAP algorithm with the *exact* Bayesian computation. In particular, we are interested in evaluating the accuracy of the approximation $\Pr(\vec{\boldsymbol{\gamma}}_l \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}})$ and the induced SNP-level PIP values from the adaptive DAP algorithm. The simulation setting

mimics multi-SNP fine-mapping analysis at the final stage of our proposed inference procedure.

For the exact Bayesian computation with reasonable computational cost, we have to limit the number of candidate SNPs in a locus. Specifically, in each simulation, we randomly select genotypes of $p = 15$ neighboring *cis*-SNPs of a gene from the GEUVADIS data set. We then uniformly select 1 to 5 QTNs and generate the phenotype measure using a multiple linear regression model.

We apply both the adaptive DAP algorithm and the exact Bayesian posterior computation on a total of 1,250 simulated data sets using the identical prior specification. The exact computation evaluates all $2^{15} = 32,768$ association models for each simulated data set. We apply the adaptive DAP algorithm by varying the threshold value for 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 adaptive DAP by computing the ratio C^*/C in each simulated data set. Utilizing all SNPs of all the simulated data sets, we also calculate the root-mean-square error (RMSE) to characterize the precision of the PIP approximations. The results indicate that for stringent λ values, the DAP can indeed estimate the normalizing constant with very high accuracy (Table 2.1 and Figure 2.4), which ensures the high precision of the estimated PIPs. As the λ threshold is relaxed, the approximation of C becomes less accurate in some cases; nevertheless, we observe that the overall precision level of the approximate PIPs is still reasonably high.

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}

Table 2.1: Numerical comparison of the exact calculation and the adaptive DAP algorithm at different threshold values in the second simulation study

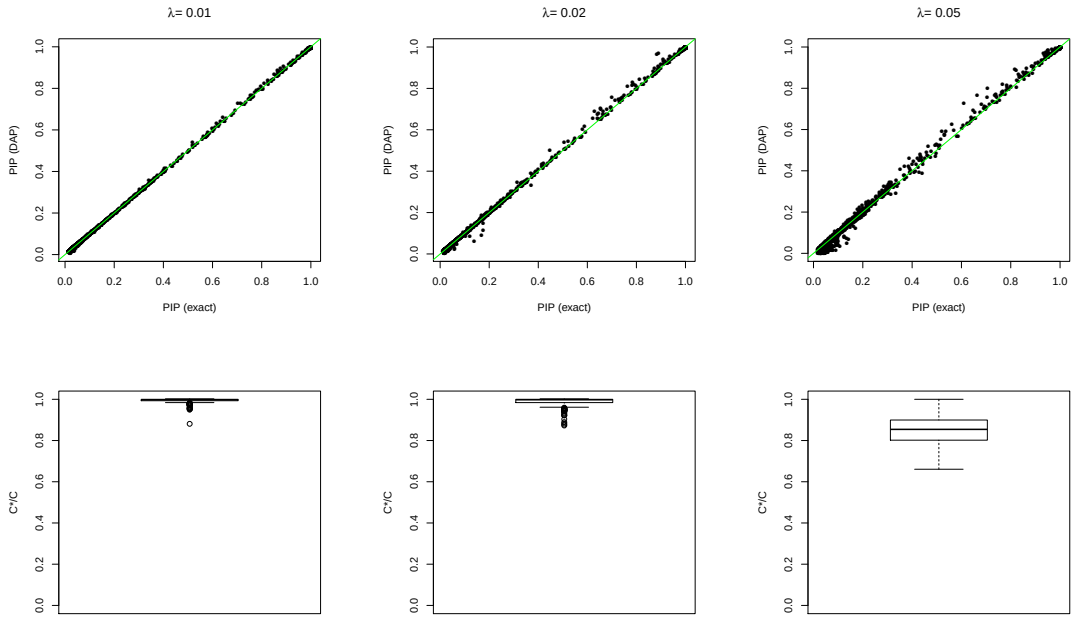

Figure 2.4: Assessment of the accuracy of the adaptive DAP algorithm at different threshold values. In the top panel, the individual PIP approximations from the DAP are compared to the exact calculations. In the bottom panel, the distribution of C^*/C is plotted. The simulation results are obtained for threshold values $\lambda = 0.01, 0.02, 0.05$ for the DAP algorithm.

Next, we examine the derived stopping rule and the analytic estimation of the approximation error. Overall, we find that the stopping rule and the error approximation work extremely well for these simulations, and we summarize the results in Figure 2.5.

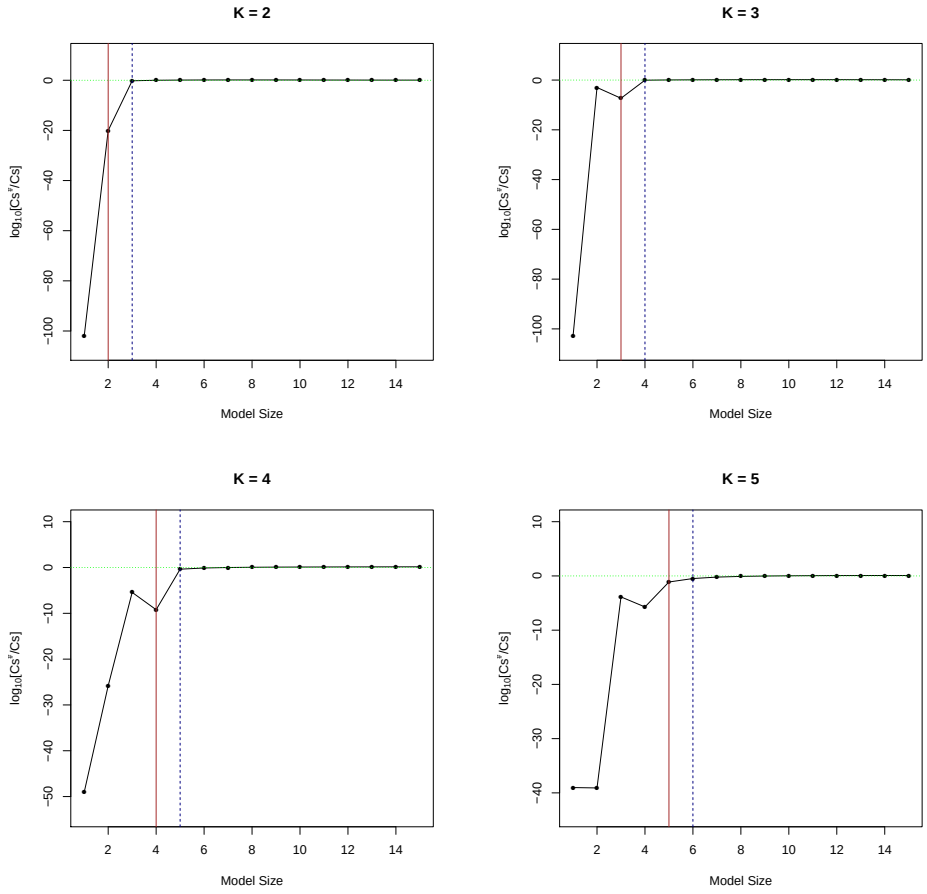

Figure 2.5: Examination of the recursive approximation of C_s by equation (A.2.4) in the simulated data sets. Each panel represents a simulated data set containing K true QTLs. The ratio of the estimated value $C_s^{\#}$ (computed using the true value of C_{s-1}) over the true value C_s is plotted on a 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 at which the adaptive DAP halts explicit exploration. As the model size s exceeds K, the estimation by $C_s^{\#}$ becomes very accurate in all settings.

Using the simulated data set, we also benchmark the average computational time for each simulation/analysis setting and present the results in Table 2.2. 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 QTNs. The DAP algorithm represents a much reduced computational time compared to the exact calculation. The general trend of the DAP running time is also clear (albeit a few small deviations): with an increasing number of true QTNs, the running

time increases, and with more relaxed λ values, the running time decreases.

	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)		

Table 2.2: Benchmark of the average computational time required for the DAP and exact computation. The running time is measured in seconds by the 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 parentheses, the "user" time is reported, which objectively reflects the actual computational cost, i.e., this measurement is not reduced by the parallelization.

Power Comparison of the Multi-SNP Analysis Algorithms

In the final simulation study, we compare the performance of the adaptive DAP with other existing algorithms in identifying multiple association signals. Specifically, we directly use the simulated multiple-population eQTL data sets from *Wen et al.* (2015b), where a genomic locus consisting of 100 relatively independent LD blocks (with 25 neighboring SNPs per block) is artificially assembled using real genotype data from the GEUVADIS project and 1 to 4 QTNs are randomly assigned to different LD blocks per simulation.

In *Wen et al.* (2015b), we compared three competing approaches, i) a single SNP analysis method, ii) a conditional analysis method, and iii) a multi-SNP analysis method based on an MCMC algorithm, regarding their abilities to correctly identify the QTN-harboring LD blocks. We run the adaptive DAP algorithm on the simulated data sets and compare the results with the three existing methods. Our results indicate that the adaptive DAP algorithm presents a significant improvement in performance (Figure 2.6) and a remarkable reduction in computational time com-

pared with the MCMC algorithm (Table 2.3), and both approaches outperform the single SNP analysis and conditional analysis approaches. In addition, Figure 2.6 also shows that with prolonged sampling steps, the MCMC outputs seemingly "converge" to the DAP results. We also run a fast version of the adaptive DAP algorithm with tuning parameter $\lambda = 0.05$ (Figure 2.7), and the results indicate that the decrease in performance from the default setting ($\lambda = 0.01$) is minimum.

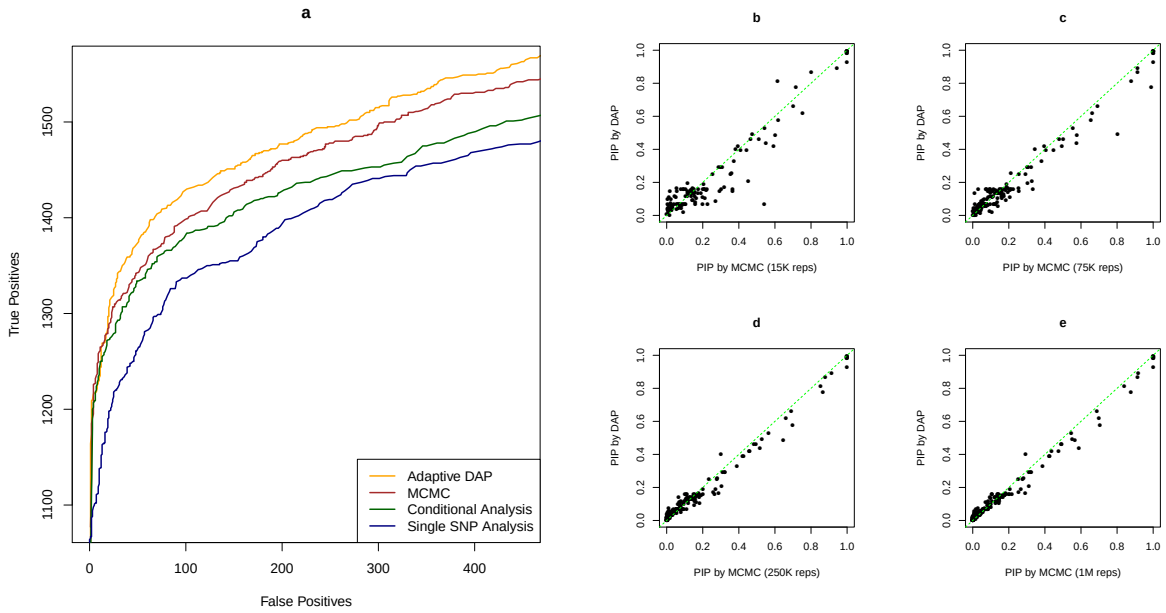

Figure 2.6: Comparison of DAP and MCMC algorithms in simulation study III. **(a)** Performance comparisons for multi-SNP QTL mapping. We apply different analytical approaches to a simulated data set reported in *Wen et al.* (2015b) to evaluate their abilities to identify multiple independent LD blocks 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.* (2015b), 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 selected 10 simulated data sets and ran 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.

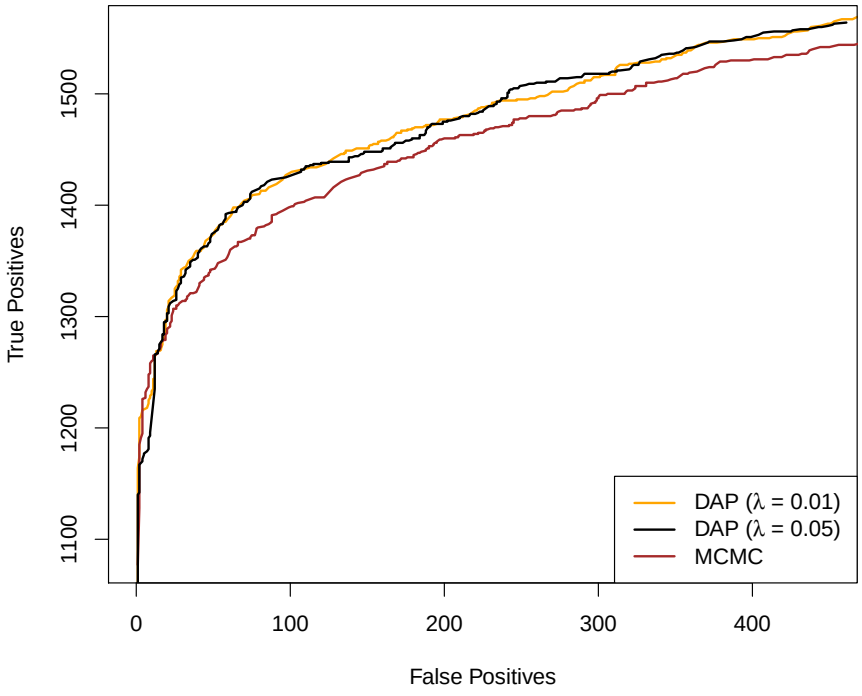

Figure 2.7: Additional comparisons for multi-SNP QTL mapping. We show the additional simulation results by running the adaptive DAP with $\lambda = 0.05$, which is most similar to the DAP outcome with the default setting ($\lambda = 0.01$) and, for the most part, still outperforms the MCMC algorithm.

	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	–

Table 2.3: Average running time and PIP comparison using MCMC runs with varying sampling steps in simulation study III. The actual running time reported from the UNIX "time" command is shown for each experiment. The DAP algorithm runs with 10 parallel threads, and the average user time (i.e., approximate running time without parallelization) is 1 minute and 8.66 seconds.

2.3.2 Re-analysis of the GEUVADIS Data

We re-analyze the cross-population eQTL data set generated from the GEUVADIS project (Web Resources) using the proposed 3-stage inference procedure. In this re-analysis, we focus on examining two types of genomic annotations that are known

to impact the enrichment of eQTNs: the SNP distance to the transcription start site (TSS) of the target gene and annotations assessing the ability of a point mutation to disrupt transcription factor (TF) binding. Following *Wen et al.* (2015b), we group all SNPs within 100 kb of a gene into 1 kb non-overlapping bins according to their distances from the TSS and use the label of the corresponding bin for each SNP to represent its distance to TSS (DTSS) as a categorical variable. In addition, a SNP is classified as a *binding SNP* if it is computationally predicted to strongly disrupt TF binding by the CENTIPEDE model using the ENCODE DNaseI data *Pique-Regi et al.* (2011), *Moyerbrailean et al.* (2016) (Web Resources). If a SNP is located in a DNaseI footprint region but there is no strong evidence for disrupting TF binding, it is classified as a *footprint SNP*; otherwise, the SNP is labeled as a *baseline SNP*. Due to the computational restraint, our previous enrichment analysis reported in *Wen et al.* (2015b) was based on a single iteration of the MCMC-within-EM (or EM-MCMC) algorithm (i.e., the E-step is carried out by the MCMC algorithm), as our main goal was enrichment *testing*. Although the evidence is sufficiently strong for testing purposes, the enrichment parameters were known to be severely underestimated.

We ran the complete DAP-1-embedded EM algorithm to perform the enrichment analysis. The full EM algorithm runs for 25 iterations to meet our convergence criteria, which require an increment ≤ 0.01 in the log-likelihood between two consecutive iterations (Figure S5). The complete EM run takes 21 minutes on a Linux box with a single 8-core Intel Xeon 2.13 GHz CPU. In comparison, the MCMC algorithm takes approximately 84 hours of computational time to fully process all 11,838 genes in a *single* E-step on the same computing system.

After a single iteration, the DAP-1-embedded EM algorithm yields point estimates

for the TF binding annotations that are very similar to our previous results reported in *Wen et al.* (2015b) (Table 2.4). As expected, the final estimates from the complete EM run have very high enrichment values: the binding SNPs have an estimated log odds ratio $\hat{\alpha}_1 = 0.94$, or fold change of 2.56, with the 95% CI $[0.84, 1.05]$, whereas the footprint SNPs have a much lower enrichment estimate (log odds ratio $\hat{\alpha}_1 = 0.53$ or fold change of 1.70, with the 95% CI $[0.40, 0.67]$). Note that the two confidence intervals are non-overlapping. In comparison, our previously reported estimates of the corresponding enrichment parameters are 0.40 (95% CI $[0.32, 0.49]$) and 0.14 (95% CI $[0.04, 0.24]$) for binding and footprint SNPs, respectively.

	Footprint SNPs		Binding Variants	
Method	α	95% C.I.	α	95% C.I.
EM-MCMC	0.14	$(0.04, 0.24)$	0.39	$(0.32, 0.49)$
EM-DAP1	0.12	$(-0.01, 0.25)$	0.41	$(0.30, 0.51)$

Table 2.4: Comparison of enrichment estimates by EM-DAP1 and EM-MCMC after a single iteration in analysis of GEUVADIS data. The binding SNPs refer to the genetic variants that are computationally predicted to disrupt TF binding, and the footprint SNPs are those simply located in the DNaseI footprint region but not predicted to affect TF binding. The enrichment estimates from both methods are very similar. The MCMC algorithm accounts for multiple independent association signals and yields slightly tighter confidence intervals, as expected. However, the EM-DAP1 is much more computationally efficient: it runs almost one thousand times faster than the EM-MCMC algorithm.

Next, we repeat the multi-SNP fine-mapping analysis using the adaptive DAP algorithm and the new set of the empirical Bayes priors obtained from the enrichment analysis. For most genes, the results (i.e., the number of independent signals for each gene) are qualitatively unchanged compared to the previous MCMC results. Nevertheless, we find that fine-mapping with the adaptive DAP is much more efficient, and the annotated SNPs, especially the binding SNPs, are further prioritized in the new fine-mapping results (Figure 2.9).

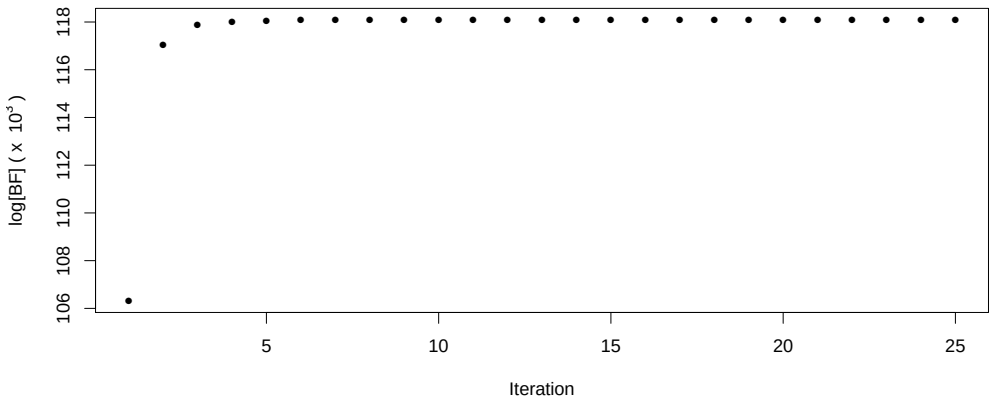

Figure 2.8: Traceplots of the marginal likelihood (in Bayes factor on the log scale) during the DAP-1-embedded EM run for analyzing the GEUVADIS data.

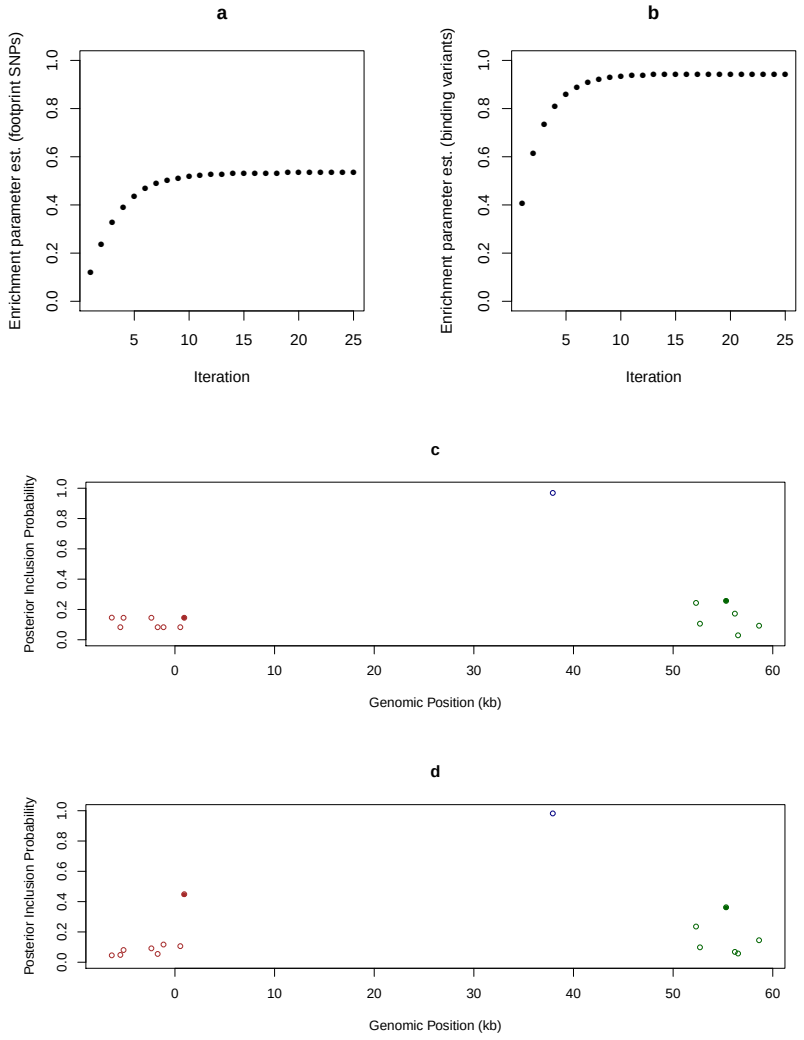

Figure 2.9: Output from the re-analysis of GEUVADIS data. **(a) - (b)** Traceplots of estimates of the enrichment parameters for binding variants and footprint SNPs during the DAP-1-embedded EM iterations for analyzing the GEUVADIS data. Both estimates are stabilized after approximately 8 iterations. **(c) - (d)** Comparison of multi-SNP *cis*-eQTL mapping with and without incorporating functional annotations. We plot the multi-SNP QTL mapping results of *LY86* [MIM 605241] 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 relative to the transcription start site. SNPs in high LD are plotted with the same color, and the filled circles indicate 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.

2.3.3 Analysis of the GTEx Data

We analyze the *cis*-eQTL data from the GTEx project (Web Resources). One of the most unique advantages of the GTEx data is that they enable the study of the commonality and specificity of the eQTLs in multiple tissues. Taking advantage of the high computational efficiency of the EM-DAP1 algorithm, we perform the enrichment analysis of the TF binding annotations, derived from the ENCODE data and the CENTIPEDE model, in eQTLs across 44 human tissues while controlling for the SNP distance to TSS. More specifically, for each gene, we consider a 2 Mb *cis* region centered at the transcription start site. For each tissue, we perform the enrichment analysis using two sets of TF binding annotations, one derived from the ENCODE LCL cell-line and the other from the ENCODE liver-related HepG2 cell-line *Moyerbrailean et al.* (2016) (Web Resources). This exercise aims to assess the impact of the cell type-specific annotations on the proposed integrative analysis.

Our results indicate that the binding variants are significantly enriched in eQTLs in all tissues regardless of the origin of the annotations. Furthermore, the point estimates of enrichment levels for binding variants are consistently higher than those for footprint SNPs, except in one occasion (small intestine tissue with LCL-derived annotations) where the two estimates are indistinguishable. Importantly, we find that the enrichment estimates in specific tissues are quantitatively correlated with the origins of the annotations. Figure 2.10 shows the results of the enrichment level estimates ($\hat{\alpha}_1$) of the binding variants in each tissue using the LCL- and HepG2-derived TF binding annotations. Most interestingly, the LCL-derived annotations yield the highest enrichment estimates in LCLs and whole blood from the GTEx data sets, whereas the liver-related HepG2-derived annotations obtain the highest enrichment estimate in the GTEx liver tissue. Overall, our results suggest that TF

binding annotations derived from different tissues must have substantial overlaps; nevertheless, the annotations from the relevant tissues may provide better functional interpretations for expression-altering causal SNPs in a specific tissue.

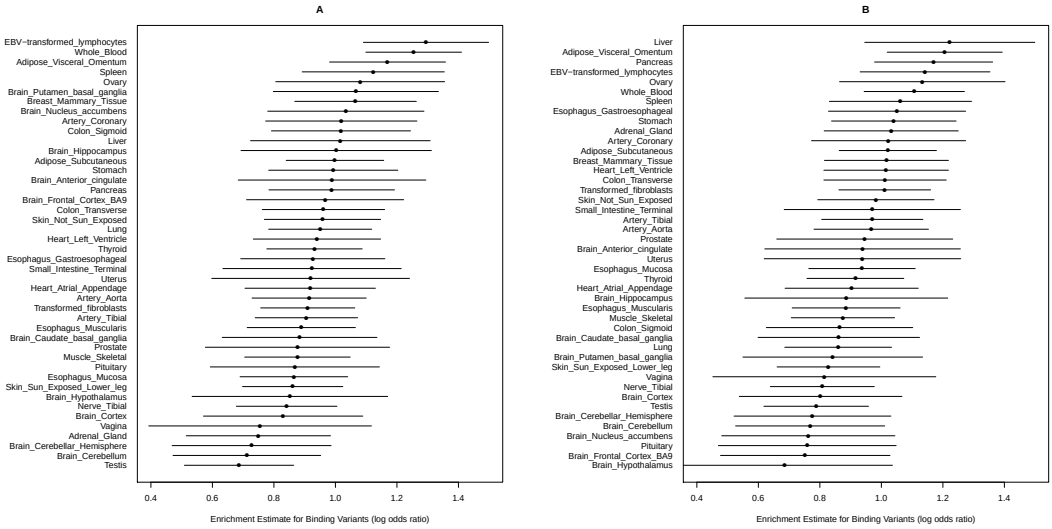

Figure 2.10: Enrichment estimates for binding variants in GTEx tissues. The estimates in panel A are based on the annotations derived from the DNaseI data of the ENCODE LCLs, whereas the estimates in panel B are based on annotations derived from the ENCODE liver-related HepG2 DNaseI data. In each panel, we plot the point estimate of the enrichment parameter and its 95% confidence interval in each tissue. The tissues are ranked in descending order according to the magnitude of the point estimates. All estimates are obtained controlling for the SNP distance from TSS. All estimates are significantly far from 0 (at the 5% level). Interestingly, when the tissue and origin of the annotations match, the point estimates for enrichment are the highest.

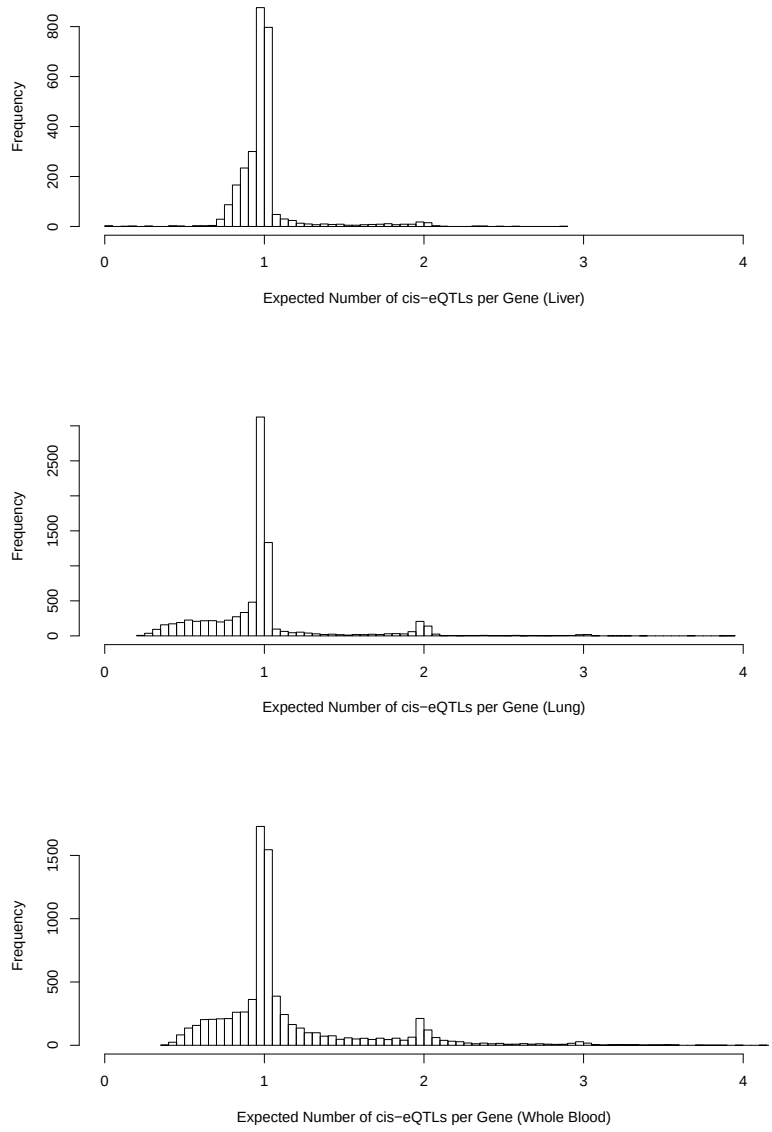

Figure 2.11: Posterior expected number of *cis*-eQTL signals per eGene in GTEx liver, lung and whole blood tissues. The top, middle and bottom panels display the histogram of the posterior expected number of *cis*-eQTLs from all the eGenes in the liver, lung and blood tissues, respectively. For most genes, we can only identify a single association signal. However, for a non-trivial number of eGenes, multiple independent association signals can be confidently identified by the adaptive DAP algorithm. The sample size is seemingly an important factor related to the ability to identify multiple independent signals in a *cis* region.

We then proceed to identify genes that harbor QTNs (i.e., eGenes) using a Bayesian FDR control procedure that we recently developed *Wen* (2016). Subsequently, we perform multi-SNP fine-mapping analysis for the identified eGenes in-

corporating the enrichment estimates using the adaptive DAP algorithm. We present the analysis results for the liver (sample size 97), lung (sample size 278) and whole blood (sample size 338). There are 2,788, 8,605 and 7,937 eGenes that are identified from the lung, liver and whole blood, respectively. We suspect that the number of differences in eGenes discovery is largely attributed to the sample sizes but is also correlated with the levels of experimental noise in measuring the gene expression in each tissue. For each fine-mapped eGene l in each tissue, we compute the posterior expected number of independent signals using $\sum_{i=1}^{p} \Pr(\gamma_{l_i} \mid \vec{\boldsymbol{y}}_l.\boldsymbol{G}_l, \hat{\vec{\boldsymbol{\alpha}}})$ and plot the histogram for each tissue in Figure 2.11. In all three tissues, we identify single eQTL signals for the vast majority of eGenes. Nonetheless, for a non-trivial number of genes, we are able to confidently identify multiple independent signals. Comparing the fine-mapping results among the three tissues, we find that the ability to identify additional independent signals is also seemingly correlated with the sample sizes.

We further examine some known individual genes to validate our integrative analysis results. In particular, we examine *SORT1* [MIM 602458], whose function is related to plasma low-density lipoprotein cholesterol (LDL-C [MIM 613589]) metabolism through modulation of hepatic VLDL secretion. Through GWAS meta-analysis and extensive functional analysis *Musunuru et al.* (2010), a single SNP, rs12740374, is identified to cause variations in LDL-C. More specifically, the major allele disrupts the binding site of C/EBP transcription factors in human hepatocytes. Our integrative fine-mapping analysis using the GTEx liver data yields a Bayesian 95% credible set, narrowed down to only two potential causal eQTNs for *SORT1*: rs12740374 (PIP = 0.473) ranks second very closely only to SNP rs7528419 (PIP = 0.526). Moreover, the direction of the genetic effect for rs7528419 fits the description provided in *Musunuru et al.* (2010). The two SNPs in the credible set are in high LD

($r^2 > 0.95$), except that the genotypes of rs12740374 in the GTEx samples are *not* directly genotyped but imputed. Upon further investigation, we find that the binding site reported by *Musunuru et al.* (2010) is not captured by the ENCODE DNaseI experiments in HepG2, and hence, rs12740374 is not correctly annotated. We then include the annotation of rs12740374 as a binding SNP based on the functional study of *Musunuru et al.* (2010) and re-run the fine-mapping analysis using the adaptive DAP. We find that rs12740374 yields the highest PIP value (PIP = 0.752) among all the candidate SNPs (the PIP for rs7528419 drops to 0.247). The lesson learned here is that the completion of the genomic annotations may have a profound impact on the integrative analysis, and efforts should be made to generate a more comprehensive set of genomic annotations by both accumulating new experimental data and integrating them with all the existing data.

2.4 Discussion

The proposed EM-DAP1 algorithm provides an efficient and flexible framework to perform enrichment analysis with respect to genomic annotations using genetic association data – there is no restriction on the types of annotations (categorical or continuous) or the number of annotations that can be simultaneously investigated. Some of the commonly applied *ad-hoc* enrichment analysis methods in the same context attempt to first classify the binary latent association status $\mathbf{\Gamma}$ for all candidate SNPs based on their single SNP testing results. However, it is worth noting that the classification based on hypothesis testing typically has very stringent controls over type I errors but is much more tolerant (in practice, it may be too tolerant) and has little control over type II errors, which are a major source of the overall mis-classification errors for $\mathbf{\Gamma}$ *Wen et al.* (2015b). As a consequence, most *ad-hoc*

procedures of this type provide poor quantification of enrichment levels. Recently, probabilistic model-based enrichment analysis approaches have been proposed based on the "one QTN per locus" assumption and applied to both molecular QTL mapping and GWAS *Pickrell* (2014). A common feature of these approaches is that they treat each locus as the exchangeable/comparable unit in the analysis: in the simplest case, each locus has the common prior probability, π_1, of harboring causal QTNs. Although the DAP-1 algorithm implicitly also makes the same assumption and enjoys the benefit of fast and efficient computation using only summary statistics, it presents some significant differences/improvements compared to the aforementioned approaches. The DAP-1 algorithm, built on the proposed hierarchical model, considers each SNP as the unit of analysis. This modeling strategy leads to a straightforward EM algorithm for parameter estimation, where the target function in the M-step is convex with well-known optimization solutions. In comparison, with the parameterization including π_1, the target function in the M-step is no longer guaranteed to be convex, which can cause convergence issues in EM estimation and prevent the simultaneous investigations of many annotations (see the details in the Appendix A.3). Furthermore, π_1 parameterization essentially assumes that genetic loci consisting of many SNPs are equally likely to harbor causal QTNs as loci consisting of only a few SNPs. From the empirical evidence produced by eQTL analysis, we find that this assumption is likely false : the genes with more *cis* candidate SNPs are more likely to harbor eQTNs *Wen et al.* (2015b). In summary, the proposed hierarchical model and the EM-DAP1 algorithm represent better alternatives.

The proposed Bayesian hierarchical model does not explicitly consider potential polygenic background. To evaluate the performance of the proposed enrichment analysis method under an explicit polygenic model, we modify the simulation set-

tings for enrichment analysis by imposing a small yet non-zero genetic effect on every candidate SNP. Under such setting, γ_i should be interpreted as an indicator whether the genetic effect of SNP i is significantly larger than the polygenic background. The simulation results (Figure 2.12) indicate that the estimates of the enrichment parameters are biased toward 0 in the presence of polygenic background: although the bias is negligible when the polygenic effects are small. We plan to extend our current work to fully account for polygenic background in our future work by considering a more appropriate model like the Bayesian sparse linear mixed model (BSLMM) *Zhou et al.* (2013).

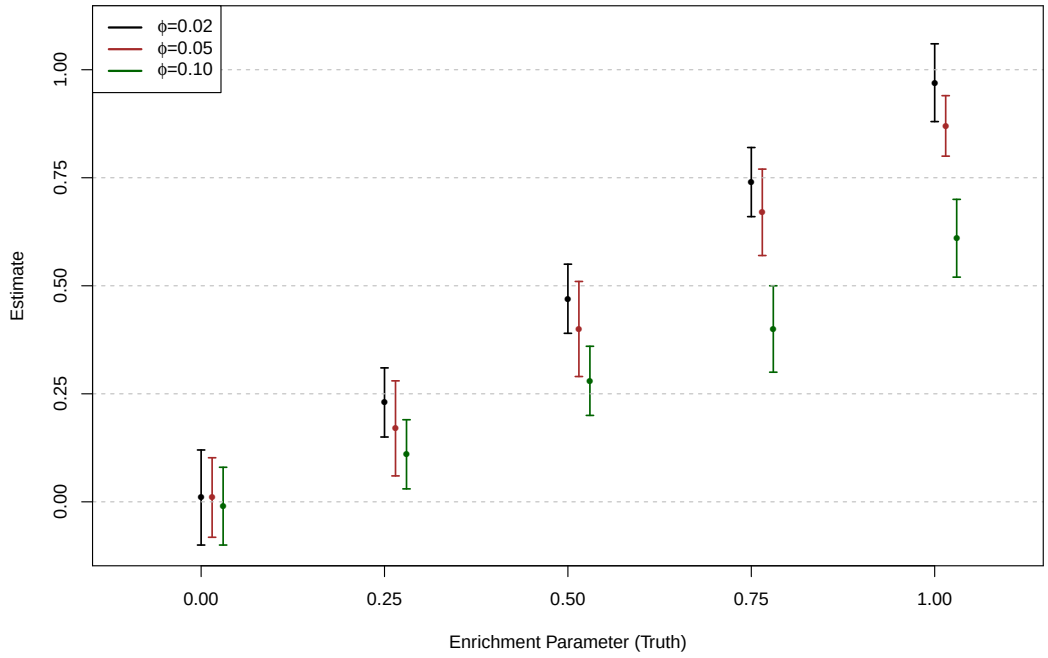

Figure 2.12: Estimates of the enrichment parameters for data simulated from polygenic models. In this experiment, the simulation scheme is mostly similar to the first simulation study described in the main text, except that in addition to the SNPs sampled to have large effects, we assign a non-zero genetic effect from an independent $N(0, \phi^2)$ distribution for all the remaining candidate SNPs. (In this case, γ_i should be interpreted as an indicator of large genetic effect.) We select $\phi = 0.02, 0.05$ and 0.1 to represent different magnitude of polygenic background. The point estimate of the $\alpha_1 \pm$ standard error (obtained from 50 simulated data sets using DAP-1-embedded EM algorithm) for each ϕ value is plotted. In all cases, the non-zero α_1 estimates are biased toward 0, however when ϕ is small ($\phi = 0.02$), the bias seems negligible.

Our analysis of multi-tissue eQTL data yields many interesting findings that are worthy of in-depth follow-up investigation. In particular, our results suggest that the cell type specificity and the completeness/accuracy of the genomic annotations may have profound impacts on the integrative association analysis in terms of different aspects as follows: the cell-type specificity of the annotations affects the global enrichment estimates and the multi-SNP analysis results of *every* subsequently fine-mapped locus, whereas mis-annotations of certain variants likely impact functional interpretations of specific loci but are not likely to alter the global enrichment estimates as long as the annotations are accurate *on average* . These findings should motivate efforts to generate a more comprehensive and accurate catalog of genomic annotations to improve the overall quality of genetic association analysis. Furthermore, it should be noted that all the annotations could have additional levels of complexity (e.g., *cis* regulatory grammar) that can be consistently analyzed within the same framework by extending our logistic prior model in a straightforward manner to allow interactions. To aid these efforts, our proposed genome-wide scale enrichment analysis has provided a principled way of assessing the tissue/cell type specificity of the genomic annotations.

2.5 Acknowledgments

We thank the GTEx consortium and the GEUVADIS RNA sequencing project for releasing valuable data in a timely fashion. This work is supported by NIH grants MH101825 (XW), HG007022(XW) and GM109215 (XW, YL, FL and RP).

2.6 Web Resources

The URLs for data presented herein are as follows:

DAP software and tutorial, `http://github.com/xqwen/dap/`

GUEVADIS data, http://www.geuvadis.org/web/geuvadis/rnaseq-project

Re-analyzed multi-SNP fine-mapping results of the GUEVADIS data, http://www-personal.umich.edu/~xwen/geuvadis/new_fm_rst/

GTEx data, http://www.gtexportal.org/home/datasets

Transcription factor binding site annotations by the extended CENTIPEDE model, http://genome.grid.wayne.edu/centisnps/

CHAPTER III

Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics

3.1 Introduction

In the past decades, genetic association analysis has become a primary analytic tool to uncover genetic risk factors in complex diseases. With the advancement of high-throughput genotyping and phenotyping technology, genome-wide association studies (GWASs) and molecular quantitative trait loci (QTL) mapping have led to discoveries of an abundance of signals through genetic association analysis. These findings have subsequently played critical roles in exploring molecular mechanisms of complex diseases and predicting risks for individual patients.

Single-SNP association testing has long been considered as the standard approach for genetic association analysis. However, the results of the single-SNP analysis are not sufficiently informative by their own and often difficult to interpret without explicit references to linkage disequilibrium (LD) patterns of candidate variants. Additionally, it has been convincingly demonstrated that single-SNP testing fundamentally lacks power in identifying multiple association signals that are close by in relatively narrow genomic regions. A simple form of multi-SNP association analysis, known as *conditional analysis*, seeks a single "best" multi-SNP association model by a step-wise forward variable selection procedure (*Yang et al.*, 2012). This approach

addresses the power issue, but a single best solution oversimplifies the intrinsic difficulty introduced by the complex LD patterns and fails to account for the uncertainty of causal associations at SNP level.

Most recently, Bayesian approaches for multi-SNP association analysis have emerged as a promising alternative. They have at least two unique advantages over the traditional frequentist methods in the practice of genetic association analysis. First, they are built upon a natural hierarchical model that enables flexible incorporation of SNP-level functional annotations through principled prior specifications. Second, they utilize probabilistic quantification to characterize the strength of association evidence at SNP level, which can fully account for the complex LD structures presented in the genotype data. The successful applications of Bayesian genetic association analysis are illustrated in a wide range of applications for GWAS and molecular QTL mapping by piMASS (*Guan and Stephens*, 2011), GUESS (*Bottolo et al.*, 2013), PAINTOR (*Kichaev et al.*, 2014), CAVIAR (*Hormozdiari et al.*, 2014), CARIVARBF (*Chen et al.*, 2015) and FINEMAP (*Benner et al.*, 2016), just to name a few. One of the significant limitations for the Bayesian approaches is the computational cost: instead of seeking a single best association model (i.e., by optimization), the Bayesian inference requires a comprehensive survey of all plausible association models (i.e., by integration). As a result, most existing Bayesian approaches do not scale well for extended genomic regions and often limited to the applications of fine-mapping analysis. Recently, we have proposed a new computational algorithm named deterministic approximation of posteriors (DAP), which is aimed to strike a balance between the commonly applied stochastic approximation algorithms (e.g., MCMC implemented in FINEMAP) and the exact computation by brute-force exhaustive search (e.g., in CAVIAR). We have shown, in,[512016 *Wen et al. Wen, Lee, Luca, and Pique-Regi*

that the DAP algorithm represents a highly efficient and accurate Bayesian inference procedure that can scale up to large-scale multi-SNP genetic association analyses in both GWAS and molecular QTL mapping.

Built upon the DAP algorithm's high computational efficiency, this chapter addresses two outstanding issues in the Bayesian multi-SNP genetic association analysis. First, we propose a novel false discovery rate (FDR) control procedure utilizing the posterior probabilities generated by our Bayesian approach. Rigorous control of type I error rate has always been an emphasis in genetic association analysis. Nevertheless, there is a lack of formal statistical procedures that can effectively control potential false discoveries in the multi-SNP analysis. Most theoretical results (*Barber et al.*, 2015, *Brzyski et al.*, 2017) on type I error control in the context of high-dimensional variable selection do not directly applicable to genetic association analysis because of the complex LD structures in the genetic association analysis. Our approach aims to fill this gap by proposing an intuitive hierarchical representation of association signals and adopting a principled Bayesian FDR control paradigm. Second, we discuss performing Bayesian multi-SNP association analysis based on summary statistics. Many authors have proposed association analysis algorithms that can work explicitly with summary-level data from single-SNP testing (*Yang et al.*, 2012, *Kichaev et al.*, 2014, *Hormozdiari et al.*, 2014, *Chen et al.*, 2015, *Benner et al.*, 2016, *Zhu et al.*, 2017). This has become an essential feature due to the nature of genetic data sharing for privacy protection. Our work on this topic focuses on understanding the analytic relationship of inference results based on individual-level data versus summary data. For example, we examine the following questions: do the two types of procedures (i.e., summary statistics vs. individual-level data) yield the same results? If not, is the inference based on summary statistics valid? Based

on the answers to these questions, we attempt to identify a set of sufficient summary statistics that can lead to *identical* inference results as individual-level data, especially in Bayesian multi-SNP analysis.

The proposed novel computational approaches for multi-SNP genetic association analysis are implemented in the software package DAP-G, which is freely available at `https://github.com/xqwen/dap/`.

3.2 Method

3.2.1 Background, model and notation

In this chapter, we focus on the problem of identifying potentially multiple genetic association signals using the following multiple linear regression model,

$$\vec{\boldsymbol{y}} = \sum_{i=1}^{p} \beta_i \vec{\boldsymbol{g}}_i + \vec{\boldsymbol{e}},\ \vec{\boldsymbol{e}} \sim \mathrm{N}(\mathbf{0}, \tau^{-1}\boldsymbol{I}). \tag{3.1}$$

In practice, we assume that linear model (3.1) is obtained after regressing out a set of controlled covariates, including the intercept, from both the outcome vector and each genotype vector of candidate genetic variants. As a result, both $\vec{\boldsymbol{y}}$ and all the $\vec{\boldsymbol{g}}_i$'s have mean 0. Furthermore, we denote the $n \times p$ design matrix $\boldsymbol{G} := [\,\vec{\boldsymbol{g}}_1\ \vec{\boldsymbol{g}}_2\ \cdots\ \vec{\boldsymbol{g}}_p\,]$, which contains genotype data of all p candidate SNPs.

The point of interest for statistical inference is to identify the genetic variants that have non-zero effects on the quantitative trait. To this end, we explicitly define a latent binary indicator for each candidate predictor i by

$$\gamma_i := \mathbf{1}(\beta_i \neq 0), \tag{3.2}$$

and $\vec{\boldsymbol{\gamma}} := \{\gamma_1, ..., \gamma_p\}$.

Based on this model, we formulate the problem of multi-SNP fine-mapping as a variable selection problem with respect to $\vec{\boldsymbol{\gamma}}$ given the observed data $(\vec{\boldsymbol{y}}, \boldsymbol{X})$. Further

details of the model are provided in Appendix B.1. Specifically, we compute the posterior probability for a given $\vec{\gamma}$ by

$$\Pr(\vec{\gamma} \mid \vec{y}, \boldsymbol{G}) = \frac{\Pr(\vec{\gamma})\mathrm{BF}(\vec{\gamma})}{\sum_{\vec{\gamma}'} \Pr(\vec{\gamma}')\mathrm{BF}(\vec{\gamma}')}, \tag{3.3}$$

where $\Pr(\vec{\gamma})$ denotes the prior probability and $\mathrm{BF}(\vec{\gamma}) = \frac{P(\vec{y}|\boldsymbol{G},\vec{\gamma})}{P(\vec{y}|\boldsymbol{G},\vec{\gamma}=\mathbf{0})}$ denotes the Bayes factor/marginal likelihood for $\vec{\gamma}$. Subsequently, the SNP-level posterior inclusion probability (PIP), which quantifies the strength of association for each SNP, can be marginalized from the posterior distribution, $\Pr(\vec{\gamma} \mid \vec{y}, \boldsymbol{G})$.

Overview of the DAP algorithm

For any given $\vec{\gamma}$ value, the prior and the Bayes factor can be analytically computed. The computational difficulty lies in the evaluation of the normalizing constant, i.e., $\sum_{\vec{\gamma}'} \Pr(\vec{\gamma}')\mathrm{BF}(\vec{\gamma}')$: it is infeasible to enumerate all possible values of $\vec{\gamma}$ for a large number of candidate SNPs. The algorithm of deterministic approximation of posteriors (DAP) is designed to tackle this problem directly and can efficiently operate on a genomic region containing tens of thousands of candidate SNPs. (For larger regions or genome-wide analysis, it requires to segment the genome into LD blocks for separate processing.) The fundamental idea behind the DAP algorithm is based on the fact that noteworthy genetic association signals are typically *sparse* for any given genomic locus. Thus, only a very small number of candidate models (namely, the plausible models) make a substantial contribution to the normalizing constant. The DAP algorithm utilizes an efficient deterministic search strategy to identify the plausible models and approximates the normalizing constants based on the proven statistical principle known as *sure independence screening* (SIS,[152008*Fan and Lv*]). The approximation error to the true normalizing constant is also estimated in the search process, which plays a role in adjusting the estimated normalizing constant. In com-

parison, the commonly applied Markov Chain Monte Carlo (MCMC) algorithm is also designed to explore the plausible models but in a stochastic fashion. Because of the sampling space is enormous and consists of discrete models, it is unrealistic to expect that the MCMC algorithm reaches convergence with a limited computing resource. As a result, we find that the DAP algorithm often outperforms conventional MCMC algorithms in the setting of genetic association analysis. The new DAP-G algorithm is built upon the existing DAP algorithm and enjoys the same computational efficiency in the posterior inference of multi-SNP genetic association analysis.

3.2.2 False discovery rate control for genetic association signals

Hierarchical representation of genetic association discoveries

Quantifying strength and uncertainty of genetic association signals is a long-standing problem in statistical genetics. The intrinsic difficulty lies in the fact that, with few exceptions, causal genetic associations may not be statistically identifiable at individual SNP level; Instead, each association signal is typically represented by a group of genetic variants whose genotypes are highly correlated. We argue that quantification and representation of a potential association signal should be dealt in a natural hierarchy, in which the following issues can be addressed:

1. the (un)certainty of the existence of an independent association signal;
2. the SNPs that are causally responsible for the association signal and their individual uncertainties

To demonstrate, we consider a hypothetical example from[522017]*Wen et al.Wen, Pique-Regi, and Luca* : one of the two perfectly linked SNPs is causally associated with the complex trait of interest, and both SNPs are uncorrelated with the remaining candidate SNPs. In an ideal analysis, a precise characterization of the genetic association discovery

should reflect that i) there is overwhelming evidence for the existence of an association signal; ii) there is a maximum degree of uncertainty to distinguish the causal variant between the two linked SNPs. We argue that the inference result of an ideal Bayesian analysis, which assigns PIP = 0.5 to each SNP, precisely encodes this information. The sum of the PIPs (= 1) indicates the sure existence of an association signal. Nevertheless, the two SNPs are equally likely to be the causal variant and not distinguishable solely based on the association data. (Further, if there exists additional information on the functional annotations of each SNP, it can be incorporated into the prior specifications that make the two SNPs distinguishable and potentially break the tie for the PIPs.)

This simple example illustrates the superiority of the probabilistic representation by Bayesian inference, which can carry comprehensive information from genetic association analysis. Nevertheless, we note that although almost all Bayesian multi-SNP analysis approaches generate SNP-level PIPs, there is no principled approach to summarizing the probabilistic evidence at the signal level, to the best of our knowledge. In a practical setting, it can be challenging to identify SNPs that are responsible for a single association signal (we will call the collection of such SNPs a signal cluster, henceforth). Identifying signal clusters require simultaneously examining both the overall evidence from multiple "similar" association models (e.g., when SNPs from the same signal cluster co-exist in an association model, the overall strength of evidence diminishes) and the pattern of LD.

The probabilistic quantification of association signals at both signal and SNP levels has multiple benefits. First, it allows rigorous control of the false discovery rate (FDR) at the signal cluster level (even though it can be challenging to pinpoint the causal association at the SNP level). Second, it allows constructions of Bayesian

credible sets for suitable signal clusters, which is proven particularly attractive in genetic association analysis (*Maller et al.*, 2012). Such credible sets provide a refined list of candidate SNPs for the underlying causal variants and can be critically valuable for the design of downstream molecular validation experiments.

Identification of signal clusters

In the DAP-G algorithm, we integrate the functionality of automatic identification of signal clusters into the deterministic model search procedure.

Let $\vec{\gamma}^{(i,j)}, \vec{\gamma}^{(\bar{i},j)}$ and $\vec{\gamma}^{(i,\bar{j})}$ denote three related association models that only differ in the values of γ_i and γ_j. Specifically, both SNP i and SNP j are assumed associated in $\vec{\gamma}^{(i,j)}$, whereas only SNP i is assumed associated in $\vec{\gamma}^{(i,\bar{j})}$ and only SNP j is assumed associated in $\vec{\gamma}^{(\bar{i},j)}$. We deem that SNP i and SNP j belong to the same signal cluster if and only if

1. the genotype R^2 between SNP i and SNP j is greater than a pre-defined threshold;
2. the overall association evidence favors a single inclusion of SNP i or j, but not both, i.e.,

$$\Pr(\vec{\gamma}^{(\bar{i},j)})\,\mathrm{BF}(\vec{\gamma}^{(\bar{i},j)}) \approx \Pr(\vec{\gamma}^{(i,\bar{j})})\,\mathrm{BF}(\vec{\gamma}^{(i,\bar{j})}) \gg \Pr(\vec{\gamma}^{(i,j)})\,\mathrm{BF}(\vec{\gamma}^{(i,j)}).$$

The first condition simply requires that SNPs within the same signal cluster are in LD and we use a rather relaxed threshold, i.e., $R^2 = 0.25$, by default. In the second condition, $\Pr(\vec{\gamma}^{(\bar{i},j)})\,\mathrm{BF}(\vec{\gamma}^{(\bar{i},j)}) \approx \Pr(\vec{\gamma}^{(i,\bar{j})})\,\mathrm{BF}(\vec{\gamma}^{(i,\bar{j})})$ implies that the SNP i and SNP j makes similar contribution to the marginal likelihood with everything else being equal. However, when both SNPs within the same signal cluster co-exist in an association model, i.e., in $\vec{\gamma}^{(i,j)}$, the likelihood is expected to be saturated, and the inequality is due to the prior "penalty" for assuming an additional causal SNP

that is redundant. Essentially, this definition attempts to ensure that each signal cluster harbors precisely one independent association signal.

Based on above criteria, the DAP-G algorithm explicitly searches for redundant SNP representations of the same association signals and group them into signal clusters. When evaluating the approximate normalizing constant, each signal cluster is treated as an independent unit, and association models containing multiple SNPs from the same inferred signal cluster are explicitly avoided. We evaluate a signal-level PIP, denoted by SPIP, for each signal cluster by summing over the SNP-level PIPs from the member SNPs, i.e.,

$$\text{SPIP}_i = \sum_{j \in C_i} \Pr(\gamma_j = 1 \mid \vec{\boldsymbol{y}}, \boldsymbol{G}), \tag{3.4}$$

where the C_i denotes the set of SNPs representing the k-th signal cluster. Note that our definition of the signal cluster and the search algorithm guarantees SPIP a valid probability distribution (i.e., strictly bounded by $[0, 1]$).

Control signal-level false discovery rate

The signal-level PIPs enable a straightforward Bayesian FDR control procedure to guard against false positive findings. Specifically, the complement of SPIP is interpreted as the false discovery probability of signal cluster i and also known as the *local fdr* of the signal i, i.e.,

$$\mathit{lfdr}_i = 1 - \text{SPIP}_i. \tag{3.5}$$

The use of local fdr for multiple hypothesis testing is well established in the statistical literature (*Efron*, 2012), and its result is asymptotically concordant to the frequentist testing approach utilizing p-values (*Wen*, 2018). Briefly, the following null hypothesis

$$H_0: \text{ cluster } i \text{ does not contain an association signal,}$$

is rejected, if $lfdr_i$ is less than or equal to a pre-defined threshold t. Moreover, the threshold t is determined by the pre-defined FDR control level α, such that the average *lfdr* value from all rejected hypotheses is no greater than α. More precisely,

$$t = \arg\max_{\lambda} \left(\frac{\sum_{lfdr_i < \lambda} lfdr_i}{[\sum_i \mathbf{1}\{lfdr_i < \lambda\}] \vee 1} \leq \alpha \right). \tag{3.6}$$

FDR control has become standard statistical approach for type I error control in molecular QTL mapping, where abundant association signals can be identified with modest sample size. Some also advocate direct specification of threshold value t (*Efron*, 2012), e.g., setting $t = 0.05$, which, in this case, is more stringent/conservative than controlling the overall FDR at 5%.

For a signal whose local fdr $\leq t$, it is straightforward to construct a $(1-t)\%$ Bayesian credible set by selecting a minimum subset of SNPs, such that their cumulative SNP-level PIPs reaches $1-t$. The Bayesian credible intervals have been widely applied in GWAS since its introduction by *Maller et al.* (2012) in this context.

3.2.3 Inference using summary-level data

In many practical settings, individual-level genotype data may not be available, and association analyses have to rely on summary statistics. In this section, we discuss inference procedures to fit the proposed Bayesian hierarchical model utilizing only summary-level information. In comparison to the existing approaches in the literature (*Chen et al.*, 2015, *Zhu et al.*, 2017), we address this problem from a distinct point of view of statistical data reduction. In particular, we attempt to identify the *sufficient*, or near sufficient, summary statistics, which could potentially lead to a minimum or no loss of inference accuracy (comparing to using complete individual-level data). Moreover, we aim to examine if the commonly applied approaches, which utilize z-scores from single-SNP association testing, are optimal in multi-SNP

association analysis.

The proposed Bayesian inference procedure depends on the observed genotype-phenotype data through the evaluation of the marginal likelihood, i.e., the Bayes factor,

$$\text{BF} := \frac{p(\vec{\boldsymbol{y}} \mid \boldsymbol{G}, \vec{\boldsymbol{\gamma}})}{p(\vec{\boldsymbol{y}} \mid \boldsymbol{G}, \vec{\boldsymbol{\gamma}} \equiv \boldsymbol{0})}. \tag{3.7}$$

For an arbitrary $\vec{\boldsymbol{\gamma}}$, *Wen* (2014) discusses an general analytic form of the Bayes factor with model (3.1) as a special case. (Note that, we take *Wen* (2014) as the starting point, because its results can be generalized to other designs of genetic association analysis, e.g., meta-analysis.) More specifically, if the residual error variance parameter τ is known, the analytic expression is exact; otherwise, it becomes an approximation by plugging in a point estimate of τ. The summary statistics required to compute the analytic form of BF include $\boldsymbol{G}'\boldsymbol{G}$ (a $p \times p$ matrix), $\boldsymbol{G}'\vec{\boldsymbol{y}}$ (a p-vector) and a point estimate of τ if τ is not known (Appendix B.2.1). Under our formulation of the regression model (i.e., all $\vec{\boldsymbol{g}}_i$'s are centered), the matrix $\boldsymbol{G}'\boldsymbol{G}$ can be factored into

$$\boldsymbol{G}'\boldsymbol{G} = \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{\Lambda},$$

where $\boldsymbol{R}$ denotes the $p \times p$ sample correlation matrix between the p candidate SNPs, and $\boldsymbol{\Lambda}$ is a diagonal matrix defined by

$$\boldsymbol{\Lambda} := \text{diag}\left(\sqrt{\vec{\boldsymbol{g}}_1'\vec{\boldsymbol{g}}_1}, \ldots, \sqrt{\vec{\boldsymbol{g}}_p'\vec{\boldsymbol{g}}_p}\right).$$

In the absence of individual-level data, some authors (*Liu et al.*, 2014) have argued explicit sharing $\boldsymbol{G}'\boldsymbol{G}$ for genomic regions of particular interests in multi-SNP fine-mapping analysis, many (*Kichaev et al.*, 2014, *Benner et al.*, 2016, *Zhu et al.*, 2017) have proposed to estimate $\boldsymbol{R}$ and $\boldsymbol{\Lambda}$, from an appropriate population panel. Henceforth, we assume that $\boldsymbol{G}'\boldsymbol{G}$ is either provided or accurately estimated, and focus

on the complete recovery of the information encoded in the p-vector, $\boldsymbol{G}'\vec{\boldsymbol{y}}$, from the summary statistics obtained in single-SNP testing.

In case that τ is known, we show that $\boldsymbol{G}'\vec{\boldsymbol{y}}$ can be accurately recovered given z-statistics and $\boldsymbol{\Lambda}$. This is because

$$z_i = \sqrt{\frac{\tau}{\vec{\boldsymbol{g}}_i'\vec{\boldsymbol{g}}_i}} \cdot \vec{\boldsymbol{g}}_i'\vec{\boldsymbol{y}}, \tag{3.8}$$

and

$$\vec{\boldsymbol{z}} = \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} = \tau^{\frac{1}{2}}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\vec{\boldsymbol{y}}. \tag{3.9}$$

Therefore, it follow that

$$\boldsymbol{G}'\vec{\boldsymbol{y}} = \tau^{-\frac{1}{2}}\boldsymbol{\Lambda}\vec{\boldsymbol{z}}. \tag{3.10}$$

We note that Equation (3.9) directly leads to the z-score distribution utilized by FINEMAP and CaviarBF (Appendix B.2.3). For some specific type of normal priors on effect s $\vec{\boldsymbol{\beta}}$, which are explicitly scaled by Λ matrix, the required summary statistics can be reduced to $(\boldsymbol{R}, \vec{\boldsymbol{z}})$.

In practice, it is unrealistic to assume the knowledge of τ and τ is required to be estimated from the data. Note that even if the priors on genetic effects $\vec{\boldsymbol{\beta}}$ are scaled by τ, as in the case of FINEMAP and CaviarBF, τ still explicitly enters into the Bayes factor computation (Appendix B.2.3). Precisely, Equation (3.9) should be modified to

$$\check{\vec{z}} = \boldsymbol{T}^{\frac{1}{2}}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\vec{\boldsymbol{y}}, \tag{3.11}$$

where $\boldsymbol{T}$ represents the following $p \times p$ diagonal matrix

$$\boldsymbol{T} = \mathrm{diag}(\check{\tau}_1, \ldots, \check{\tau}_p), \tag{3.12}$$

and each $\check{\tau}_i$ represents the estimate of τ from the simple regression model testing the association of SNP i. For the first time, we provide rigorous justification to show that, under a specific prior specification, the summary statistics, $(\boldsymbol{R}, \boldsymbol{\Lambda}, \check{\vec{z}})$, can be used to *approximate* required Bayes factor as an application of Laplace's method (Appendix B.2.3). More importantly, our derivation and numerical experiments in Appendix B.2.3 also indicates that the residual error variance can be (sometimes severely) over-estimated in applying z-scores to approximate Bayes factors, especially when multiple independent signals co-exist. As a result, overestimation of the noise levels can lead to reduced power in uncovering true association signals.

To remedy the conservativeness of the z-score based inference procedure, we propose a new analytic strategy that enables more flexible and accurate approximation of marginal likelihood. Our approach requires following summary-level information,

1. estimated effect size and its standard error, $(\hat{b}_i, \text{se}(\hat{b}_i))$, from single-SNP analysis for each SNP i (note that, $z_i = \hat{b}_i/\text{se}(\hat{b}_i)$);
2. sample size of the study, n;
3. total sum of squares (SST) of the quantitative trait: $\text{SST} = \sum_{i=1}^{n} y_i^2$, assuming $\vec{\boldsymbol{y}}$ is pre-centered.

Let $\hat{\boldsymbol{b}} := (\hat{b}_1, ..., \hat{b}_p)$ and $\hat{\boldsymbol{s}} := (\text{se}(\hat{b}_1), ..., \text{se}(\hat{b}_p))$. We show that the complete summary statistics $(\boldsymbol{R}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, n, \text{SST})$ are sufficient to accurately recover $\boldsymbol{G}'\vec{\boldsymbol{y}}$ and $\boldsymbol{G}'\boldsymbol{G}$. Furthermore, they allow estimating the corresponding MLE (or RMLE) of τ given $\vec{\boldsymbol{\gamma}}$, which leads to a more accurate approximation of Bayes factors. The detailed justification and derivation are provided in Appendix B.2.2. The major benefits of the proposed approach are

1. It allows accurately estimating τ from the data matching any given $\vec{\boldsymbol{\gamma}}$ value;

2. It allows work with an arbitrary type of the normal prior on genetic effect size (with or without scaling by τ and/or $\mathbf{\Lambda}$).

Beyond the setting described by the model (3.1) for a single genetic association analysis, the proposed approach can be straightforwardly extended to multi-SNP analysis in a meta-analysis or trans-ethnic genetic association analysis using summary-level statistics *Wen et al.* (2015a). (For this purpose, the second point above is particularly important.) From both simulations and real data analysis, we find that the ability to dynamically estimate τ according to the selected candidate SNPs can significantly improve the signal-to-noise ratios required for discovering multiple genuine genetic association signals. This factor likely explains the observation that approaches utilizing individual-level data typically outperform the existing approaches utilizing only z-scores. Our proposed strategy bridges this gap: if the LD information (namely, $\boldsymbol{R}$) is sufficiently accurate, the results based on the summary-level information are *identical* to those based on individual-level genotype data.

3.3 Results

3.3.1 Simulation studies

We set up a simulation scenario mimicking *cis*-eQTL mapping in a practical setting. In particular, we use the real genotype data from 343 European individuals from the GUEVADIS project (*Lappalainen et al.*, 2013). We artificially construct a genomic region of 1,001 SNPs. The region is divided into 91 LD blocks, and each block contains 11 SNPs. All LD blocks are selected from chromosome 1, and the consecutive blocks are at least 1Mb apart. With this construction scheme, the LD only presents within each block, and the SNP genotypes are mostly uncorrelated across blocks (Supplementary Figure A8). We simulate a quantitative phenotype

according to a sparse linear model. Specifically, with probability 0.05, an LD block is selected and a causal association is randomly assigned to one of its 11 member SNPs. On average, 4.75 genuine associations are expected from the whole region. The genetic effect of a causal SNP is independently drawn from a normal distribution, $\mathrm{N}(0, 0.6^2)$, and the residual error for each sample is independently simulated from $\mathrm{N}(0, 1)$. Those particular parameters are selected such that the distribution of single SNP testing z-statistics from the simulated data matches the characteristics of the empirical distribution observed from the *cis*-eQTL analysis from multiple real eQTL data sets, namely GEUVDIS and GTEx (Supplementary Figure A9). We generate 1,000 independent data sets using this scheme.

The simulated data sets are analyzed using three methods:

1. DAP-G with sufficient summary statistics, i.e., $(\boldsymbol{R}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, n, \mathrm{SST})$;
2. DAP-G with single SNP testing z-scores, i.e., $(\boldsymbol{R}, \boldsymbol{\Lambda}, \check{\vec{z}})$;
3. FINEMAP with single SNP testing z-scores, i.e., $(\boldsymbol{R}, \boldsymbol{\Lambda}, \check{\vec{z}})$

The software package FINEMAP (*Benner et al.*, 2016) implements a particular version of MCMC algorithm using the shotgun stochastic search scheme. Moreover, it utilizes the summary information $(\boldsymbol{R}, \boldsymbol{\Lambda}, \check{\vec{z}})$ as input to compute the same approximate Bayes factors as in CAVIARBF. Because of its superior computational efficiency and accuracy compared to other available methods (see *Benner et al.* (2016) for details), we considered it the state-of-the-art for multi-SNP genetic association analysis using summary-level information.

We use the default priors for both DAP-G and FINEMAP, which are slightly different. DAP-G employs a more conservative default prior with respect to the simulated data sets, which assumes a single causal variant is expected *a priori*. FINEMAP, designed for fine-mapping analysis, assumes $\Pr(\vec{\boldsymbol{\gamma}} = \mathbf{0}) = 0$ by default.

In comparison, $\Pr(\vec{\gamma} = \mathbf{0}) = (1 - 1/1001)^{1001} = 0.368$ for DAP-G. As a result, we conclude that our simulated data scheme, in this case, slightly favors FINEMAP.

None of the methods assumes the knowledge of the artificial LD blocks constructed in the simulated data, i.e., LD information is inferred from the genotype data through $\boldsymbol{R}$ in all three approaches.

Power for signal discovery

We first examine the power of all methods in uncovering the LD blocks that harbors a causal association signal.

Because the concept of a signal cluster is not defined in FINEMAP, we compute the cumulative PIPs for each constructed LD blocks and use this quantity to rank the blocks within each method. Although this approach does not always guarantee a valid probability for each pre-defined block, especially for FINEMAP, we find very few false positives from the blocks with cumulative PIPs > 1 for all three methods. We construct and compare the receiver operating characteristic (ROC) curves based on the ranking of the pre-defined LD blocks across all simulations. In addition to the three aforementioned approaches, we also rank the LD blocks by their the minimum p-values from the single SNP testing of their member SNPs. This approach is commonly used to identify eGenes (i.e., genes harboring at least an eQTL in their cis region) in *cis*-eQTL mapping.

Figure A1 shows the comparison of the ROC curves in a practically meaningful range, i.e., the false positive rate $< 50\%$. In this set of simulations, the best performer is the DAP-G algorithm running with the sufficient summary statistics: for any false positive rate threshold, it always identifies more true positives than any of the approach in comparison. The difference in performance between the DAP-G algorithms using different summary statistics input confirms our theoretical argument

for the superiority of the sufficient summary statistics over the z-scores. Although the DAP-G with z-score input and FINEMAP compute the approximate Bayes factor the same way, there is very noticeable difference reflected by the ROC curves. We suspect this difference is mainly attributed to the convergence issue of the MCMC algorithm employed in FINEMAP: if the MCMC run can be extended (significantly) longer, we expect these two sets of results would eventually converge. Finally, it is clear that all multi-SNP analysis approaches outperform the single-SNP method in identifying genetic loci that harbor association signals by a large margin in this simulation setting.

Calibration of SNP-level PIP

Next, we inspect the calibration of SNP-level PIPs obtained from the different methods. The calibration of Bayesian posterior probabilities refers to the frequency property in repeated observations. For example in our specific context, it is expected that among many SNPs assigned PIP = 0.50, half of them are genuinely associated if the PIPs are indeed calibrated. The calibration of the posterior probabilities indicates the robustness of the model and the accuracy of the Bayesian computation.

For each method examined, we group all SNPs across simulated data sets into 10 bins according to their reported PIP values (namely, $[0, 0.1), [0.1, 0.2), ..., [0.9, 1.0]$). We then compute the proportion of truly associated SNPs in each bin. We expect that the frequency value is aligned to the average PIP value for each bin for calibrated SNP-level posterior probabilities.

Figure A2 shows that, among three methods compared, DAP-G running with sufficient summary statistics yield most calibrated SNP-level posterior probabilities. As expected, the PIPs by DAP-G using z-scores as input are slightly conservative. For FINEMAP, the posterior probabilities in some high-value PIP bins are shown to

be anti-conservative, indicating potential convergence issues in MCMC runs.

FDR control at signal level

We then proceed to inspect the performance of FDR control at the signal cluster level by DAP-G. We performed the proposed Bayesian FDR control procedure using the inferred SPIP values. We label a true discovery if an inferred signal cluster indeed contains a causal SNP and the corresponding SPIP is greater than a pre-defined threshold and a false discovery otherwise. Subsequently, we compute the realized false discovery proportion (FDP) and the power with respect to the corresponding FDR control threshold. We repeat this procedure for a set of FDR levels ranging from 0.01 to 0.25. The detailed results are shown in Table A1. In all cases, the FDR's of the signal clusters are conservatively controlled at all pre-defined levels. Furthermore, by utilizing sufficient summary statistics the power of discovering association signals is consistently higher than using z-scores.

Computational efficiency

Our implementation of the DAP-G algorithm is highly efficient: we observe that DAP-G runs magnitude faster than the state-of-the-art FINEMAP program. The speed-up is mainly due to the nature of deterministic search algorithm. Additionally, the implemented functionality of parallel processing for the DAP-G deterministic search procedure (via the OpenMP library) also contributes to the improved computational efficiency. For a dataset contains 5 independent signals, DAP-G runs about 1.5 seconds with 4 parallel threads and correctly identifies 4 of the 5 signals. In comparison, FINEMAP also achieves the same accuracy, and the runtime is benchmarked at 1 minute and 45 seconds on the same computer. The total user time for analyzing the complete set of 1,000 simulated data sets are 34 minutes 48 seconds

and 741 minutes 6 seconds for DAP-G and FINEMAP, respectively. With 4 data sets being simultaneously analyzed on an eight-core Xeon 2.13 GHz Linux system, the real time of the complete analysis for DAP-G and FINEMAP are 7 minutes 50 seconds and 190 minutes 36 seconds, respectively.

3.3.2 Multi-SNP analysis of *cis*-eQTLs in GTEx whole blood samples

In this section, we illustrate a complete process of *cis*-eQTL mapping of the GTEx whole blood samples (version 6p) using the proposed DAP-G algorithm. The GTEx whole blood data include 338 individuals for which dense genotyping are performed. The expressions of 22,749 protein-coding and lincRNA genes are measured by RNA-seq experiments. The individual-level genotype-phenotype data are available for analysis. We followed the procedures described in *GTEx Consortium* (2017) to perform pre-processing and quality control of the genotype and expression data. For *cis*-eQTL mapping, we focus on the candidate genetic variants located within a 1Mb radius of the transcription start site (TSS) of each gene. On average, there are 7,118 candidate genetic variants per gene and no further SNP filtering procedure is taken before the multi-SNP association analysis.

We take an empirical Bayes approach to estimate the prior inclusion probability for each SNP. The estimation procedure, implemented in the software package TORUS *Wen* (2016), utilizes the single-SNP association testing results across all genes. Additionally, it can incorporate SNP-level annotation data. We classify the candidate SNPs into 21 categories according to their distances to the TSS (DTSS) of corresponding genes and allow the priors vary in different categories. This decision is motivated by the previous observations (in almost all eQTL studies) that the abundance of *cis*-eQTLs is strongly associated with SNP DTSS. Our estimated priors by DTSS bins (Figure A3) from the GTEx data clearly confirms this pattern.

We then proceed to analyze all 22,749 genes using DAP-G. On a computing cluster and with 30 to 50 genes simultaneously analyzed, the processing of the complete data set takes about 14 hours. First, we compute the posterior expected number of *cis*-eQTLs for each gene by

$$\mathrm{E}(\sum_{i}^{p} \gamma_i \mid \vec{\boldsymbol{y}}, \boldsymbol{G}) = \sum_{i}^{p} \Pr(\gamma_i = 1 \mid \vec{\boldsymbol{y}}, \boldsymbol{G}). \tag{3.13}$$

Figure A4 shows the histogram of the expected number of *cis*-eQTLs across all genes, which indicates that we are able to confidently identify multiple independent *cis*-eQTLs for a good proportion of genes. Applying the proposed FDR control procedure, we identify 9,056 independent *cis*-eQTL signals from 7,135 unique genes by controlling FDR at 5% level. A subset of 6,328 signals from 5,123 unique genes exceeds the more stringent threshold at 5% local fdr, for which we can construct 95% credible sets. There is a substantial variation in the size of the 95% credible sets (Figure A5). The median size of the credible sets is 7, and the mean is 14.9. The average pairwise r^2 between SNPs in a credible set is 0.85 (median = 0.89). The largest credible set observed in this data set represents a *cis*-eQTL signal for gene *KANSL1* (ensembl id: ENSG00000120071) located at chromosome 17 (SPIP $\sim$ 1.0), which consists of 354 tightly linked SNPs (average pairwise $r^2 = 0.90$). Even for a single gene, we sometimes observe various sizes of credible sets. Figure A6 shows gene *TMTC1* (ensembl id: ENSG0000133687) for which we confidently identify 4 independent *cis*-eQTL signals. Interestingly, two of the signals have relatively small 95% credible sets containing 1 and 4 SNPs, respectively; while the credible sets for the other two signals are noticeably larger, containing 20 and 32 SNPs, respectively. These results reinforce our observations that causal associations can be complicated to identify even if the evidence for the existence of an association signal (e.g., SPIP) is overwhelming.

To compare results with summary statistics based inference, we extract summary-level information from the complete data in two forms: the sufficient summary statistics, $(\boldsymbol{R}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, n, \mathrm{SST})$, and commonly used $(\boldsymbol{R}, \hat{\tilde{z}})$. As predicted by our theoretical arguments, we find that the inference results based on sufficient summary statistics are identical to the analysis of individual-level data, whereas noticeable discrepancy can be observed from the inference results applying z-score based summary statistics. Figure A7 shows the comparison of SNP-level PIPs among the three different inputs for gene *TMTC1*. Particularly when z-scores are used as input, we note that the SPIPs for the third and the fourth signals in the original analysis for *TMTC1* are severely under-estimated and the 95% credible sets can no longer be constructed. In comparison, the SPIPs for the first two signals are still close to 1, and the corresponding credible sets mostly remain the same. We find these results are also consistent with our observations from the simulation studies.

In summary, we find our *cis*-eQTL mapping analysis by DAP-G is highly efficient. The multi-SNP analysis results are more informative and more natural to interpret in comparison to the standard single-SNP analysis. We provide the complete analysis results in, which include the quantification of all *cis*-eQTL signal clusters and corresponding credible sets.

3.4 Discussion

In this chapter, we have described a powerful and efficient computational approach to perform multi-SNP genetic association analysis. Within the Bayesian framework, we have introduced a new paradigm to comprehensively represent a complex genetic association signal in a natural hierarchy that accounts for LD structures and easy to interpret. With the probabilistic quantification of the strength of association

evidence, we have shown rigorous FDR control can be straightforwardly applied. From the perspective of data reduction, we have derived the sufficient summary statistics, $(\boldsymbol{R}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{s}}, n, \text{SST})$, that result in identical inference with individual-level data in quantitative trait mapping. Furthermore, we are also able to establish the theoretical connection to the commonly applies inference based on summary-level data, which is shown to be a conservative approximation to the exact inference using individual-level data.

In *cis*-eQTL mapping, we have illustrated that multi-SNP analysis can completely replace the need for reporting single SNP analysis findings because of its informativeness and efficiency. We believe the same argument can be made regarding the analysis of GWAS data. We acknowledge that almost all multi-SNP genetic association analysis approaches, including ours, do not computationally scale beyond a genomic region up to 4 Mb and most commonly applied for fine-mapping analysis instead of genome-wide scan. Many have shown (*Berisa and Pickrell*, 2016, *Wen et al.*, 2016) that it is effective to apply a divide-and-conquer strategy that segments genome according to population-specific LD blocks and performs multi-SNP analysis independently on each LD block. This strategy may be necessary if genomic annotations are incorporated into GWAS analysis and an unbiased enrichment analysis contrasting annotated functional SNP versus unannotated is desired. Additionally, with sample sizes of GWAS reach to the bio-bank scale, improved power for uncovering modest genetic association signals has become critically important. As shown in our simulation study, especially Figure A1, identifying critical regions through filtering via single SNP testing may not be the best practice.

With the previous results on computing Bayes factors in complex linear systems (*Wen*, 2014), our results presented in this chapter can be straightforwardly extended

to accommodate many different study designs for studying complex and molecular traits. The important applications include multi-SNP analysis in meta-analysis setting and eQTL mapping across multiple tissues, just to name a few. With the availability of the analytic forms of approximate Bayes factors under these complicated settings, it is now possible to perform rigorous FDR control and carry out the computation through summary statistics.

Genetic association analysis is not and should never be the end point of scientific discovery. It is therefore critically important to disseminate the findings in genetic association analysis to the downstream analysis and experimental work. From this perspective, Bayesian approaches are generally advantageous mainly because of their use of probabilistic quantification to summarize association results comprehensively. This point has been illustrated by the co-localization analysis of molecular QTL and GWAS signals, where most existing approaches (*Giambartolomei et al.*, 2014, *Hormozdiari et al.*, 2016, *Wen et al.*, 2017) all require posterior probabilities from both complex and molecular trait-association analyses. For integrative analysis requiring results from genetic association analysis, e.g., SNP-level eQTL annotations, the probabilistic quantification of association results is more appropriate than the binary classification based on some stringent type I error threshold. This is because the latter approach fundamentally ignores the potential type II errors (which also contribute to the misclassifications) and can introduce severe bias in the integrative analysis.

It is worth pointing out that the equivalency of analysis results by individual-level and using summary-level data is based on the assumption that the correlation matrix between candidate variants, $\boldsymbol{R}$, is estimated accurately. The deviation from this assumption can cause noticeable discrepancy between the two types of analyses.

We acknowledge that the estimation of $\boldsymbol{R}$ from an appropriate population is an important problem and refer the readers to some important recent works on this topic (*Zhu et al.*, 2017).

Web Resources

DAP-G software and tutorial, `http://github.com/xqwen/dap/`

GTEx data, `http://www.gtexportal.org/home/datasets`

Simulation data and code, `https://github.com/xqwen/dap/tree/master/dap-g_paper/simulation` Multi-SNP fine-mapping results of the GTEx whole blood data, `https://github.com/xqwen/dap/tree/master/dap-g_paper/gtex_v6p`

CHAPTER IV

Measuring Reproducibility Accounting for Reproducibility

4.1 Introduction

The advancement of high-throughput technologies has allowed researchers to study transcriptomes and other metabolomics with large sets of data, including information on numerous candidate genetic variants and various molecular expression levels. However, utilizing data from high-throughput technologies often confronts by several challenges in practice. One of these challenges is assessing the level of concordance between high-throughput assays. Since batch effects and other unknown systematic errors heavily affect signals from these assays, it is crucial to measure reproducibility between signals from these assays.

In addition, many association studies are finding genetic variants that are associated with complex traits. However, some novel findings are not replicable in other studies, either due to issues of powers or systematic differences between studies. Therefore, by measuring reproducibility between these results, we seek to distinguish strong and replicable signals from study-specific signals and non-signals.

One of the commonly used methods to measure reproducibility for high-throughput assays is computing the Spearmans pairwise rank correlation coefficient among significant results. However, this approach has several limitations, including the de-

pendence on the choice of thresholds and not emphasizing the importance of the consistency of top-ranked signals. *Li et al.* (2011) suggested an alternative method of rank correlation, called the Irreproducibility Discovery Rate (IDR), which applies the concept of false discovery rate (FDR) to this setting. The IDR approach assumes two different dependent structures for the models of spurious and genuine signals, and fits a Gaussian copula mixture model to estimate IDR and local IDR (analogous to local FDR). The IDR approach resolves several issues of rank correlation method and improve the power of identifying genuine signals. However, we find that fitting the model using software provided by *Li et al.* (2011) often fails to converge, even with a large number of iterations, and is sensitive to the user-provided parameters.

In addition, none of the existing methods to measure reproducibility utilizes the directional information of estimated effects, because this information is lost during rank transformation. Considering the directional consistency can provide additional information to classify strong, genuine signals. For example, if a specific allele is a genuine and strong regulate variants for a target gene, its association with an expression level would generally increase (or decrease) the expression level across studies. Assuming its regulation effect is strong enough to overcome study-specific confounders, it should be identified as reproducible signals. By incorporating directional consistency of effects between studies, we can measure reproducibility with improved accuracy and so better identify reproducible signals.

In this chapter, we utilize the Bayesian frameworks by *Wen* (2016) on summary statistics to measure irreproducibility so address these limitations of existing methods. First, we propose a visualization and quantification tool to aid the assessment of reproducibility using only rank information. Second, we propose a Bayesian approach to control IDR rigorously within a local FDR framework. Our approach utilizes the

quantitative information, specifically regression coefficients and its standard errors, and is more powerful than rank-based approaches.

4.2 Models and Methods

In this section, we first describe a visualization tool to characterize the reproducibility of two datasets based on rank information. We then proceed to propose a probabilistic hierarchical model to fully utilize the information of directional consistency of effects presented in multiple datasets.

4.2.1 Visualization of Reproducibility between Studies

We consider a scenario where we compare results of two association studies. We aim to visualize the degree of concordance of the two association results via a scatter plot.

Specifically, for each testing unit (e.g., a gene-SNP pair in single-SNP eQTL analysis), we utilize summary statistics such as Bayes factors from association studies. We rank n overlapping testing units within each study. For a testing unit i, we have a pair of ranks (i_1, i_2) for studies 1 and 2. The higher rank typically implies stronger evidence for association studies.

For study j, the rank transformation of the effect size estimate is as follows:

$$z_{i,j} = -\Phi^{-1}(\frac{i_j}{n+1}), \tag{4.1}$$

where Φ denotes the cumulative probability distribution function of the standard normal distribution. We plot $(z_{i,1}, z_{i,2})$ for all testing units and refer to the resulting scatterplot as a *rank copula plot*. In the process, we also estimate the two-dimensional empirical kernel density.

Rank copula plots are designed to visualize the degree of concordance and non-concordance between two studies based on rank information. When results from two

studies are extremely non-concordant, $z_{i,1}$ and $z_{i,2}$ tend to be uncorrelated, consequently the rank copula plot would resemble a scatterplot of independent bivariate normal distribution. For highly reproducible results, $z_{i,1}$ and $z_{i,2}$ tend to be highly correlated. Hence, most points should scatter around the line of slope 1, which can be approximated by a bivariate normal with correlation coefficient closed to 1. Rank copula plots from real data are different from scatterplots of two extreme bivariate normal distributions. Since the correlations between $z_{i,1}$ and $z_{i,2}$ are expected to be high for stronger signals and low for weak and non-signals, the rank copula plots from real data shows a pattern mixing the extreme scenarios.

We consider two measures to quantify the degree of concordance shown in rank copula plots. These measures are the empirical Kullback-Leibler (KL) divergence and the empirical mutual information (*Kullback and Leibler*, 1951).

KL divergence is a measure that quantifies the difference of an observed distribution from an expected distribution. In the calculation of the empirical KL, KL, the observed distribution is the empirical distribution of derived from the transformed ranks. For the expected distribution, we consider two reference distributions that represents extreme concordance and extreme non-concordance. For the empirical mutual information, we measure it by calculating the empirical KL divergence with the same reference datasets as KL, and it is denoted by $MI - KL$ in this chapter. The observed distribution for the empirical mutual information is the joint distribution of the observed distribution and the reference distributions.

Let $KL_{concordance}$ denote the empirical KL divergence computed from contrasting observed rank distribution and the reference concordance distribution. This quantity is computed using a reference dataset drawn from a bivariate normal with a correlation coefficient close to 1. Similarly, we define $KL_{non-concordance}$ as the empir-

ical KL divergence computed from the observed rank distribution and the reference non-concordance distribution, drawn from an independent bivariate normal.

In practice, we calculate the density functions by setting bins on two-dimensional spaces and count the number of testing units in each bin. Both $KL_{concordance}$ and $KL_{non-concordance}$ is calculated as follows:

$$\text{KL} = \sum_t P_{obs,t} \times \log(\frac{P_{obs,t}}{P_{ref,t}}), \tag{4.2}$$

where $P_{obs,t}$ is the density of observed dataset, and $P_{ref,t}$ is the density function of reference dataset in bin t.

MI-KLs are also calculated using the same reference distributions, and denoted by $MI-KL_{concordance}$ and $MI-KL_{non-concordance}$ respectively. When $P_{ref,obs}$ denotes the joint distribution of reference and observed distributions, the empirical MI-KL is calculated as follows:

$$\text{MI-KL} = \sum_t P_{ref,obs,t} \times \log(\frac{P_{ref,obs,t}}{P_{ref,t} \times P_{obs,t}}). \tag{4.3}$$

Since KL is not a bounded measure and has no explicit interpretation, we suggest using KL mainly to compare the degree of concordance and non-concordance among different pairs of studies. For example, when we compare the degree of concordance of eQTLs between same tissues and different tissues, we can compute $KL_{concordance}$ and $KL_{non-concordance}$ for two comparisons using the same reference datasets. Then we can compare $KL_{concordance}$ of same tissues and $KL_{concordance}$ of different tissues, and decide which result has the higher degree of concordance.

4.2.2 Measuring Reproducibility Accounting for Directional Consistency

Motivation

Most existing methods measuring reproducibility use rank-based statistics and do not consider the directional consistency of the estimated effects. However, directional consistency is highly informative for measuring reproducibility. For example, for a genuine eQTL, we expect that the same allele would always increase (or decrease) the level of expression of a targeted gene in different studies. Therefore, the estimated effects of the allele would have the same direction of effects across studies.

In practice, we always expect some heterogeneity due to biological variation. However, the degree of heterogeneity of reproducible signals should be constrained by the directional consistency.

To meet the directional consistency requirement, we propose a novel probabilistic hierarchical model to quantify the degree of concordance between studies.

Overview of Method

Basic Ideas of Model We define a reproducible signal as a genuine signal that is reproducible across studies. To classify reproducible signals, we consider the case of two studies. Maximum likelihood estimates (MLEs) of the effect size for a given testing unit are denoted by $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

We assume a prior that defines the level of heterogeneity expected from a reproducible signal, i.e.:

$$P(\beta_1 \text{ and } \beta_2 \text{ have different signs} \mid \text{a reproducible signal }),$$

where β_1 and β_2 are the unobserved true effects. This prior enables the computation of a posterior probability of being reproducible,

$$P(\text{ a signal is reproducible } |\hat{\beta}_1, \hat{\beta}_2), \tag{4.4}$$

which has a natural interpretation for classifying the reproducibility of the signal.

Statistical Model The likelihood of the observed data can be described by :

$$\hat{\beta}_i|\beta_i \sim N(\beta_i, \sigma_i^2), \tag{4.5}$$

while β_i is the unobserved true effect in study i. β_i is study-specific, since it may be affected by study-specific confounders. These confounders could be the characteristics of samples, data processing procedures or other systematic errors.

We assume the prior distribution on β_i, as follows:

$$\beta_i|\beta \sim N(\beta, k^2\beta^2), \tag{4.6}$$

where β is the true underlying biological effect of the testing unit, assumed to be the same for between studies. The parameter k quantifies the heterogeneity of effects between studies. In the special case $k = 0$, the model becomes a fixed-effect meta-analysis model. This type of prior is referred to as the curved exponential family normal prior in *Wen et al.* (2014). With this prior, the probability of β_i having a different sign from β only depends on k and not β as follows:

$$\Pr(\beta_i \text{ is having a different sign from } \beta \mid \text{ a genuine signal }) = \Phi(-\frac{1}{|k|}). \tag{4.7}$$

Finally, we assume the prior on β is given by:

$$\beta \sim N(0, \omega^2), \tag{4.8}$$

where ω quantifies the effect size of β. In the special case of $\omega = 0$, both β and β_i are strictly becomes 0, which describes a theoretical null model.

These parios specification imply that

(4.9)

$$\Pr(\beta_1 \text{ and } \beta_2 \text{ have a same sign} \mid \text{a genuine signal }) = \{1 - \Phi(-\frac{1}{k})\}^2 + \Phi(-\frac{1}{k})^2.$$

We consider limiting the probability

$$\Pr(\beta_1 \text{ and } \beta_2 \text{ have a same sign} \mid \text{a reproducible signal }) \geq 0.98, \tag{4.10}$$

which implies $k \leq 0.43$, following the expression 4.9. We note that reproducible signals are a subset of genuine signals. Genuine signals cannot be identified in some studies due to study-specific variations. However, reproducible signals are those who are strong enough to be identified in both studies. Our method provides a measure to classify reproducible signals from signals that are not reproducible (irreproducible signals) and non-signals.

By using a set of different k and ω, testing units can be partitioned into three possible scenarios. The first scenario describes non-signals, i.e. $\omega = 0$. The second scenario describes weak signals, considered as irreproducible. In such case, ω is nonzero and $k \geq 0.43$. Finally, the third scenario describes all the reproducible signals, which not only have $\omega \neq 0$, but also $k \leq 0.43$.

Overview of Computational Procedures

Computational Overview For computational purposes, we use a set of grid (k, ω) values to represent the complete model space to cover all three scenarios. We calculate the corresponding Bayes Factor, for each grid based on $\hat{\beta}$ and $\hat{\sigma^2}$, following the approach suggested by *Wakefield* (2009) and *Wen et al.* (2014).

Let $p_1, p_2, \ldots p_L$ denote the prior probabilities on all (k, ω) grid values. For a target SNP, we compute the posterior probability as follows:

$$PP_{rep} = \frac{\sum_{m \in \Omega_{III}} p_m BF_m}{\sum_o p_o BF_o}, \tag{4.11}$$

where the set Ω_{III} denotes the grid values representing the reproducible scenario. We take an empirical Bayes approach to estimate the priors $p_1, p_2, \ldots p_L$ from data. Specifically, we use the EM algorithm implemented in the software TORUS (*Wen*, 2017) to find maximum likelihood estimates (MLEs) for $p_1, p_2, \ldots p_L$. Note that PP_{irr}, $1 - PP_{rep}$, can be interpreted as the local false discovery rate (local FDR), and we can apply the FDR control procedure suggested by *Efron et al.* (2007).

4.3 Results

4.3.1 Visualization of Reproducibility between Studies

We drew the rank copula plot for the comparison of gene-level eQTLs between FUSION skeletal muscle study (*Scott et al.*, 2016) and skeletal muscle tissue from GTEx project (*GTEx Consortium*, 2017). FUSION skeletal muscle study analyzes eQTLs with 267 Finnish individuals for skeletal muscle tissue, while the GTEx project (v6) analyzes eQTLs for 51 tissues with 570 donors. The sample size of skeletal muscle tissue of GTEx project (v6) is 430 and that of blood tissue is 393.

We used 19,037 overlapping genes for the comparison between skeletal muscle tissues from two studies. For the comparison of between FUSION skeletal muscle tissue and GTEx blood tissue, we used 17,579 overlapping genes. We then computed a gene-level Bayes factor for each gene by applying the statistical method proposed in *Veyrieras et al.* (2008), following the procedure described in *Wen* (2016).

In Figure 4.1, the plot on the left side shows the degree of concordance of rank-transformed gene-level Bayes factors between skeletal muscle tissues from FUSION skeletal muscle study and the GTEx project. The plot on the right side displays the

degree of concordance between FUSION skeletal muscle study and blood tissues from the GTEx project. In the comparison of skeletal muscle tissues, most of gene-level eQTLs are expected to be shared between tissues. Since the large portion of eQTLs including muscle-specific eQTLs would be shared between the studies, we expect the higher degree of concordance between muscle tissues than muscle and blood tissues. As expected, the left plot in Figure 4.1 shows a more concordant pattern than the right plot.

While the comparison between different tissues (muscle and blood) shows the lower level of concordance than same tissues (muscle), the plot on the right in Figure 4.1 still shows the moderate degree of reproducibility. This corresponds to the biological fact that while there are tissue-specific eQTLs, the many eQTLs are shared across tissues.

To quantify the degree of concordance shown in Figure 4.1, we calculated the empirical MI-KL and KL. The calculated measures are displayed in Table 4.1. The $KL_{concordance}$ of two muscle tissues are smaller than $KL_{concordance}$ of muscle and blood tissues. Also, the $KL_{non-concordance}$ of two muscle tissues are greater than $KL_{non-concordance}$ of muscle and blood tissues. These KL values confirm the conclusion from Figure 4.1 that the comparison of same tissues show the higher degree of concordance than that of different tissues.

We find that $MI-KL$s deliver the mixed conclusions. Both $MI-KL_{concordance}$ and $MI-KL_{non-concordance}$ of muscle tissues are smaller than those of muscle and blood tissues. We suggest that the interpretation of $MI-KL$ should be carefully done.

Studies	Deviation from Extreme Concordance		Deviation from Extreme non-concordance	
	MI-KL	KL	MI-KL	KL
FUSION muscle vs GTEx muscle	0.3401	10.5080	1.1580	0.7840
FUSION muscle vs GTEx blood	0.4312	14.2436	1.3897	0.4159

Table 4.1: Empirical mutual information based on Kullback-Leibler divergence(MI-KL) and Kullback-Leibler divergence(KL), calculated based on gene-level eQTL analyses from FUSION skeletal muscle, GTEx muscle and GTEx blood tissues. The reference dataset for extreme concordance is drawn from bivariate normal distribution with a correlation coefficient 0.99. The reference dataset for extreme non-concordance is drawn from independent bivariate normal distribution. The observed datasets are gene-level Bayes factors.

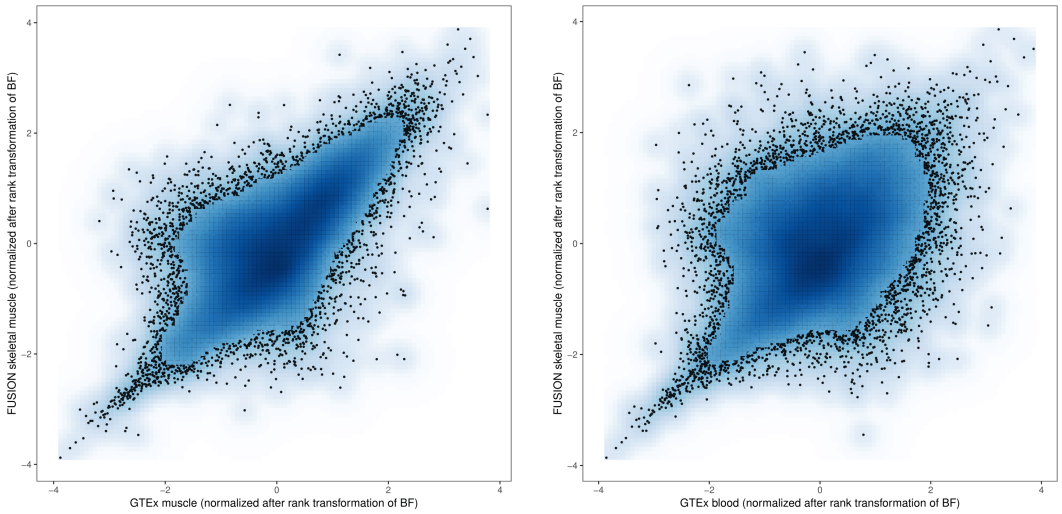

Figure 4.1: Examples of rank copula plots between studies. The datasets used in these plots are the results of gene-level eQTL analyses from FUSION skeletal muscle, GTEx muscle and GTEx blood tissues. Two-dimensional kernel density is estimated and plotted based on transformed ranks of gene-level Bayes factors.

4.3.2 Measuring Reproducibility with the Direction of Effects

Simulation Study

We conducted a simulation study to assess the performance of the proposed method on identifying reproducible signals.

Details of Simulation Study For each dataset, we generated 10,000 overlapping testing units across two studies. The estimated effects of each testing unit are z-scores drawn from the mixture of two bivariate normal distributions. Specifically, 5,000 z-scores were drawn from independent bivariate normal distribution and representing

non-signals. The remaining 5,000 z-scores, representing signals, were drawn from the following distributions :

$$
\begin{aligned}
z_i &\sim N(\bar{z}, \phi^2),\ i = 1, 2 \\
\bar{z} &\sim N(0, M+1), M > 0
\end{aligned}
\tag{4.12}
$$

where z_i is a z-score drawn for a testing unit in study i. ϕ is a parameter that controls the heterogeneity between studies, and M is a non-centrality parameter that controls the underlying true effect size for each testing unit. The z-scores drawn from this distribution are signals, and their reproducibility was decided by ϕ.

We carried out simulations for different combinations of M and ϕ. M is taken from $\{1, 3, 8, 11\}$ and ϕ from $\{0.5, 1, 1.5, 2\}$. We then applied the proposed method to each simulated dataset.

For each generated dataset, we applied the MeSH (*Wen et al.*, 2014) software to calculate BF for each grid point (k, ω). Minimum, maximum and the number of total points of ω are derived from each dataset, following the guide from *Stephens* (2016). The grid points of k are selected to cover all three scenarios, including the threshold defined in Section 4.2.2, 0.43. The weights of each BF were estimated using the EM algorithm implemented in the software package TORUS (*Wen et al.*, 2015a). We then performed Bayesian model averaging to calculate PP_{rep}. The FDR control procedure is performed on the local FDRs derived from PP_{rep}.

Simulated results are displayed in Figures 4.2, 4.3 and 4.4. Figure 4.2 simultaneously demonstrates the parameter settings and the reproducibility of each testing unit in rank copula plots. Figure 4.3 highlights the clear distinction between reproducible signals and other scenarios. Especially in Figure 4.2, we can see that the proposed method can classify reproducible signals (purple dots) from irreproducible signals (blue green dots) and non-signals (red dots) in the spectrum of effect sizes.

Figure 4.3 are scatter plots with the generated z-scores, highlighting the identified reproducible signals as red dots. These plots depict the signs of simulated z-scores. These plot also clearly show that our method can identify z-scores showing the directional consistency as reproducible signals.

As expected, Figure 4.4 demonstrates that with the same underlying effect size, the heterogeneity between studies (ϕ) has an influence on the proportion of reproducible signals identified, as the model suggested.

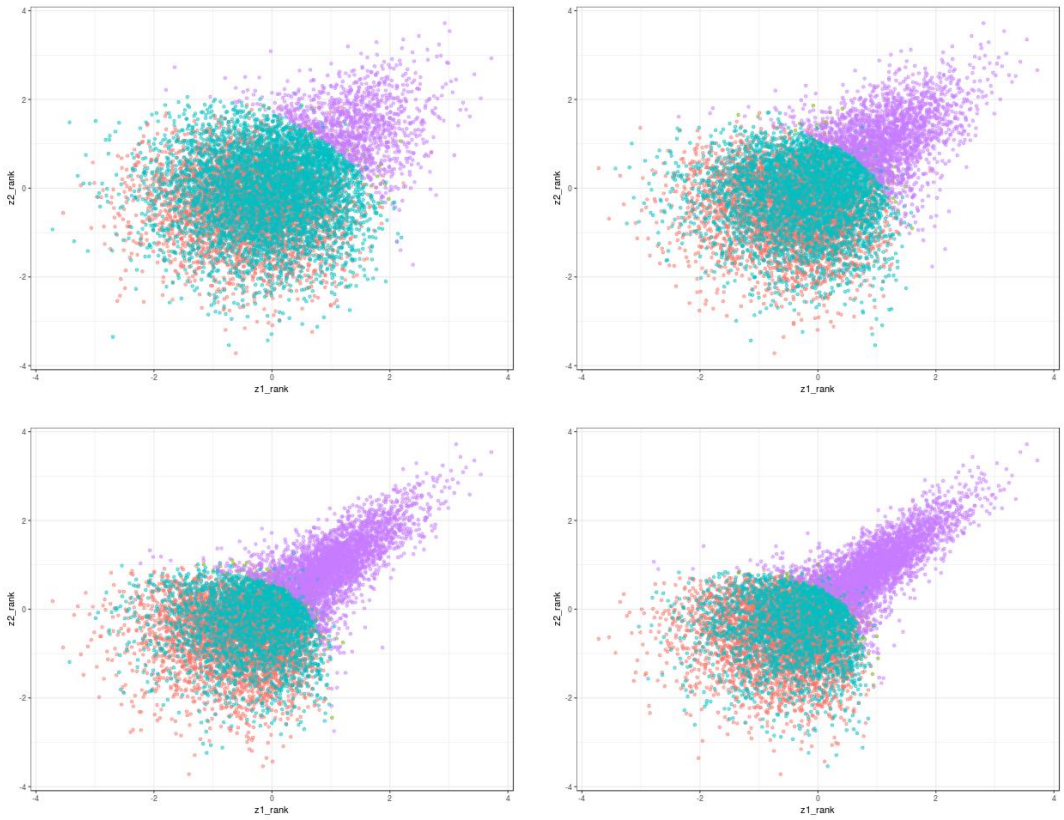

Figure 4.2: Rank copula plots show the result of simulations studies from datasets generated with various M. Ms are set as 1,3,8 and 11 from the upper left plot to the lower right plot. ϕ is set as 1 for all plots. Orange and green color dots represent identified irreproducible and reproducible signals respectively, when z-scores are generated from the null model. Blue green and purple color dots represent identified irreproducible and reproducible signals respectively, when z-scores are generated from the alternative model.

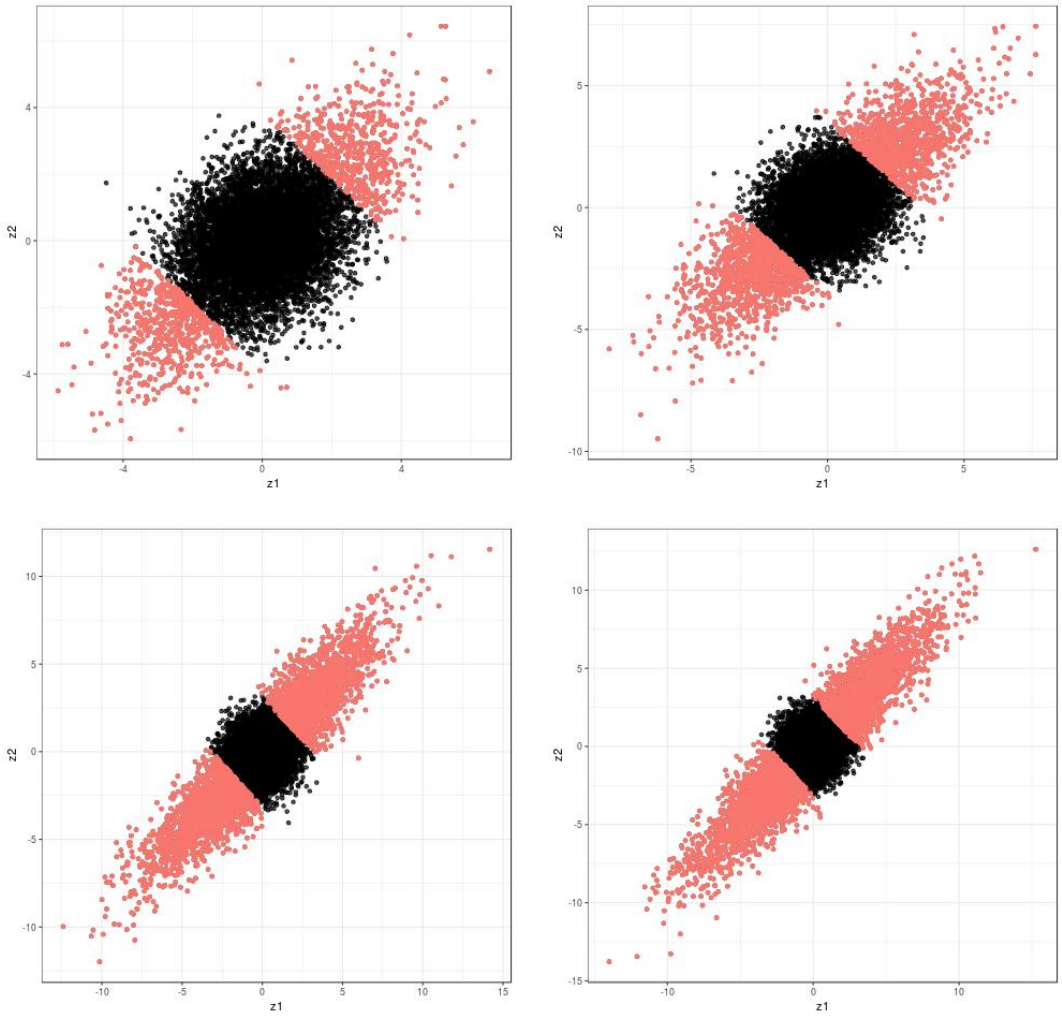

Figure 4.3: Scatter plots show the result of simulations studies from datasets generated with various M. Ms are set as being 1,3,8 and 11 from the upper left plot to the lower right plot. ϕ is set as 1 for all plots. Each black dot represents the pair of z-scores of each testing unit, which is overlaid by a red dot when the testing unit is categorized as reproducible signals after the FDR control.

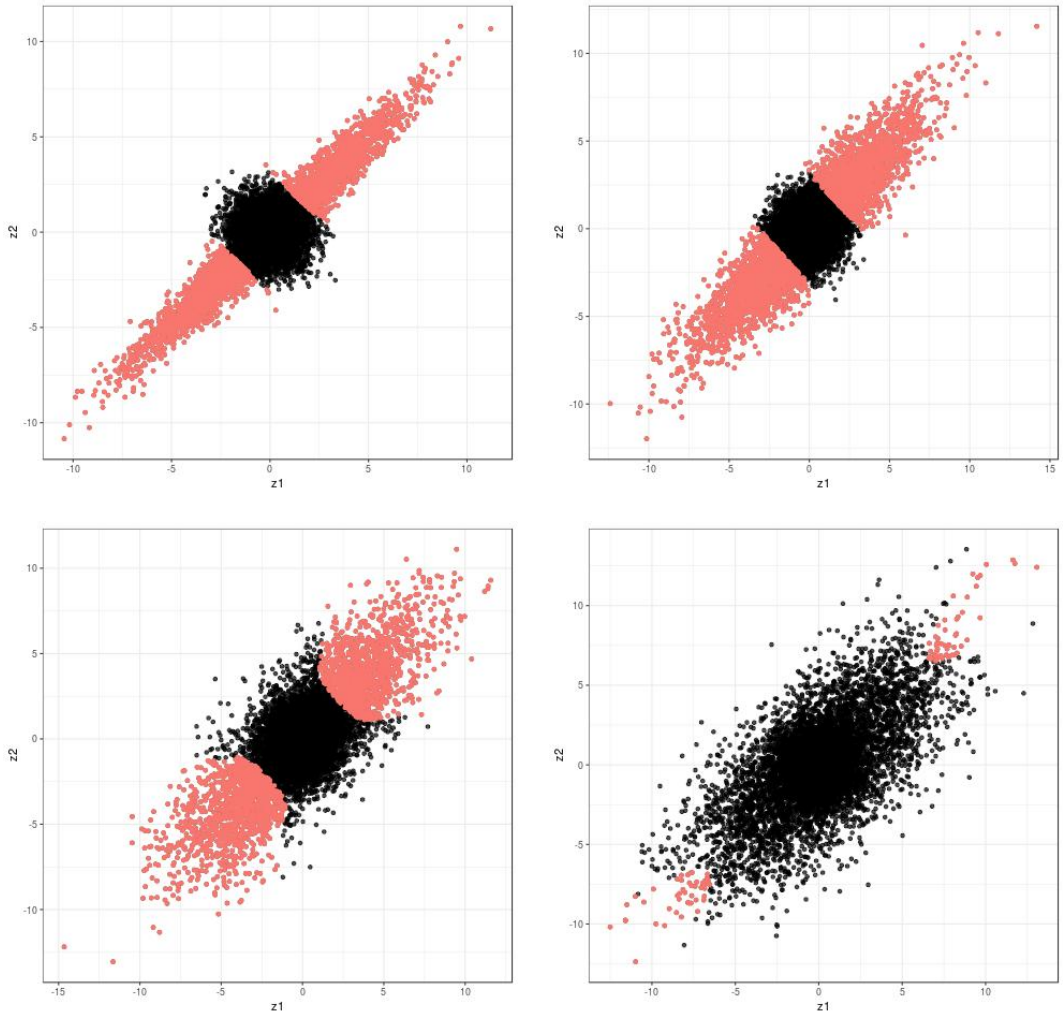

Figure 4.4: Scatter plots show the result of simulations studies from datasets generated with M being set as 8, when ϕs are set as 0.5,1,1.5 or 2 from the upper left plot to the lower right plot. Each black dot represents the pair of z-scores of each testing unit, which is overlaid by a red dot when the testing unit is categorized as reproducible signals after the FDR control.

Real Data Analysis

We applied the proposed method to single-SNP muscle eQTLs from FUSION skeletal muscle and the GTEx project (*Scott et al.*, 2016, *GTEx Consortium*, 2017). From these datasets, we used 191,057,404 gene-SNP pairs and 19,037 unique genes that exist in both studies. In this application, we aim to identify genes that have at least one causal SNP associated with, i.e. eGenes, that are reproducible in both studies. To achieve this purpose, we first calculated the posterior probabilities of

SNP being reproducible signal for each gene-SNP pair. We then summed it over across all SNPs within the testing windows for each gene, to calculate the posterior probability of reproducible eGene. The details on the real data analysis can be found in Section C.1.

After FDR control, we find out that 11,025 genes are identified as reproducible eGenes among 18,946 tested genes. Figure 4.5 depicts the overall results in forms of a rank copula plot and a histogram. The left plot is the same rank copula plot with the left plot of Figure 4.1, overlaid by red dots representing the reproducible eGenes. Compared with Figure 4.1, the left plot of Figure 4.5 shows that most genes on the upper right part are identified as eGenes. This confirms the common belief that most of eGenes with strong signals are reproducible between same tissues.

The right plot of Figure 4.5 strengthens the conclusion. since the most of identified reproducible eGenes have PP_{irr}s close to 0, as the red bar in the most left side in the histogram has the highest number of counts.

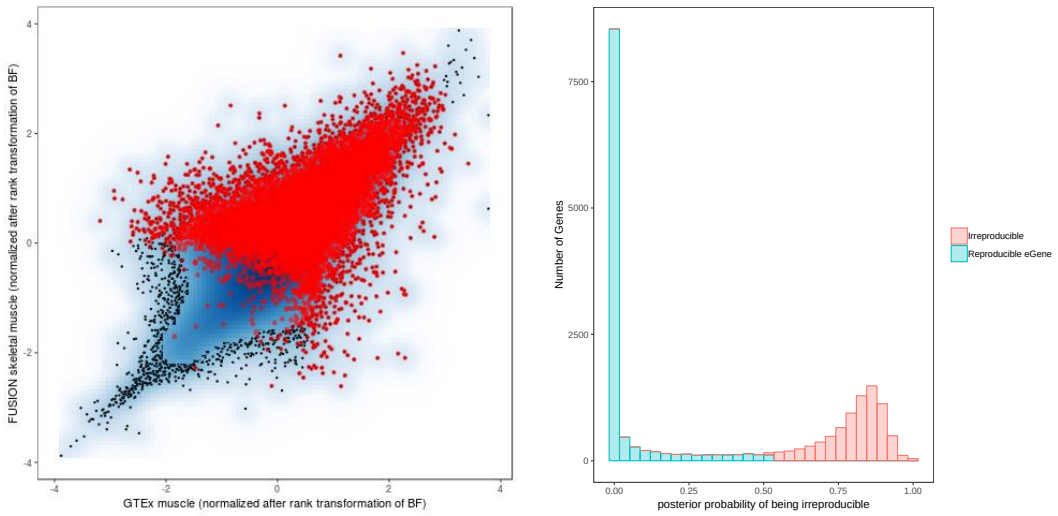

Figure 4.5: Left plot shows the rank copula plot of eGene Discoveries from FUSION skeletal muscle and GTEx muscle tissues. Red-colored dots are identified reproducible eGene in both studies, after the FDR control. Right plot is a histogram of PP_{irr}s, the posterior probabilities of being irreproducible. Red bars display reproducible eGenes and green bars display irreproducible genes.

4.4 Discussion

In association studies utilizing high-throughput data and assays, it is important to assess the degree of concordance between studies. Quantifying the degree of concordance can be utilized for measuring the quality of data processing procedure in high-throughput assays. It also can provide valuable information on identifying the evidence of strong and genuine signals from association studies. While there have been several methods to measure reproducibility between studies, none of the existing methods utilizes information from the directional consistency of estimated effects.

In this chapter, we first provide a visualization tool to demonstrate the degree of concordance between two studies, and quantify it using the empirical KL divergence. After that, we have proposed the method to measure reproducibility between studies under the Bayesian framework. Unlike the existing rank-based methods, the proposed method takes account of the direction consistency of estimated effects from studies. We assume that a strong and genuine signal should have the same direction of effects across studies, but in practice, there are some heterogeneity due to biological variations. By limiting this heterogeneity giving a set of priors, the proposed method can classify reproducible signals from irreproducible signals or non-signals.

The proposed method can successfully identify reproducible signals, as demonstrated by simulation studies and real data applications. Especially, the simulation studies show that the method distinguishes reproducible signals from others in various settings of effect sizes and the heterogeneity between studies.

However, there are some limitations utilizing the proposed visualization tool and method. For example, we find that the computed values of KL and $MI - KL$ are

sensitive to the choice of reference distributions and the size of bins in calculating empirical densities. To compare these values between different sets of studies, it is recommended to use the same reference distributions and the same size of bins.

We also find that the proposed method is sensitive to the choice of grid points of parameters k and ω and the single best way to set grid points is unknown. Especially, if we set the maximum value k as a huge value and consider a large number of grid points, it requires huge computational resources and makes computational infeasible in practice. While it can be avoided by bounding the limit of k, the better way to set the grid points should be explored in further study.

APPENDICES

APPENDIX A

Appendix of Chapter 2

A.1 Selection of Priority SNPs in Adaptive DAP

We give a detailed account of the Bayesian conditional analysis procedure for selecting high-priority SNPs in the adaptive DAP algorithm. For a given locus l, the procedure starts with model size partition $s = 1$. Let $\vec{\gamma}^*$ denote the model with the highest posterior probability in the size partition $s - 1$ in locus l, i.e.,

$$\vec{\gamma}^* = \text{argmax}_{\{||\vec{\gamma}||=s-1\}} \Pr(\vec{\gamma}_l = \vec{\gamma}) \text{BF}(\vec{\gamma}).$$

For each SNP i that is *not* included in the current best model, we compute a Bayes factor for the expanded model, $\vec{\gamma}_i^\dagger = \vec{\gamma}^* \cup \{\gamma_{l_i} = 1\}$. Assuming that there is exactly one additional QTL and that 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

$$\text{PIP}_i^* = \frac{\text{BF}(\vec{\gamma}_i^\dagger)/\text{BF}(\vec{\gamma}^*)}{\sum_j \text{BF}(\vec{\gamma}_j^\dagger)/\text{BF}(\vec{\gamma}^*)} = \frac{\text{BF}(\vec{\gamma}_i^\dagger)}{\sum_j \text{BF}(\vec{\gamma}_j^\dagger)}. \tag{A.1.1}$$

The resulting quantity is a well-defined posterior probability and is solely determined by the relative likelihood values of the expanded models. In particular, it should be

noted that (A.1.1) fully accounts for LD between SNPs: e.g., if two SNPs are in perfect LD, they would possess identical values that correctly reflect the uncertainty (i.e., they are indistinguishable). The procedure requires $p - s$ evaluations of Bayes factors that 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 $\text{PIP}_i^* \geq \lambda$. For $s \geq 2$, we then enumerate all s-combinations from the resulting set of priority SNPs to compute C_s^*. During this enumeration, we also record the new $\vec{\boldsymbol{\gamma}}^*$ for the increased model size.

Intuitively, the threshold parameter λ is related to the precision of the approximate PIPs. The selection procedure roughly estimates the probability, $\Pr(\gamma_{l_i} = 1 \mid \vec{\boldsymbol{y}}, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}, ||\vec{\boldsymbol{\gamma}}_l|| = s)$, for SNP i. Note the relationship

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

The following can be concluded:

1. If $\Pr(\gamma_{l_i} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}, ||\vec{\boldsymbol{\gamma}}_l|| = s) < \lambda$ for a given SNP at all s values, then it must be the case that the overall PIP $< \lambda$.

2. The loss of precision of the PIP of SNP i due to the selection screening for a particular size partition must be $< \lambda$.

A.2 Stopping Rule and Estimation of the Approximation Error in Adaptive DAP

When a non-associated SNP is added to 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

unnecessary explicit explorations of the model partitions once the sizes of the models are considered saturated. 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 of the unexplored size partitions (i.e., the approximation error) is then estimated by an analytic combinatorial approximation.

To explain the stopping rule and the combinatorial approximation, we assume that there are K detectable true QTNs. 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 QTNs (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 $\sum_{\vec{\gamma}} \Pr(\vec{\gamma}_l = \vec{\gamma}; ||\vec{\gamma}_l|| = s)\mathrm{BF}(\vec{\gamma})$ from this particular category by the equation

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

where $\widetilde{\Pr}(\vec{\gamma}_l; ||\vec{\gamma}_l|| = s)$ 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 QTNs. 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, increase the marginal likelihood/Bayes factor. Hence, we obtain

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

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

(A.2.1)

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

In the last step, we approximate the quantities $\frac{\widetilde{\Pr}(\vec{\gamma}_l;||\vec{\gamma}_l||=s+1)}{\widetilde{\Pr}(\vec{\gamma}_l;||\vec{\gamma}_l||=s)}$ 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}

(A.2.2)
$$\frac{p-s-K}{s+1}\,\omega\, C_s.$$

Thus, we have shown

(A.2.3)
$$\frac{p-s}{s-K+1}\,\omega\, C_s \gtrsim C_{s+1} \gtrsim \frac{p-s-K}{s+1}\,\omega\, C_s.$$

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

(A.2.4)
$$C_{s+1} \approx \frac{p-s}{s+1}\,\omega\, C_s,$$

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 this approximation is surprisingly accurate (Figure S3).

Our stopping rule is built upon the upper bound specified by the inequality (A.2.3). Specially, the adaptive DAP stops explicit exploration at partition size $s = t$ if

(A.2.5)
$$C_t^* \leq (p-t+1)\,\omega\, C_{t-1}^*.$$

The inequality essentially tests $K \geq t-1$. 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 a high accuracy of the approximation, we also add an optional criterion to the stopping rule on top of (A.2.5), 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 make no substantial contributions to the normalizing constant. This additional criterion provides practical flexibility for running the DAP: as $\kappa \to 0$, it enforces the DAP to explore all the model size partitions, whereas when κ is large, only the stopping rule (A.2.5) 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

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

A.3 Derivation of the DAP-1 Algorithm

In this section, we provide a detailed derivation for the DAP-1 algorithm. It should be noted that the derivation can be generalized to the 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 $\{\vec{\gamma}\}$, i.e.,

$$C - \sum_{||\vec{\gamma}||\leq 1} \Pr(\vec{\gamma}_l = \vec{\gamma})\mathrm{BF}(\vec{\gamma}) \to 0. \tag{A.3.1}$$

Consequently, it follows that

$$\Pr(\vec{\gamma}_l = \vec{\gamma} \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}) \approx \begin{cases} \frac{\Pr(\vec{\gamma}_l=\vec{\gamma}|\vec{\boldsymbol{\alpha}})\mathrm{BF}(\vec{\gamma})}{\sum_{||\vec{\gamma}'||\leq 1}\Pr(\vec{\gamma}_l=\vec{\gamma}')\mathrm{BF}(\vec{\gamma}')} & \text{if } ||\vec{\gamma}|| \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

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

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

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

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

and

$$\mathrm{BF}(\vec{\gamma}_j^{\circ}) = \mathrm{BF}_j.$$

We 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., the estimated genetic effect of the target SNP and its standard error *Wakefield* (2009), *Wen et al.* (2014), and it is computationally trivial.

Finally, we note that given the restrained model space, the PIP of SNP j, $\Pr(\gamma_j \mid \vec{\boldsymbol{y}}, \boldsymbol{G}, \vec{\boldsymbol{\alpha}})$, coincides with $\Pr(\vec{\gamma}_j^{\circ} \mid \vec{\boldsymbol{\alpha}})$. Given all of the above, it follows from simple

algebra that

$$\begin{aligned}\Pr(\gamma_i = 1 \mid \vec{\boldsymbol{y}}, \boldsymbol{G}, \vec{\boldsymbol{\alpha}}) &= \frac{\sum_{k=1}^{p} e^{\alpha_0 + \sum_{l=1}^{q} \alpha_l d_{kl}} \mathrm{BF}_k}{1 + \sum_{k=1}^{p} e^{\alpha_0 + \sum_{l=1}^{q} \alpha_l d_{kl}} \mathrm{BF}_k} \cdot \frac{e^{\sum_{l=1}^{q} \alpha_l d_{il}} \mathrm{BF}_i}{\sum_{k=1}^{p} e^{\sum_{l=1}^{q} \alpha_l d_{kl}} \mathrm{BF}_k} \\ &= \left[1 - \Pr(\vec{\boldsymbol{\gamma}}_l = \mathbf{0} \mid \vec{\boldsymbol{y}}, \boldsymbol{G}, \vec{\boldsymbol{\alpha}})\right] \cdot \frac{e^{\sum_{l=1}^{q} \alpha_l d_{il}} \mathrm{BF}_i}{\sum_{k=1}^{p} e^{\sum_{l=1}^{q} \alpha_l d_{kl}} \mathrm{BF}_k},\end{aligned} \tag{A.3.2}$$

where the first term assesses 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 (A.3.2) bears great similarity to the previously proposed Bayesian approaches *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 does not contain a QTL. Conditioning on the case that the locus does contain a QTL, the prior for SNP i being the causal SNP is assigned

$$\Pr(\gamma_i = 1 \mid \vec{\boldsymbol{\gamma}}_l \neq 0, \vec{\boldsymbol{\delta}}) = \frac{e^{\sum_{l=1}^{q} \delta_l d_{il}}}{\sum_{k=1}^{p} e^{\sum_{l=1}^{q} \delta_l d_{kl}}}, \tag{A.3.3}$$

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

$$\Pr(\gamma_i = 1 \mid \vec{\boldsymbol{y}}, \boldsymbol{G}_l, \pi_0, \vec{\boldsymbol{\delta}}) = \left[1 - \Pr(\vec{\boldsymbol{\gamma}}_l = \mathbf{0} \mid \vec{\boldsymbol{y}}, \boldsymbol{G}_l, \pi_0)\right] \cdot \frac{e^{\sum_{l=1}^{q} \delta_l d_{il}} \mathrm{BF}_i}{\sum_{k=1}^{p} e^{\sum_{l=1}^{q} \delta_l d_{kl}} \mathrm{BF}_k}. \tag{A.3.4}$$

Despite the algebraic similarity, the parameters (π_0 and $\vec{\boldsymbol{\delta}}$) in (A.3.4) cannot be directly interpreted as $\vec{\boldsymbol{\alpha}}$ in our logistic priors, partly due to the conditional nature of the prior specification (A.3.3). Furthermore, in enrichment analysis, the M-step of the EM algorithm becomes much more involved for optimizing the objective function jointly with respect to $(\pi_0, \vec{\boldsymbol{\delta}})$. 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.

A.4 Factorization of the posterior probability by LD blocks

For integrative association analysis for loci spanning very large genomic regions, especially in GWAS settings, we recommend an additional approximate factorization, $\Pr(\vec{\boldsymbol{\gamma}} \mid \vec{\boldsymbol{y}}, \boldsymbol{G}, \vec{\boldsymbol{\alpha}}) \approx \sum_{k=1}^{L} \Pr(\vec{\boldsymbol{\gamma}}_{[k]} \mid \vec{\boldsymbol{y}}, \boldsymbol{G}, \vec{\boldsymbol{\alpha}})$, before applying the DAP to each genomic region independently. We provide the necessary mathematical justification for this factorization.

It is sufficient to show that

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

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

To show that

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

we note the previous result on the Bayes factors *Wen* (2014),

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

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

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

If $\gamma_i = 1$, β_i is assigned a normal prior, whereas if $\gamma_i = 0$, $\beta_i = 0$ with probability 1 (or is represented by a degenerated normal distribution, $\beta_i \sim \mathrm{N}(0, 0)$). Equivalently, we write

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

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

Without loss of generality, we assume that both the phenotype vector $\vec{\boldsymbol{y}}$ and the genotype vectors $\vec{\boldsymbol{g}}_1, ..., \vec{\boldsymbol{g}}_p$ 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 *Wen* (2014) that

$$\text{(A.4.1)} \quad \text{BF}(\vec{\boldsymbol{\beta}}; \boldsymbol{W}) = |\boldsymbol{I} + \tau \boldsymbol{G}'\boldsymbol{G}\boldsymbol{W}|^{-\frac{1}{2}} \cdot \exp\left(\frac{1}{2}\vec{\boldsymbol{y}}'\boldsymbol{G}\left[\boldsymbol{W}(\boldsymbol{I} + \tau \boldsymbol{G}'\boldsymbol{G}\boldsymbol{W})^{-1}\right]\boldsymbol{G}'\vec{\boldsymbol{y}}\right).$$

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 $\text{Var}(\boldsymbol{G})$. In other words, $\boldsymbol{G}'\boldsymbol{G}$ can be viewed as a noisy observation of $n\text{Var}(\boldsymbol{G})$. Using population genetic theory, Wen and Stephens *Wen and Stephens* (2010) show that $\text{Var}(\boldsymbol{G})$ is extremely banded. Based on this result, Berisa and Pickrell *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

$$\text{(A.4.2)} \qquad \widehat{\boldsymbol{G}'\boldsymbol{G}} = \boldsymbol{G}'_{[1]}\boldsymbol{G}_{[1]} \oplus \cdots \oplus \boldsymbol{G}'_{[L]}\boldsymbol{G}_{[L]},$$

where "$\oplus$" denotes the direct sum of the matrices. It is important to note that (A.4.2) 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 (A.4.2) into (A.4.1), it follows that

$$\text{(A.4.3)} \qquad \text{BF}(\vec{\boldsymbol{\beta}}; \boldsymbol{W}) = \prod_{k=1}^{L} \text{BF}_{[k]},$$

where

(A.4.4)

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

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, we integrate out the residual error variance parameter τ for each $\mathrm{BF}_{[k]}$ by applying the Laplace approximation *Wen* (2014). This step results in plugging in a point estimate of τ (e.g., based on $\vec{\boldsymbol{y}}$ and $\boldsymbol{G}_{[k]}$ for each block k) into the expression (A.4.4). Taken together, we have shown that

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

and consequently,

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

A.5 Average Accuracy of PIP Estimates using DAP-1

In this section, we provide some mathematical arguments to justify that DAP-1 (or adaptive DAP with less stringent threshold values) algorithm can provide *on average* accurate estimate. Specifically, we write the expression for the exact calculation of the PIP for SNP k at locus l as

$$\Pr(\gamma_{l_k} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}) = \sum_{s=1}^{p} \frac{C_i}{C} \cdot \Pr(\gamma_{l_k} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}, ||\vec{\boldsymbol{\gamma}}_l|| = s). \tag{A.5.1}$$

In the case of DAP-1, we essentially use the following expression to approximate the PIP,

$$\Pr(\gamma_{l_k} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}) \approx \frac{C_1}{C_0 + C_1} \cdot \Pr(\gamma_{l_k} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}, ||\vec{\boldsymbol{\gamma}}_l|| = 1). \tag{A.5.2}$$

Note that in genetic association analysis, the vast majority of SNPs have overall PIPs $\to 0$ within any given locus; hence, it must be the case that for such a SNP k,

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

Therefore, even $C_1 + C_0$ approximates C poorly, and (A.5.2) still provides an adequately accurate PIP estimation for the majority of SNPs that are not QTNs. The same argument can also be applied to candidate QTNs with very strong evidence for associations, especially when the "primary" association signals have strengths of associations that are orders of magnitude higher than the remaining candidate SNPs within a locus (e.g., $\Pr(\gamma_{l_k} = 1 \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \vec{\boldsymbol{\alpha}}, ||\vec{\boldsymbol{\gamma}}_l|| = s) \to 1$ for all s). Therefore, the only SNPs whose PIPs are poorly approximated by DAP-1 are those secondary QTL signals (if there are any), but in most practical cases, it can be assured that such SNPs are small in number.

APPENDIX B

Appendix of Chapter 3

B.1 Details on Bayesian hierarchical linear model

Recall that we assume the linear model (3.1),

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

and define $\gamma_i = \mathbf{1}(\beta_i \neq 0)$. The γ_i's are assumed independent *a priori* with the following prior distribution,

$$\gamma_i \sim \text{Bernoulli}\,(\eta_i). \tag{B.1.1}$$

In case that an m-dimensional annotation, $\vec{\boldsymbol{\delta}}_i$, is available for each SNP i, we incorporate this quantitative information into the prior specification through a logistic function, i.e.,

$$\text{logit}(\eta_i) = \alpha_0 + \vec{\boldsymbol{\alpha}}' \vec{\boldsymbol{\delta}}_i. \tag{B.1.2}$$

We estimate the enrichment parameters $(\alpha_0, \vec{\boldsymbol{\alpha}})$ from the observed data using an EM algorithm detailed in *Wen* (2016). In the absence of the annotation data, the logistic

prior reduces to a single intercept term. The prior for the effect size parameter β_i is assumed the following form

$$\beta_i \mid \gamma_i = 1 \sim \sum_{k=1}^{K} \pi_k \mathrm{N}(0, \phi_k^2). \tag{B.1.3}$$

By including a grid of values for ϕ_k, this mixture prior attempts to capture a spectrum of genetic effect sizes ranging from modest to strong. By default, π_k is set to $1/K$. It is also possible to estimate individual π_k values by an EM algorithm (e.g., the one implemented in TORUS). The marginal priors on β_i's are known as spike-and-slab in the statistical literature.

Finally, we assume a Γ prior for the parameter τ that controls residual error variance in the linear model, i.e.,

$$\tau \sim \Gamma(\kappa/2, \lambda/2). \tag{B.1.4}$$

For inference, we assume the limiting form of this prior as $\kappa, \lambda \to 0$.

B.2 Inference using summary statistics

B.2.1 Sufficient statistics for likelihood computation

Our result is derived from the analytic expression of Bayes factors and their approximations in a general complex linear model system reported in *Wen* (2014), where the multiple linear regression model discussed in this paper is a trivial special case. Assuming for a given $\vec{\boldsymbol{\gamma}}$ value, the linear regression model (3.1) is reduced to q assumed associated SNPs (i.e., q entries of the $\vec{\boldsymbol{\gamma}}$ vector are 1) and we adjust $\boldsymbol{G}$ to denote the $q \times q$ design matrix specific to the value of $\vec{\boldsymbol{\gamma}}$. Let the q-vector $\vec{\boldsymbol{\beta}}$ denote the genetic effect sizes of the q SNPs. We assume a general prior distribution for $\vec{\boldsymbol{\beta}}$,

$$\vec{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}), \tag{B.2.1}$$

where $\boldsymbol{W}$ is a $q \times q$ positive semi-definite matrix. In this case, the Bayes factor can be computed by

$$\text{(B.2.2)} \quad \text{BF}(\boldsymbol{W}) = |\boldsymbol{I} + \boldsymbol{V}^{-1}\boldsymbol{W}|^{-\frac{1}{2}} \exp\left(\frac{1}{2}\hat{\vec{\beta}}'\boldsymbol{V}^{-1}\left[\boldsymbol{W}(\boldsymbol{I} + \boldsymbol{V}^{-1}\boldsymbol{W})^{-1}\right]\boldsymbol{V}^{-1}\hat{\vec{\beta}}\right),$$

For multiple linear regression model, we note that

$$\text{(B.2.3)} \quad \begin{aligned} \hat{\vec{\beta}} &= (\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\vec{\boldsymbol{y}}, \\ \boldsymbol{V}^{-1} &= \tau\boldsymbol{G}'\boldsymbol{G}, \end{aligned}$$

and (B.2.2) can be simplified to

$$\text{(B.2.4)} \quad \text{BF}(\boldsymbol{W}) = |\boldsymbol{I} + \tau\boldsymbol{G}'\boldsymbol{G}\boldsymbol{W}|^{-\frac{1}{2}} \exp\left(\frac{\tau^2}{2}\vec{\boldsymbol{y}}'\boldsymbol{G}\left[\boldsymbol{W}(\boldsymbol{I} + \tau\boldsymbol{G}'\boldsymbol{G}\boldsymbol{W})^{-1}\right]\boldsymbol{G}'\vec{\boldsymbol{y}}\right).$$

If τ is known, the above expression is exact, and the computation relies on the observed data only through the summary statistics $(\boldsymbol{G}'\vec{\boldsymbol{y}}, \boldsymbol{G}'\boldsymbol{G})$.

When τ is unknown, *Wen* (2014) shows the above analytic form becomes an approximation via Laplace's method, i.e.,

(B.2.5)

$$\text{BF}(\boldsymbol{W}) = |\boldsymbol{I} + \check{\tau}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{W}|^{-\frac{1}{2}} \exp\left(\frac{\check{\tau}^2}{2}\vec{\boldsymbol{y}}'\boldsymbol{G}\left[\boldsymbol{W}(\boldsymbol{I} + \check{\tau}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{W})^{-1}\right]\boldsymbol{G}'\vec{\boldsymbol{y}}\right) \cdot \left[1 + o\left(\frac{1}{n}\right)\right].$$

In particular, the point estimate $\check{\tau}$ is an affine combination of the MLEs of τ estimated under the null model (denoted by $\tilde{\tau}$) and the full model (denoted by $\hat{\tau}$). More specifically,

$$\text{(B.2.6)} \quad \begin{aligned} \tilde{\tau} &= \frac{n}{\vec{\boldsymbol{y}}'\vec{\boldsymbol{y}}} \\ \hat{\tau} &= \frac{n}{(\vec{\boldsymbol{y}} - \boldsymbol{G}\hat{\vec{\beta}})'(\vec{\boldsymbol{y}} - \boldsymbol{G}\hat{\vec{\beta}})} = \frac{n}{\vec{\boldsymbol{y}}'\vec{\boldsymbol{y}} - \vec{\boldsymbol{y}}'\boldsymbol{G}\,(\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\vec{\boldsymbol{y}}} \\ \check{\tau} &= \alpha\tilde{\tau} + (1-\alpha)\hat{\tau},\ 0 \le \alpha \le 1. \end{aligned}$$

In other words, plugging in any value between $\hat{\tau}$ and $\tilde{\tau}$ for τ corresponds to a valid Laplace approximation of $\text{BF}(\boldsymbol{W})$. Note that this result is essential for the justification of the use of single-SNP testing z-scores when τ is unknown.

In addition to $\boldsymbol{G}'\vec{\boldsymbol{y}}$ and $\boldsymbol{G}'\boldsymbol{G}$, estimating $\tilde{\tau}$ and/or $\hat{\tau}$ requires two more summary statistics, sample size n and SST $= \vec{\boldsymbol{y}}'\vec{\boldsymbol{y}}$, when τ is unknown. Thus, we conclude that the sufficient statistics required to compute the Bayes factors are $(\boldsymbol{G}'\vec{\boldsymbol{y}}, \boldsymbol{G}'\boldsymbol{G}, n, \text{SST})$.

B.2.2 Recovering sufficient statistics

When τ is known, single SNP testing z-statistics along with $\boldsymbol{R}$ and $\boldsymbol{\Lambda}$ are sufficient to recover the required sufficient statistics for Bayes factor computation, i.e.,

$$\begin{aligned} \boldsymbol{G}'\boldsymbol{G} &= \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{\Lambda} \\ \boldsymbol{G}'\vec{\boldsymbol{y}} &= \tau^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{Z}. \end{aligned} \tag{B.2.7}$$

Here we focus our discussion on recovering the sufficient statistics in a realistic setting where τ is not known. In particular, we assume that for each SNP i, the effect size estimate

$$\hat{b}_i = \frac{\vec{\boldsymbol{g}}_i'\vec{\boldsymbol{y}}}{\vec{\boldsymbol{g}}_i'\vec{\boldsymbol{g}}}, \tag{B.2.8}$$

and its standard error, $\hat{s}_i = \text{se}(\hat{b}_i)$. Additionally, we only assume the knowledge of $\boldsymbol{R}$ (but not $\boldsymbol{\Lambda}$), n and SST .

We show the following procedure can recover required sufficient statistics assuming $\boldsymbol{R}$ is accurate. For each SNP i,

1. Compute $z_i = \hat{b}_i/\hat{s}_i$
2. Compute R^2 for the corresponding simple linear regression model by

$$R_i^2 = \frac{z_i^2}{z_i^2 + n - 2}$$

3. Find the estimated residual error variance from the corresponding simple linear regression model by

$$\hat{\sigma}_i^2 = \text{SST}\,(1 - R_i^2)/(n-2)$$

.

4. Compute $\vec{\boldsymbol{g}}_i'\vec{\boldsymbol{g}}_i = (1/\hat{s}_i)^2\,\hat{\sigma}_i^2$

5. Compute $\vec{\boldsymbol{g}}_i'\vec{\boldsymbol{y}} = \hat{b}_i \cdot \vec{\boldsymbol{g}}_i'\vec{\boldsymbol{g}}$

Subsequently, we obtain that

$$\boldsymbol{G}'\vec{\boldsymbol{y}} = \begin{bmatrix} \vec{\boldsymbol{g}}_1'\vec{\boldsymbol{y}} \\ \vdots \\ \vec{\boldsymbol{g}}_q'\vec{\boldsymbol{y}} \end{bmatrix}, \tag{B.2.9}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(\sqrt{\vec{\boldsymbol{g}}_1'\vec{\boldsymbol{g}}_1}, ..., \sqrt{\vec{\boldsymbol{g}}_q'\vec{\boldsymbol{g}}_q}), \tag{B.2.10}$$

and

$$\boldsymbol{G}'\boldsymbol{G} = \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{\Lambda}.$$

Consequently, any appropriate form of τ estimate can be obtained.

B.2.3 Connection to previous results

In this section, we show that our results are connected to the existing literature, assuming τ is known.

Result for known τ Assume that $\boldsymbol{W}$ is full-rank, it follows that

$$\mathrm{BF}(\boldsymbol{W}) = |\boldsymbol{I} + \tau\boldsymbol{G}'\boldsymbol{G}\boldsymbol{W}|^{-\frac{1}{2}}\,\exp\left(\frac{\tau^2}{2}\vec{\boldsymbol{y}}'\boldsymbol{G}\left(\boldsymbol{W}^{-1} + \tau\boldsymbol{G}'\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\vec{\boldsymbol{y}}\right). \tag{B.2.11}$$

Plugging in Equation (B.2.7) results in

$$\mathrm{BF}(\boldsymbol{W}) = |\boldsymbol{I} + \tau\boldsymbol{G}'\boldsymbol{G}\boldsymbol{W}|^{-\frac{1}{2}}\,\exp\left(\frac{1}{2}\boldsymbol{Z}'\left[(\tau\boldsymbol{\Lambda}\boldsymbol{W}\boldsymbol{\Lambda})^{-1} + \boldsymbol{R}\right]^{-1}\hat{\vec{z}}\right). \tag{B.2.12}$$

In particular, *Chen et al.* (2015) uses a specific form of prior, which scales the effect size of each SNP by its genotype variance and τ, namely,

$$\boldsymbol{W} = \frac{n\phi^2}{\tau}\boldsymbol{\Lambda}^{-2}. \tag{B.2.13}$$

It follows from the Sylvester's determinant theorem that

$$|\boldsymbol{I} + \tau \boldsymbol{G}'\boldsymbol{G}\boldsymbol{W}| = |\boldsymbol{I} + (n\phi^2)\boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{\Lambda}^{-1}| = |\boldsymbol{I} + (n\phi^2)\boldsymbol{R}|, \tag{B.2.14}$$

and

$$\mathrm{BF}(\phi^2) = |\boldsymbol{I} + (n\phi^2)\boldsymbol{R}|^{-\frac{1}{2}} \exp\left(\frac{1}{2}\vec{\boldsymbol{z}}' \left[(n\phi^2)^{-1}\boldsymbol{I} + \boldsymbol{R}\right]^{-1} \vec{\boldsymbol{z}}\right), \tag{B.2.15}$$

which only requires $(\boldsymbol{R}, \vec{\boldsymbol{z}})$ and is exactly the same result presented by *Chen et al.* (2015).

Result for unknown τ Here we justify the use of the analytic form of Equation (B.2.15) when τ is not known. With the specific prior (B.2.13), the Bayes factor for a given τ value is

$$\mathrm{BF}(\phi^2; \tau) = |\boldsymbol{I} + (n\phi^2)\boldsymbol{R}|^{-\frac{1}{2}} \exp\left(\frac{\tau}{2}\vec{\boldsymbol{y}}'\boldsymbol{G}\boldsymbol{\Lambda}^{-1} \left[(n\phi^2)^{-1}\boldsymbol{I} + \boldsymbol{R}\right]^{-1} \boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\vec{\boldsymbol{y}}\right). \tag{B.2.16}$$

Wen (2014) shows that the desired Bayes factor with respect to arbitrary prior density $p(\tau)$ can be approximated by the Laplace's method. The resulting approximation is given by

$$\mathrm{BF}(\phi^2) = \mathrm{BF}(\phi^2; \check{\tau}) \cdot \left[1 + o\left(\frac{1}{n}\right)\right], \tag{B.2.17}$$

where $\check{\tau}$ can be any affine combination of $\tilde{\tau}$ (the MLE of τ from the null model) and $\hat{\tau}$ (the MLE of τ from the full model). Note that the quadratic form,

$$\vec{\boldsymbol{y}}'\boldsymbol{G}\boldsymbol{\Lambda}^{-1} \left[(n\phi^2)^{-1}\boldsymbol{I} + \boldsymbol{R}\right]^{-1} \boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\vec{\boldsymbol{y}},$$

is positive definite, the approximate Bayes factor (ABF) is monotonically increasing with respect to the value of $\check{\tau}$. More specifically, all valid ABFs justified by this approximation satisfy

$$\mathrm{BF}(\phi^2; \tilde{\tau}) \leq \mathrm{BF}(\phi^2; \check{\tau}) \leq \mathrm{BF}(\phi^2; \hat{\tau}). \tag{B.2.18}$$

Alternatively, we can represent the ABF result as a function of multi-dimensional z-scores. First, we define a p-vector $\vec{\boldsymbol{\tau}} := (\tau_1, \tau_2, ..., \tau_p)$, and denote

$$\begin{aligned} \hat{\vec{\tau}} &= (\hat{\tau}, ..., \hat{\tau}) \\ \tilde{\vec{\tau}} &= (\hat{\tau}, ..., \tilde{\tau}) \end{aligned} \tag{B.2.19}$$

Let $\boldsymbol{T}(\vec{\boldsymbol{\tau}}) := \text{diag}(\vec{\boldsymbol{\tau}})$, $\vec{\boldsymbol{z}}(\vec{\boldsymbol{\tau}}) := \boldsymbol{T}(\vec{\boldsymbol{\tau}})^{\frac{1}{2}} \boldsymbol{\Lambda} \boldsymbol{G}' \vec{\boldsymbol{y}}$, and we define

$$\text{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \vec{\boldsymbol{\tau}}) = |\boldsymbol{I} + (n\phi^2)\boldsymbol{R}|^{-\frac{1}{2}} \exp\left(\frac{1}{2} \vec{\boldsymbol{z}}(\vec{\boldsymbol{\tau}})' \left[(n\phi^2)^{-1} \boldsymbol{I} + \boldsymbol{R} \right]^{-1} \vec{\boldsymbol{z}}(\vec{\boldsymbol{\tau}}) \right). \tag{B.2.20}$$

Here we attempt to link the analytic expression, $\text{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \check{\vec{\tau}})$, to the well-defined approximate Bayes factor $\text{BF}(\phi^2; \check{\tau})$.

Let $f(\vec{\boldsymbol{z}}) = \vec{\boldsymbol{z}}' \left[(n\phi^2)^{-1} \boldsymbol{I} + \boldsymbol{R} \right]^{-1} \vec{\boldsymbol{z}} = \vec{\boldsymbol{z}}' \boldsymbol{A} \vec{\boldsymbol{z}}$, it follows that

$$\frac{\partial f}{\partial |z_i|} = \frac{\partial f}{\partial \vec{\boldsymbol{z}}} \frac{\partial |\vec{\boldsymbol{z}}|}{\partial z_i} = 2A_{ii}|z_i|, \ i = 1, 2, ..., p. \tag{B.2.21}$$

Because matrix $\boldsymbol{A}$ is positive definite, it follows that $A_{ii} = \vec{\boldsymbol{e}}_i' \boldsymbol{A} \vec{\boldsymbol{e}}_i > 0, \ \forall i$, where $\vec{\boldsymbol{e}}_i$ denotes the unit vector with the i-th entry being set to 1. Equation (B.2.21) indicates that $f(\vec{\boldsymbol{z}})$ is monotonically increasing with respect to each individual $|z_i|$.

In practice, $\check{\vec{z}} := \vec{\boldsymbol{z}}(\check{\vec{\tau}}) = \boldsymbol{T}^{\frac{1}{2}} \boldsymbol{\Lambda}^{-1} \boldsymbol{G}' \vec{\boldsymbol{y}}$ is used to evaluate $\text{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \check{\vec{\tau}})$, where $\boldsymbol{T} = \text{diag}(\check{\tau}_1, \ldots, \check{\tau}_p)$ and each $\check{\tau}_i$ represents the MLE of τ estimated from the simple regression model testing the association of SNP i. It should be clear that

$$\hat{\tau} \geq \check{\tau}_i \geq \tilde{\tau}, \ \forall i. \tag{B.2.22}$$

Let $\hat{\vec{z}} := \vec{\boldsymbol{z}}(\hat{\vec{\tau}})$ and $\tilde{\vec{z}} := \vec{\boldsymbol{z}}(\tilde{\vec{\tau}})$, it follows that

$$|\hat{z}_i| \geq |\check{z}_i| \geq |\tilde{z}_i|, \ \forall i, \tag{B.2.23}$$

Consequently, it implies that

$$\text{BF}(\phi^2; \hat{\tau}) \equiv \text{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \hat{\vec{\tau}}) \geq \text{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \check{\vec{\tau}}) \geq \text{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \tilde{\vec{\tau}}) \equiv \text{BF}(\phi^2; \tilde{\tau}). \tag{B.2.24}$$

By the intermediate value theorem, there must exists $0 \leq \alpha \leq 1$, and

$$\check{\tau} = \alpha\tilde{\tau} + (1-\alpha)\hat{\tau}, \tag{B.2.25}$$

such that

$$\mathrm{BF}(\phi^2; \check{\tau}) = \mathrm{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \check{\vec{\tau}}). \tag{B.2.26}$$

Therefore, $\mathrm{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \check{\vec{\tau}})$ is valid approximation of Bayes factor under the prior (B.2.13) by the argument of Laplace's method.

Numerical Illustration

If the association model under consideration (i.e., $\vec{\boldsymbol{\gamma}}$) contains no true association signal, or the genetic effects of the suspected associations are small, we expect that $\hat{\tau} \approx \check{\tau} \approx \tilde{\tau}$. As a result, we also expect

$$\mathrm{BF}(\phi^2; \hat{\tau}) \approx \mathrm{BF}_{\vec{\boldsymbol{z}}}(\phi^2; \check{\vec{\tau}}) \approx \mathrm{BF}(\phi^2; \tilde{\tau}). \tag{B.2.27}$$

To illustrate, we simulate a quantitative traits for 343 individuals with 3 independent genetic variants, i.e.,

$$y_i = 0.05x_1 + 0.05x_2 + 0.05x_3 + e_i,\ e_i \sim \mathrm{N}(0,1). \tag{B.2.28}$$

Assuming $\phi^2 = 1$, the comparison of the three approximate Bayes factors is shown in Table A2.

In an alternative scenario where the data contain multiple modest to strong association signals, directly applying the z-score approximation can result in an equivalent $\check{\tau}$ value that severely over-estimates residual error, hence under-estimate the Bayes factor. To illustrate, we use the same simulated genotype data and the following linear model,

$$y_i = 0.25x_1 + 0.25x_2 + 0.25x_3 + e_i,\ e_i \sim \mathrm{N}(0,1). \tag{B.2.29}$$

As expected, the resulting approximate Bayes factors shows difference in order of magnitude (Table A3), albeit all approximations show overwhelming evidence of association.

B.2.4 Figure Legends

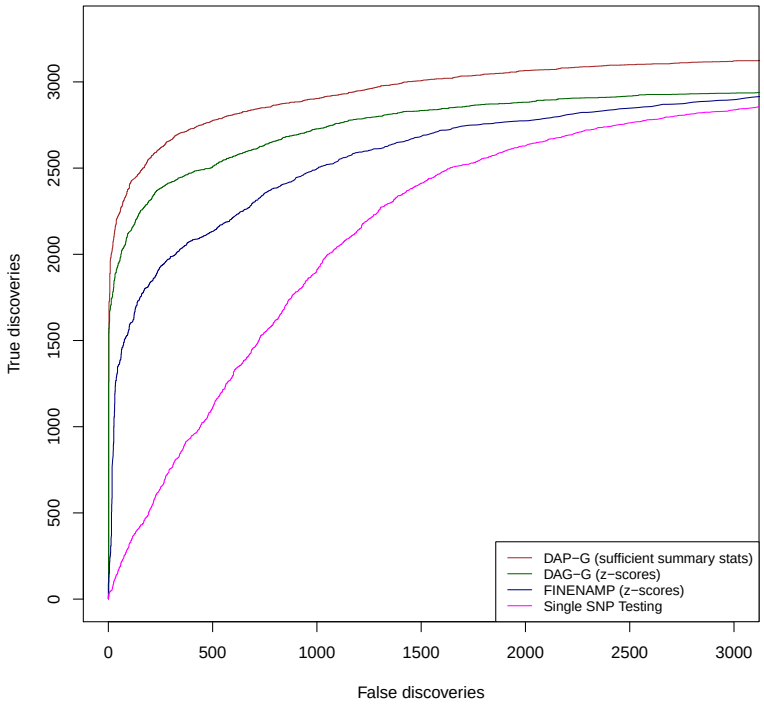

Figure A1: Power comparison in simulation studies. We examined the performance of 4 different methods in identifying the LD blocks that harbor true association signals. The methods compared include DAP-G using sufficient summary statistics (brown line), DAP-G using single SNP testing z-scores (dark green line), FINEMAP using single SNP testing z-score (navy blue line) and the single-SNP testing approach (magenta line). 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.

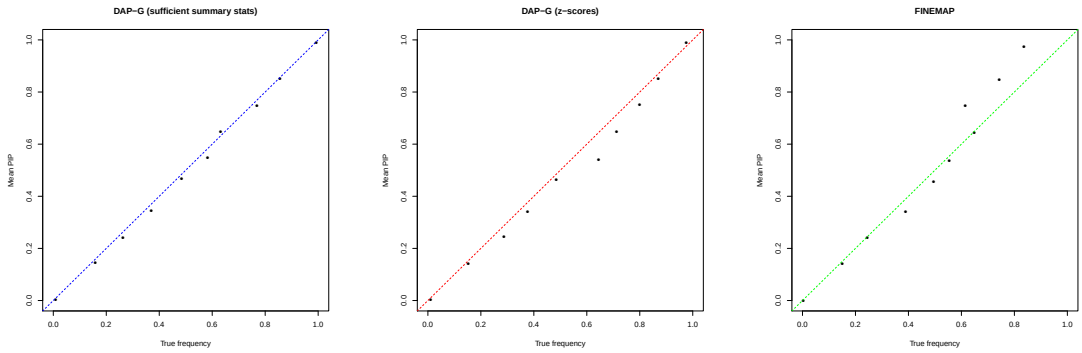

Figure A2: Calibration of SNP PIPs in the simulation study. PIPs from three Bayesian multi-SNP analysis methods (DAP-G with sufficient summary statistics, DAP-G with z-scores and FINEMAP with z-scores) are examined. PIPs from each method are classified into 10 equal-length frequency bins, the average PIP versus the corresponding true proportion (i.e., frequency) of causal SNPs for each bin is then plotted for each bin. If the PIPs are calibrated, we expect all points are aligned in the diagonal line. Points deviating from the diagonal line indicate that the PIPs may not be calibrated. More specifically, points below the diagonal line imply that the corresponding PIPs are conservative and points above the diagonal line suggest the PIPs are anti-conservative.

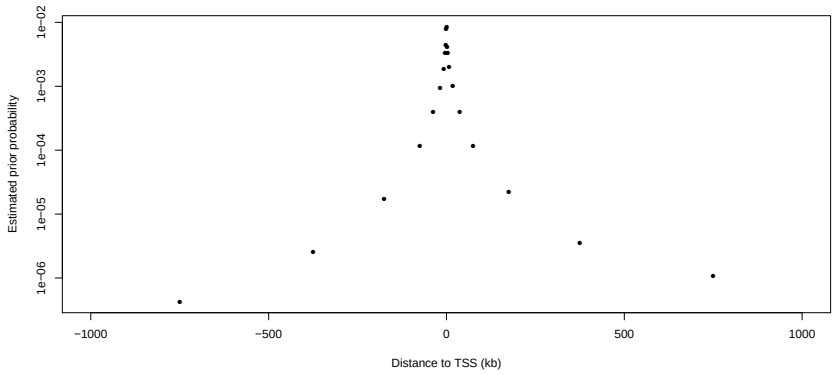

Figure A3: Relationship between estimated *cis*-eQTL priors and the SNP distances to transcription start sites (DTSS). All *cis* candidate SNPs are classified into 21 unequal-length bins according to their DTSS values. An EM algorithm implemented in the software package TORUS is used to estimate the prior inclusion probability for SNPs in each bin. Note that the quantitative distance information for the distance bins is *not* used by the EM algorithm. Each point on the plot represents the middle point of a distance bin, and its corresponding estimated prior. The result displays a clear pattern of fast decay of the abundance of eQTLs away from transcription start sites.

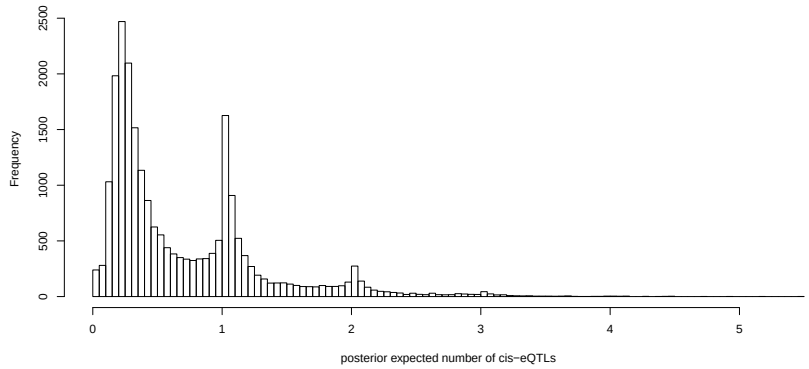

Figure A4: Histogram of posterior expected number of *cis*-eQTLs for 22,749 protein-coding and lincRNA genes analyzed in the GTEx whole blood data.

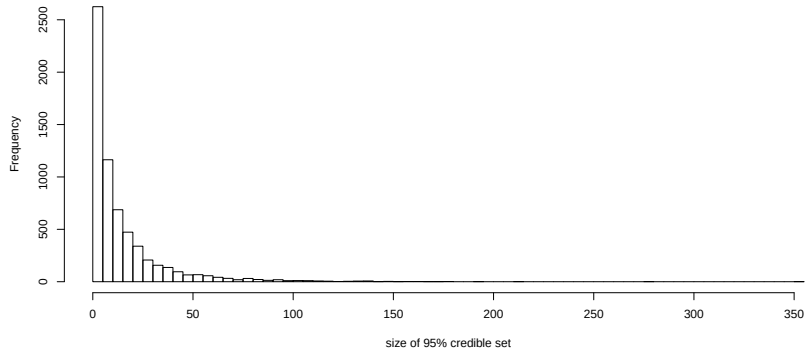

Figure A5: Histogram of the size of 95% credible sets constructed for 6,328 independent whole blood *cis*-eQTLs using GTEx samples.

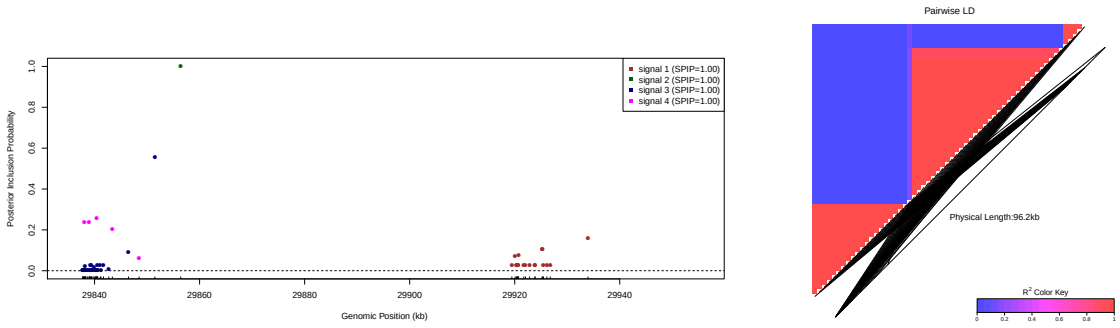

Figure A6: *cis*-eQTLs identified for gene *TMTC1*. The left panel shows 4 independent association signals are confidently identified in the *cis*-region of gene *TMTC1* (all SPIPS $\rightarrow$ 1). Each colored point represents a member SNP in the corresponding 95% credible set. The size of the credible sets differs according to different LD patterns. The right panel plots the LD pattern (R^2) between the plotted SNPs. There is high LD within each signal cluster and very weak LD between the clusters.

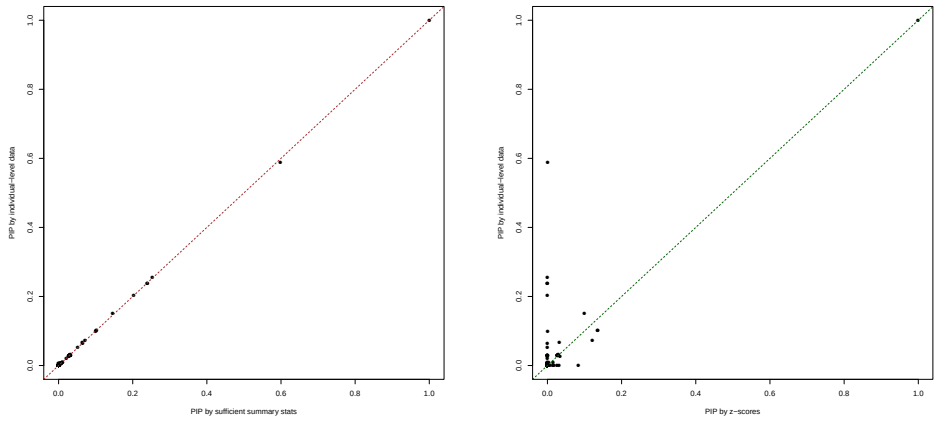

Figure A7: Comparison of PIPs computed from individual-level data versus summary statistics. The PIPs for 863 *cis* candidate SNPs for gene *TMTC1* are plotted. All PIPs are computed by DAP-G. The left panel shows the PIPs computed from sufficient summary statistics, and they are identical to the PIPs computed from individual-level data. The right panel shows the PIPs computed from z-scores, which are noticeably conservative, for most cases, in comparison to the PIPs computed from the individual-level data.

B.2.5 Tables

	DAP-G (z-scores)		DAP-G (sufficient summary stats)	
FDR control level	FDP	power	FDP	power
0.01	0.009	0.452	0.002	0.529
0.05	0.013	0.501	0.010	0.576
0.10	0.021	0.535	0.030	0.605
0.15	0.040	0.560	0.064	0.623
0.20	0.067	0.580	0.107	0.634
0.25	0.100	0.598	0.148	0.646

Table A1: Realized signal-level false discovery proportion (FDP) and power in simulation studies. In all cases, the actual FDP values are below the target FDR control levels. As expected, the powers of DAP-G using sufficient summary statistics are consistently higher than the using the z-score based summary statistics.

$\log_{10}\mathrm{BF}(\phi^2=1;\hat{\tau})$	$\log_{10}\mathrm{BF}_{\vec{z}}(\phi^2=1;\check{\tau})$	$\log_{10}\mathrm{BF}(\phi^2=1;\tilde{\tau})$
-0.776	-0.822	-0.870

Table A2: Comparison of different approximate Bayes factors under weak association

$\log_{10}\mathrm{BF}(\phi^2=1;\hat{\tau})$	$\log_{10}\mathrm{BF}_{\vec{z}}(\phi^2=1;\check{\tau})$	$\log_{10}\mathrm{BF}(\phi^2=1;\tilde{\tau})$
16.244	13.247	12.09

Table A3: Comparison of different approximate Bayes factors under modest association

B.2.6 Supplemental Figures

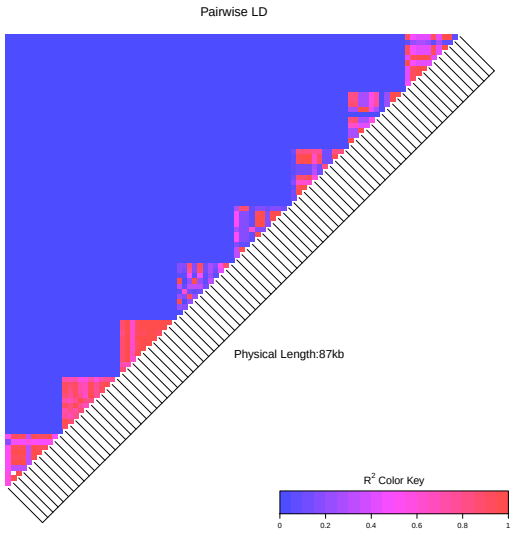

Figure A8: LD structures from 8 randomly selected blocks in the simulation study. R^2 values are plotted for 88 SNPs from 8 artificially constructed blocks. The 8 blocks are randomly selected from a total of 91 blocks used in the simulations. All genotype data are real and from GUEVADIS study. By our construction, LD patterns within each block vary but the LD between blocks is consistently weak.

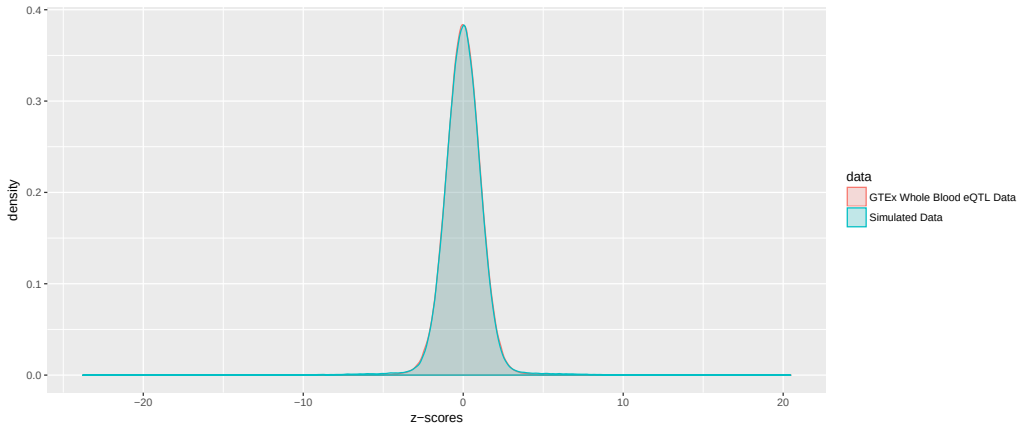

Figure A9: Comparison of single SNP z-scores between simulated data and GTEx whole blood eQTL data. The effect size parameters in the simulation studies are chosen to mimic the observed *cis*-eQTL data. The density of z-scores computed from the simulated data overlay almost entirely with the observed z-score distribution from the GTEx whole blood data.

APPENDIX C

Appendix of Chapter 4

C.1 Identifying Reproducible eGenes

We assume that each eGene has one causal SNP associated with. In this setting, the posterior probability that gene o being reproducible eGene can be measured as follows: First, the probability of a gene-SNP pair s being reproducible is calculated as follows:

$$\begin{aligned} &Pr(\beta_s \in L \text{ and } \beta_s \neq 0|\text{Data}) = \\ &Pr(\beta_s \neq 0|\text{Data}) \times Pr(\beta_s \in L_r|\text{Data}, \beta_s \neq 0), \end{aligned} \tag{C.1.1}$$

where L_r is a set of grid points that represent of reproducible scenarios.

C.1.1 Computation of the Probability of being Signal

The probability of a gene-SNP pair s being a genuine signal, $Pr(\beta_s \neq 0|\text{Data})$, can be estimated using the probabilistic hierarchical model discussed in *Wen* (2016). Specifically, for S SNPs within a testing window for a gene, genotype-phenotype association is:

$$\vec{\boldsymbol{y}} = \mu\mathbf{1} + \sum_{s=1}^{S} \beta_s \vec{\boldsymbol{g}}_s + \vec{\boldsymbol{e}},\ \vec{\boldsymbol{e}} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}), \tag{C.1.2}$$

where $\vec{\boldsymbol{g}}_s$ and β_s are genotype and an effect size for SNP s.

In this model, we set a binary variable explaining the status of β_s, γ_s as follows :

$$\gamma_s = \begin{cases} 1 & \beta_s \neq 0 \\ 0 & \beta_s = 0. \end{cases} \tag{C.1.3}$$

With the definition of γ_s, a causal SNP as a SNP with $\gamma_s = 1$. Also, $\vec{\boldsymbol{\gamma}} := (\gamma_1, ..., \gamma_S)$.

For the above model, the log prior odds on γ_s is set as follows:

$$\log\left[\frac{\Pr(\gamma_s = 1)}{\Pr(\gamma_s = 0)}\right] = \alpha_0 + \sum_{q=1}^{Q} \alpha_q d_{sq}, \tag{C.1.4}$$

where $\vec{\boldsymbol{d}}_s := (d_{s1}, \ldots, d_{sQ})$ are Q genomic annotations and $\alpha_0, \ldots, \alpha_Q$ are enrichment parameters for SNP s. Assuming there is only one SNP being QTL in a testing window and no annotation is available, the only enrichment parameter in the model is α_0 and the model space of $\{\vec{\boldsymbol{\gamma}} : ||\vec{\boldsymbol{\gamma}}|| \leq 1\}$ contains only the null model, $\vec{\boldsymbol{\gamma}} = \mathbf{0}$, and all single-SNP association models.

With these notations and the genotype $\boldsymbol{G} := (\vec{\boldsymbol{g}}_1, ..., \vec{\boldsymbol{g}}_Q)$, Consequently, it follows that

$$\Pr(\vec{\boldsymbol{\gamma}}_l = \vec{\boldsymbol{\gamma}} \mid \vec{\boldsymbol{y}}_l, \boldsymbol{G}_l, \alpha_0) \approx \frac{\Pr(\vec{\boldsymbol{\gamma}}_l = \vec{\boldsymbol{\gamma}} \mid \alpha_0)\mathrm{BF}(\vec{\boldsymbol{\gamma}})}{\sum_{||\vec{\boldsymbol{\gamma}}'||\leq 1} \Pr(\vec{\boldsymbol{\gamma}}_l = \vec{\boldsymbol{\gamma}}' \mid \alpha_0)\mathrm{BF}(\vec{\boldsymbol{\gamma}}')}, \tag{C.1.5}$$

where

$$\Pr(\vec{\boldsymbol{\gamma}} = 0 \mid \alpha_0) = \pi_0 = \left(\frac{1}{1 + \exp(\alpha_0)}\right)^Q, \tag{C.1.6}$$

and for a single-SNP association models that is not null,

$$\Pr(\vec{\boldsymbol{\gamma}}_j^{\circ} \mid \alpha_0) = \exp(\alpha_0) \prod_{q=1}^{Q} \left(\frac{1}{1 + exp(\alpha_0)}) \right)^{-1} = \pi_0 \exp(\alpha_0).$$

$Pr(\beta_s \neq 0|\text{Data})$ can be calculated as follows:

$$\begin{aligned} \Pr(\gamma_s = 1 \mid \vec{\boldsymbol{y}}, \boldsymbol{G}, \alpha_0) &= \frac{\sum_{q=1}^{Q} e^{\alpha_0} \, \mathrm{BF}_q}{1 + \sum_{q=1}^{Q} e^{\alpha_0} \, \mathrm{BF}_q} \cdot \frac{\mathrm{BF}_s}{\sum_{q=1}^{Q} \mathrm{BF}_q} \\ &= \left[1 - \Pr(\vec{\boldsymbol{\gamma}}_l = \mathbf{0} \mid \vec{\boldsymbol{y}}, \boldsymbol{G}, \alpha_0)\right] \cdot \frac{\mathrm{BF}_s}{\sum_{q=1}^{Q} \mathrm{BF}_q}, \end{aligned} \tag{C.1.7}$$

where the Bayes factor for each SNP s, BF_s is calculated as

$$\mathrm{BF}_s = \sum_{m=1}^{M} p_m BF_{s,m}, \tag{C.1.8}$$

where $BF_{s,m}$ is the estimated Bayes factor for grid point $l_m = (k_m, \omega_m)$ in SNP s and the MLE of p_m is estimated using TORUS (*Wen*, 2017).

C.1.2 Computation of the Probability of being Reproducible

We define a set of grid points that describes "reproducible" situation as L_r. In this setting, the probability of being a reproducible pair given data and a signal, $Pr(\beta_s \in L_r|\text{Data}, \beta_s \neq 0)$, can be calculated as follows for a gene-SNP pair s:

$$Pr(\beta_s \in L_r|\text{Data}, \beta_s \neq 0) = \frac{\sum_{l_m \in L_r} p_m BF_{s,m}}{\sum_{l_m^*: k_m^* \neq 0 \& \omega_m^* \neq 0} p_m^* BF_{s,m^*}}, \tag{C.1.9}$$

where

$$p_{l_m}^* = \begin{cases} (1 - \pi_0) \times p_m & l_m \neq (0,0) \\ \pi_0 & l_m = (0,0) \end{cases} \tag{C.1.10}$$

C.1.3 Probability of Reproducible eGENE

The posterior probability of being reproducible eGene can be calculated by summing over C.1.1 across S SNP within the testing windows of each gene:

$$Pr(\beta_o \neq 0|\text{Data}) = \sum_{s=1}^{S} Pr(\beta_s \in L \text{ and } \beta_s \neq 0|\text{Data}). \tag{C.1.11}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

Albert, F. W., and L. Kruglyak (2015), The role of regulatory variation in complex traits and disease, *Nature Reviews Genetics*, *16*(4).

Ardlie, K. G., et al. (2015), The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans, *Science*, *348*(6235), 648–660.

Banovich, N. E., X. Lan, G. McVicker, B. Van de Geijn, J. F. Degner, J. D. Blischak, J. Roux, J. K. Pritchard, and Y. Gilad (2014), Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels, *PLOS Genetics*, *10*(9), e1004,663.

Barber, R. F., E. J. Candès, et al. (2015), Controlling the false discovery rate via knockoffs, *The Annals of Statistics*, *43*(5), 2055–2085.

Benner, C., C. C. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen (2016), Finemap: efficient variable selection using summary data from genome-wide association studies, *Bioinformatics*, *32*(10), 1493–1501.

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

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

Bottolo, L., et al. (2013), Guess-ing polygenic associations with multiple phenotypes using a gpu-based evolutionary stochastic search algorithm, *PLoS genetics*, *9*(8), e1003,657.

Brzyski, D., C. B. Peterson, P. Sobczyk, E. J. Candès, M. Bogdan, and C. Sabatti (2017), Controlling the rate of gwas false discoveries, *Genetics*, *205*(1), 61–75.

Chen, W., B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, G. A. Poland, and D. J. Schaid (2015), Fine mapping causal variants with an approximate bayesian method using marginal test statistics, *Genetics*, *200*(3), 719–736.

Degner, J. F., et al. (2012), Dnase [thinsp] i sensitivity qtls are a major determinant of human expression variation, *Nature*, *482*(7385), 390–394.

Efron, B. (2012), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press.

Efron, B., et al. (2007), Size, power and false discovery rates, *The Annals of Statistics*, *35*(4), 1351–1377.

ENCODE Project Consortium (2012), An integrated encyclopedia of dna elements in the human genome, *Nature*, *489*(7414), 57–74.

Fan, J., and J. Lv (2008), Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Findlay, G. M., E. A. Boyle, R. J. Hause, J. C. Klein, and J. Shendure (2014), Saturation editing of genomic regions by multiplex homology-directed repair, *Nature*, *513*(7516), 120–123.

Flutre, T., X. Wen, J. Pritchard, and M. Stephens (2013), A statistical framework for joint eQTL analysis in multiple tissues, *PLOS Genetics*, *9*(5), e1003,486.

Gaffney, D. J., et al. (2012), Dissecting the regulatory architecture of gene expression qtls, *Genome Biol*, *13*(1), R7.

Giambartolomei, C., D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol (2014), Bayesian test for colocalisation between pairs of genetic association studies using summary statistics, *PLoS genetics*, *10*(5), e1004,383.

GTEx Consortium (2017), Genetic effects on gene expression across human tissues, *Nature*, *550*(7675), 204.

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

Hormozdiari, F., E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin (2014), Identifying causal variants at loci with multiple signals of association, *Genetics*, *198*(2), 497–508.

Hormozdiari, F., et al. (2016), Colocalization of gwas and eqtl signals detects target genes, *The American Journal of Human Genetics*, *99*(6), 1245–1260.

Kichaev, G., W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, and B. Pasaniuc (2014), Integrating functional data to prioritize causal variants in statistical fine-mapping studies, *PLOS Genetics*, *10*(10), e1004,722.

Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *The annals of mathematical statistics*, *22*(1), 79–86.

Kundaje, A., et al. (2015), Integrative analysis of 111 reference human epigenomes, *Nature*, *518*(7539), 317–330.

Lappalainen, T., et al. (2013), Transcriptome and genome sequencing uncovers functional variation in humans, *Nature*, *501*, 506–511.

Li, Q., J. B. Brown, H. Huang, P. J. Bickel, et al. (2011), Measuring reproducibility of high-throughput experiments, *The annals of applied statistics*, *5*(3), 1752–1779.

Liu, D. J., et al. (2014), Meta-analysis of gene-level tests for rare variant association, *Nature genetics*, *46*(2), 200.

Maller, J. B., et al. (2012), Bayesian refinement of association signals for 14 loci in 3 common diseases, *Nature genetics*, *44*(12), 1294–1301.

McVicker, G., et al. (2013), Identification of genetic variants that affect histone modifications in human cells, *Science*, *342*(6159), 747–749.

Moyerbrailean, G. A., C. T. Harvey, C. A. Kalita, X. Wen, F. Luca, and R. Pique-Regi (2016), Which genetics variants in dnase-seq footprints are more likely to alter binding?, *PLoS Genetics*, *12*(2), e1005,875.

Musunuru, K., et al. (2010), From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus, *Nature*, *466*(7307), 714–719.

Patwardhan, R. P., C. Lee, O. Litvin, D. L. Young, D. Pe'er, and J. Shendure (2009), High-resolution analysis of dna regulatory elements by synthetic saturation mutagenesis, *Nature biotechnology*, *27*(12), 1173–1175.

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., J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard (2011), Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data, *Genome research*, *21*(3), 447–455.

Savic, D., S. Park, K. Bailey, G. Bell, and M. Nobrega (2013), In vitro scan for enhancers at the tcf7l2 locus, *Diabetologia*, *56*(1), 121–125.

Scott, L. J., et al. (2016), The genetic regulatory signature of type 2 diabetes in human skeletal muscle, *Nature communications*, *7*.

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

Stephens, M. (2016), False discovery rates: a new deal, *Biostatistics*, *18*(2), 275–294.

Veyrieras, J.-B., S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard (2008), High-resolution mapping of expression-QTLs yields insight into human gene regulation, *PLOS Genetics*, *4*(10), e1000,214.

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. (2016), Molecular qtl discovery incorporating genomic annotations using bayesian false discovery rate control, *The Annals of Applied Statistics*, *10*(3), 1619–1638.

Wen, X. (2017), Robust bayesian fdr control using bayes factors, with applications to multi-tissue eqtl discovery, *Statistics in Biosciences*, *9*(1), 28–49.

Wen, X. (2018), A unified view of false discovery rate control: Reconciliation of bayesian and frequentist approaches, *arXiv preprint arXiv:1803.05284*.

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

Wen, X., M. Stephens, 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., F. Luca, and R. Pique-Regi (2015a), Cross-population joint analysis of eqtls: fine mapping and functional annotation, *PLoS genetics*, *11*(4), e1005,176.

Wen, X., F. Luca, and R. Pique-Regi (2015b), Cross-population joint analysis of eqtls: Fine mapping and functional annotation, *PLOS Genetics*, *11*(4), e1005,176.

Wen, X., Y. Lee, F. Luca, and R. Pique-Regi (2016), Efficient integrative multi-snp association analysis via deterministic approximation of posteriors, *The American Journal of Human Genetics*, *98*(6), 1114–1129.

Wen, X., R. Pique-Regi, and F. Luca (2017), Integrating molecular qtl data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization, *PLoS genetics*, *13*(3), e1006,646.

Wilson, M. A., E. S. Iversen, M. A. Clyde, S. C. Schmidler, and J. M. Schildkraut (2010), Bayesian model search and multilevel inference for snp association studies, *The annals of applied statistics*, *4*(3), 1342.

Yang, J., et al. (2012), Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits, *Nature genetics*, *44*(4), 369.

Zhou, X., P. Carbonetto, and M. Stephens (2013), Polygenic modeling with bayesian sparse linear mixed models, *PLoS Genet*, *9*(2), e1003,264.

Zhu, X., M. Stephens, et al. (2017), Bayesian large-scale multiple regression with summary statistics from genome-wide association studies, *The Annals of Applied Statistics*, *11*(3), 1561–1592.