bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CoMM-S^2: a collaborative mixed model using summary statistics in transcriptome-wide association studies

Yi Yang[1,2], Xingjie Shi[2,3], Yuling Jiao[4], Jian Huang[5], Min Chen[6], Xiang Zhou[7], Lei Sun[8], Xinyi Lin[2,9,10], Can Yang[11], and Jin Liu[2*]

[1]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[2]Centre for Quantitative Medicine, Program in Health Services and Systems Research, Duke-NUS Medical School, Singapore 169857
[3]Department of Statistics, Nanjing University of Finance and Economics, Nanjing, China
[4]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China
[5]Department of Applied Mathematics, Hong Kong Polytechnics University, Hong Kong, China
[6]Academy of Mathematics and Systems Science, The Chinese Academy of Sciences ,Beijing, China
[7]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
[8]Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, Singapore
[9]Singapore Clinical Research Institute, Singapore
[10]Singapore Institute for Clinical Sciences, A*STAR, Singapore
[11]Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China

Abstract

Motivation: Although genome-wide association studies (GWAS) have deepened our understanding of the genetic architecture of complex traits, the mechanistic links that underlie how genetic variants cause complex traits remains elusive. To advance our understanding of the underlying mechanistic links, various consortia have collected a

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

vast volume of genomic data that enable us to investigate the role that genetic variants play in gene expression regulation. Recently, a collaborative mixed model (CoMM) [42] was proposed to jointly interrogate genome on complex traits by integrating both the GWAS dataset and the expression quantitative trait loci (eQTL) dataset. Although CoMM is a powerful approach that leverages regulatory information while accounting for the uncertainty in using an eQTL dataset, it requires individual-level GWAS data and cannot fully make use of widely available GWAS summary statistics. Therefore, statistically efficient methods that leverages transcriptome information using only summary statistics information from GWAS data are required.

Results: In this study, we propose a novel probabilistic model, CoMM-S^2, to examine the mechanistic role that genetic variants play, by using only GWAS summary statistics instead of individual-level GWAS data. Similar to CoMM which uses individual-level GWAS data, CoMM-S^2 combines two models: the first model examines the relationship between gene expression and genotype, while the second model examines the relationship between the phenotype and the predicted gene expression from the first model. Distinct from CoMM, CoMM-S^2 requires only GWAS summary statistics. Using both simulation studies and real data analysis, we demonstrate that even though CoMM-S^2 utilizes GWAS summary statistics, it has comparable performance as CoMM, which uses individual-level GWAS data.

Contact: jin.liu@duke-nus.edu.sg

Availability and implementation: The implement of CoMM-S^2 is included in the *CoMM* package that can be downloaded from https://github.com/gordonliu810822/CoMM.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the last decade, genome-wide association studies (GWAS) have achieved remarkable success in identifying genetic susceptibility variants for a variety of complex traits/diseases [39]. However, the biology of how genetic variants affect complex traits remains unclear. Recent expression quantitative trait loci (eQTL) studies indicate that regulatory information play an important role in mediating the complex traits/diseases [26]. Measured comprehensive cellular traits can serve as reference data and provide investigators with an avenue to examine the role that genetic variants play in gene expression regulation. For example, the Genotype-Tissue Expression (GTEx) Project [23] has provided DNA sequencing data (about

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

12.5 million variants) from 449 individuals and collected gene-expression measurements of 44 tissues from these individuals; the number of subjects increases to 620 in over 48 tissues in the recent V7 release. Although the sample sizes of these reference datasets are limited, they provide an important avenue for one to study how genetic variants regulate human gene expression in different tissues.

In the absence of identical cohorts in eQTL and GWAS datasets, various authors have proposed statistical methods that allow one to leverage regulatory information on the cellular mechanisms in a GWAS analysis. These methods can be broadly grouped into two categories. The first group consists of methods that require the use of individual-level GWAS data, and include methods such as PrediXcan [10] and CoMM [42]. Because methods in this category require the availability of all individual-level genotype and phenotype data, their application can be complicated by restrictions on data sharing and storage. In contrast to the first group of methods that utilize individual-level data, the second group of methods uses GWAS summary statistics; for example one could apply the second group of methods to GWAS results that are publicly available from GWAS repositories, such as the NHGRI-EBI GWAS Catalog [4]. Examples of these methods include TWAS [13], S-PrediXcan [1], and UTMOST [17]. Among these methods, TWAS and S-PrediXcan use transcriptome data from a single tissue while UTMOST can be applied to cross-tissue analysis. Generally, TWAS-type methods (PrediXcan, S-PrediXcan, TWAS and UTMOST) proceed with three steps. First, the expression reference panel is used to fit predictive models for each gene using genetic variants in the vicinity of a gene. Next, levels of gene expression for the individuals in the GWAS data are predicted using these models. Finally, associations between the predicted expression levels and the complex trait are examined by simple linear regressions. Consequently, TWAS-type methods do no account for the uncertainty associated with the first step. In contrast, CoMM accounts for the uncertainty by combining the three steps in

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a unified probabilistic framework.

Compared with methods using individual-level GWAS data, methods using GWAS summary statistics face an additional difficulty: the summary statistics do not contain any information of linkage disequilibrium (LD), which plays an important role in prioritizing variants in GWAS. TWAS used an imputation method to impute the expression-trait association statistics directly from GWAS summary statistics [13] while S-PrediXcan derived a test statistic using pre-calculated weights to expression and a reference panel to estimate correlation (LD) among cis-variants. To make use of summary statistics rigorously, it is important to develop a probabilistic model. [16] first proposed an approximated distribution for z-scores in CAVIAR. Later, [45] formalized this distribution by introducing a regression with summary statistics (RSS) likelihood in a Bayesian framework and they further showed that the difference between the RSS log-likelihood and the one from individual-level data was constant. Although the approximated RSS-type distribution has been extended in several works including RSS-E [46] and REMI [18], both of these works including RSS are designed for one-sample studies. In the analysis using two different samples, such as PrediXcan, TWAS and CoMM, the questions become how to combine a RSS distribution for GWAS summary statistics with that for eQTL data.

To overcome the limitation of not accounting for uncertainty and further extend CoMM using GWAS summary statistics, we propose a probabilistic model, a Collaborative Mixed Models for GWAS summary statistics – CoMM-S^2. Unlike TWAS and S-PrediXcan, our method accounts for the uncertainty in the 'imputed' gene expression. The key idea is to build a joint probabilistic model for GWAS summary statistics and individual-level eQTL data, and use the 1KG data as a reference panel to estimate LD. We has also developed an efficient variational Bayesian expectation-maximization accelerated using parameter expansion (PX-VBEM), where the calibrated evidence lower bound is used to conduct likelihood

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ratio tests for genome-wide gene associations with complex traits/diseases. We illustrate the performance of CoMM-S^2 with extensive simulation studies and real data applications of 10 traits in NFBC1966 dataset and summary statistics from 14 traits/diseases. The results demonstrate that CoMM-S^2 performs better than competing methods.

2 Methods

2.1 Notation

Suppose that we have an individual-level eQTL dataset $\mathcal{D}_1 = \{\mathbf{Y}, \mathbf{W}_1\}$ that consists of n_1 samples, g genes and m genetic variants, and where $\mathbf{Y} \in \mathbb{R}^{n_1 \times g}$ is a matrix of gene expression and $\mathbf{W}_1 \in \mathbb{R}^{n_1 \times m}$ is a genotype matrix. In addition, we have the GWAS summary statistics $\mathcal{D}_2 = \{\widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{s}}\}$, where $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^m$ and $\widehat{\mathbf{s}} \in \mathbb{R}^m$ are the effect sizes and standard errors from the single-variant analysis for all genetic variants. We further assume that the individual-level GWAS data corresponding to $\mathcal{D}_2$ has the phenotype vector $\mathbf{z} \in \mathbb{R}^{n_2 \times 1}$ and the centered genotype matrix $\mathbf{W}_2 \in \mathbb{R}^{n_2 \times m}$. The sample sizes of $\mathcal{D}_1$ and $\mathcal{D}_2$ are generally distinct, with $\mathcal{D}_1$ having a smaller sample size ($\approx 10^2$) compared to the sample size of $\mathcal{D}_2$ ($\approx 10^4 \sim 5 \times 10^5$). We will examine gene expression levels for each gene individually. Let $\mathbf{y}_j$, the j-th column of $\mathbf{Y}$, be the gene expression level of the j-th gene and let $\mathbf{W}_{1j}$ be a genotype matrix containing its nearby genetic variants (within either 50 kb upstream of the transcription start site or 50 kb downstream of the transcription end site, in this study), respectively. We standardize the genotype data $\mathbf{W}_{1j} = [\mathbf{w}_{1j1}, \ldots, \mathbf{w}_{1jm_j}] \in \mathbb{R}^{n_1 \times m_j}$ to mean zero and unit variance. Correspondingly, summary statistics for genetic variants using the centered genotype ($\mathbf{W}_{2j}$) within the j-th gene is $\{\widehat{\boldsymbol{\gamma}}_j, \widehat{\mathbf{s}}_j\}$, where $\widehat{\boldsymbol{\gamma}}_j \in \mathbb{R}^{m_j}$, $\widehat{\mathbf{s}}_j \in \mathbb{R}^{m_j}$, and m_j is the number of variants corresponding to the j-th gene. Denote $\widehat{\mathbf{S}}_j = \mathrm{diag}(\widehat{\mathbf{s}}_j)$ a diagonal matrix for the j-th gene and $\widehat{\mathbf{R}}_j \in \mathbb{R}^{m_j \times m_j}$ the estimated correlation among genetic variants within the j-th gene.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.2 Model

We first model the relationship in the eQTL data using linear regression

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}_{1j}\boldsymbol{\gamma}_j + \mathbf{e}, \tag{1}$$

where $\boldsymbol{\gamma}_j = [\gamma_{j1}, \ldots, \gamma_{jm_j}]^T$ is an $m_j \times 1$ vector of genetic effects on the gene expression, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$ is an $n_1 \times 1$ vector of independent random noises for the gene expression levels, $\mathbf{X} \in \mathbb{R}^{n_1 \times q}$ are the design matrix for covariates including intercept, and $\boldsymbol{\beta}$ is a $q \times 1$ vector of the corresponding effect sizes for covariates. Similar to CoMM [42], the relationship between the phenotype $\mathbf{z}$ and genotype $\mathbf{W}_{2j}$ nearby the j-th gene can be modeled as

$$\mathbf{z} = \alpha_j \mathbf{W}_{2j}\boldsymbol{\gamma}_j + \mathbf{e}_2, \tag{2}$$

where $\mathbf{e}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_{e2}^2\mathbf{I})$ is an $n_2 \times 1$ vector of independent error associated with the phenotype. Here, α_j represents the effects of gene expression of gene j on the phenotype, due to genotype. Assume that individual-level data $\{\mathbf{z}, \mathbf{W}_2\}$ is inaccessible, but the summary statistics $\{\widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{s}}\}$ from the univariate linear regression are available. It can be shown that the distribution of $\widehat{\boldsymbol{\gamma}}_j$ can be approximated by [45]

$$\widehat{\boldsymbol{\gamma}}_j | \boldsymbol{\gamma}_j, \widehat{\mathbf{R}}_j, \widehat{\mathbf{S}}_j \sim \mathcal{N}(\alpha_j \widehat{\mathbf{S}}_j \widehat{\mathbf{R}}_j \widehat{\mathbf{S}}_j^{-1} \boldsymbol{\gamma}_j, \widehat{\mathbf{S}}_j \widehat{\mathbf{R}}_j \widehat{\mathbf{S}}_j), \tag{3}$$

provided that sample size n_2 to generate these summary statistics is large and the trait is highly polygenic (*i.e.*, the squared correlation coefficient between the trait and each genetic variant is close to zero). We further assume the prior distribution for $\boldsymbol{\gamma}_j$ is a Gaussian,

$$\boldsymbol{\gamma}_j \sim \mathcal{N}(0, \sigma_{\gamma_j}^2 \mathbf{I}_{m_j}), \tag{4}$$

which is widely used in genetics [44]. Taking $\boldsymbol{\gamma}_j$ as the latent variable, the complete data likelihood can be written as

$$\begin{aligned} &\Pr(\mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j, \boldsymbol{\gamma}_j | \mathbf{X}, \mathbf{W}_{1j}, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \boldsymbol{\theta}) \\ =&\Pr(\mathbf{y}_j | \boldsymbol{\gamma}_j, \mathbf{X}, \mathbf{W}_{1j}; \boldsymbol{\theta}) \Pr(\widehat{\boldsymbol{\gamma}}_j | \boldsymbol{\gamma}_j, \widehat{\mathbf{R}}_j, \widehat{\mathbf{S}}_j) \Pr(\boldsymbol{\gamma}_j) \end{aligned} \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $\boldsymbol{\theta} = \{\sigma^2_{\gamma_j}, \sigma^2_e, \alpha_j, \boldsymbol{\beta}\}$ is the collection of model parameters. By integrating out latent variable $\boldsymbol{\gamma}_j$, the marginal likelihood is

$$\begin{aligned} &\Pr(\mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j, |\mathbf{X}, \mathbf{W}_{1j}, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \boldsymbol{\theta}) \\ &= \int_{\boldsymbol{\gamma}_j} \Pr(\mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j, \boldsymbol{\gamma}_j | \mathbf{X}, \mathbf{W}_{1j}, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \boldsymbol{\theta}) d\boldsymbol{\gamma}_j \end{aligned} \tag{6}$$

2.3 Algorithm

We require a computationally efficient algorithm that is capable of fitting model (5) when the signal-noise-ratio is low. A standard expectation-maximization (EM) algorithm in not ideal for this purpose due to the slow convergence; a Newton-Raphson algorithm is also not ideal because it can be unstable because of the non-negative constraint on variance components. Additionally, a standard EM algorithm involves the inversion of $\widehat{\mathbf{R}}_j$, which may cause numerical failure as $\widehat{\mathbf{R}}_j$ is estimated from a small reference panel. Therefore, we develop a variational Bayesian (VB) EM algorithm [3] accelerated by parameter expansion [22], namely, PX-VBEM. First, the original model (1) can be expanded as follows

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta} + \tau \mathbf{W}_{1j}\boldsymbol{\gamma}_j + \mathbf{e}, \tag{7}$$

where $\tau \in \mathbb{R}$ is the expanded parameter, the likelihood for summary statistics (3) and the prior remain the same, and model parameters become $\boldsymbol{\theta} = \{\sigma^2_{\gamma_j}, \sigma^2_e, \alpha_j, \boldsymbol{\beta}, \tau\}$. Next, we sketch the variational Bayesian EM algorithm for the expanded model (7). Given a variational posterior distribution $q(\boldsymbol{\gamma}_j)$, it is easy to verify that the marginal likelihood can be decomposed into two components, the evidence lower bound (ELBO) and the KL divergence,

$$\Pr(\mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j | \mathbf{X}, \mathbf{W}_{1j}, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \boldsymbol{\theta}) = \mathcal{L}(q) + \mathbb{KL}(q||p) \tag{8}$$

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where

$$\begin{aligned}\mathcal{L}(q) &= \int_{\boldsymbol{\gamma}_j} q(\boldsymbol{\gamma}_j) \log \frac{\Pr(\mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j, \boldsymbol{\gamma}_j | \mathbf{X}, \mathbf{W}_{1j}, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \boldsymbol{\theta})}{q(\boldsymbol{\gamma}_j)} d\boldsymbol{\gamma}_j \\ \mathbb{KL}(q||p) &= \int_{\boldsymbol{\gamma}_j} q(\boldsymbol{\gamma}_j) \log \frac{q(\boldsymbol{\gamma}_j)}{p(\boldsymbol{\gamma}_j | \mathbf{X}, \mathbf{W}_{1j}, \mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \boldsymbol{\theta})} d\boldsymbol{\gamma}_j.\end{aligned} \tag{9}$$

Note that $\mathcal{L}(q)$ is the ELBO of the marginal likelihood, and $\mathbb{KL}(q||p)$ is Kullback-Leibler (KL) divergence between two distributions and satisfies $\mathbb{KL}(q||p) \geq 0$, with the equality holding if, and only if, the variational posterior probability (q) and the true posterior probability (p) are equal. Similar to the EM algorithm, we can maximize the ELBO $\mathcal{L}(q)$ by optimizing with respect to q that is equivalent to minimizing the KL divergence [2]. To make the evaluation of the lower bound computationally efficient, we use the mean-field theory [27] and assume that $q(\boldsymbol{\gamma}_j)$ can be factorized as

$$q(\boldsymbol{\gamma}_j) = \prod_{k=1}^{m_j} q(\gamma_{jk}) \tag{10}$$

This is the only assumption that we make using variational inference. This factorization (10) is used as an approximation for the posterior distribution $p(\boldsymbol{\gamma}_j | \mathbf{X}, \mathbf{W}_{1j}, \mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \boldsymbol{\theta})$. In the VB E-step, given the hidden variables γ_{ji}, $i \neq k$, the terms with γ_{jk} have a quadratic form, where i and k are indices for the i-th and the k-th genetic variants, respectively. Thus, the variational posterior distribution of γ_{jk} is a Gaussian distribution $\mathcal{N}(u_{jk}, v_{jk}^2)$. The details of derivation for the updating formula of mean $\mathbf{u}_j = [u_{j1}, \ldots, u_{jm_j}]^T$ and standard deviation $\mathbf{v}_j = [v_{j1}, \ldots, v_{jm_j}]^T$, and the ELBO $\mathcal{L}(q)$ of the marginal likelihood (9) at the old parameters $\boldsymbol{\theta}^{old}$ can be found in the supplementary document. In the VB M-step, we obtain the new updates for all parameters $\boldsymbol{\theta}$ by setting the derivative of ELBO to zero. The

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

resulting updating equations for parameters are

$$\begin{aligned}
\sigma_e^2 &= \frac{||\mathbf{y}_j^* - \tau\mathbf{W}_{1j}\mathbf{u}_j||^2 + \tau^2\sum_{k=1}^{m_j}\mathbf{w}_{1jk}^T\mathbf{w}_{1jk}v_k^2}{n_1}, \\
\sigma_{\gamma_j}^2 &= \frac{||\mathbf{u}_j||^2 + ||\mathbf{v}_j||^2}{m_j}, \\
\alpha_j &= \frac{\mathbf{u}_j^T\widehat{\mathbf{S}}_j^{-2}\widehat{\boldsymbol{\gamma}}_j}{\mathbf{u}_j^T\widehat{\mathbf{S}}_j^{-1}\widehat{\mathbf{R}}\widehat{\mathbf{S}}_j^{-1}\mathbf{u}_j + (\mathbf{v}_j\odot\mathbf{v}_j)^T\mathrm{diag}(\widehat{\mathbf{S}}_j^{-1}\widehat{\mathbf{R}}_j\widehat{\mathbf{S}}_j^{-1})}, \\
\boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}_j - \tau\mathbf{W}_{1j}\mathbf{u}_j), \\
\tau &= \frac{\mathbf{u}_j^T\mathbf{W}_{1j}^T\mathbf{y}_j^*}{\mathbf{u}_j^T\mathbf{W}_{1j}^T\mathbf{W}_{1j}\mathbf{u}_j + \sum_{k=1}^{m_j}\mathbf{w}_{1jk}^T\mathbf{w}_{1jk}v_k^2},
\end{aligned} \tag{11}$$

where $\mathbf{y}_j^* = \mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}$ is the vector of residuals after removing the fixed effects, $\odot$ denotes the element-wise multiplication of two vectors, and $\mathrm{diag}(\mathbf{A})$ denotes a vector whose elements are the diagonal entries of the square matrix $\mathbf{A}$. The corresponding PX-VBEM algorithm is summarized as Algorithm 1 in the supplementary document.

2.4 Reference panel

The CoMM-S^2 uses marginal effect sizes and their standard errors to construct probabilistic modeling for summary statistics from GWAS. Using summary-level data, we do not have any information for correlations among SNPs (*i.e.*, LD, denoted as $\mathbf{R}_j$). Here, we choose to use 1KG samples as a reference panel. We first calculate the empirical correlation matrix $\widehat{\mathbf{R}}_j^{\mathrm{emp}} = [r_{ik}] \in \mathbb{R}^{m_j\times m_j}$ with $r_{ik} = \frac{\mathbf{w}_{ji}^T\mathbf{w}_{jk}}{\sqrt{(\mathbf{w}_{ji}^T\mathbf{w}_{ji})(\mathbf{w}_{jk}^T\mathbf{w}_{jk})}}$, where $\mathbf{w}_{jk}$ is the genotype vector for the k-th genetic variant within the j-th gene. To make the estimated correlation matrix positive definite, we applied a simple shrinkage estimator [32] to obtain $\widehat{\mathbf{R}}_j$ as $\widehat{\mathbf{R}}_j = \lambda\widehat{\mathbf{R}}_j^{\mathrm{emp}} + (1-\lambda)\mathbf{I}_{m_j}$, where $\lambda \in [0,1]$ is the shrinkage intensity. Note that the shrinkage correlation matrix is the combination of the two extremes, the empirical correlation matrix $\widehat{\mathbf{R}}_j^{\mathrm{emp}}$ and the identity matrix $\mathbf{I}$. It is easy to recognize that the shrinkage correlation matrix can recover the original empirical correlation matrix $\widehat{\mathbf{R}}_j^{\mathrm{emp}}$ when $\lambda = 1$ or identity matrix $\mathbf{I}$ when $\lambda = 0$. In

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

addition, we have tested with different $\lambda \in [0.8, 0.95]$ for CoMM-S^2 and its results are quite robust.

3 Statistical Inference

3.1 Evaluate association between a a complex trait/disease and a gene

We propose the following statistical test to formally examine the association between a a complex trait/disease and a gene:

$$\mathcal{H}_0 : \alpha_j = 0 \qquad \mathcal{H}_a : \alpha_j \neq 0 \tag{12}$$

A likelihood ratio test (LRT) statistic for the j-th gene is given by

$$\begin{aligned} \Lambda_j =& 2(\log \Pr(\mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j^{\mathrm{ML}} | \mathbf{X}, \mathbf{W}_{1j}, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \hat{\boldsymbol{\theta}}^{\mathrm{ML}}) \\ & - \log \Pr(\mathbf{y}_j, \widehat{\boldsymbol{\gamma}}_j^{\mathrm{ML}} | \mathbf{X}, \mathbf{W}_{1j}, \widehat{\mathbf{S}}_j, \widehat{\mathbf{R}}_j; \hat{\boldsymbol{\theta}}_0^{\mathrm{ML}})), \end{aligned} \tag{13}$$

where $\hat{\boldsymbol{\theta}}_0^{\mathrm{ML}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{ML}}$ are vectors of parameter estimates that are obtained by maximizing the marginal likelihood, under the null hypothesis $\mathcal{H}_0$ and under the alternative hypothesis $\mathcal{H}_A$, respectively. Using standard asymptotic theory [38], the test statistics Λ_j asymptotically follows the $\chi^2_{\mathrm{df}=1}$ under the null.

As discussed in Section 2.3, to overcome the intractability of maximizing the marginal likelihood, we utilize a (PX)-VBEM algorithm where we maximize the ELBO, instead of the marginal likelihood, to obtain parameter estimates $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$. Earlier applications demonstrate that (PX)-VBEM produces practically useful and accurate posterior mean estimates [3, 8, 43, 34] (i.e. $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$). While it might seem reasonable to use the estimated posterior distribution from maximizing the ELBO to directly approximate the marginal likelihood in Equation 13, it is well-known that the (PX)-VBEM typically identifies posterior distributions that underestimate the marginal variances [40, 36]. Consequently, this is not

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a feasible approach. Instead of using the estimated posterior distribution as a proxy for the marginal likelihood in Equation 13, under the assumption that the posterior means of interest are well-estimated by (PX)-VBEM for all the parameters of interest (i.e. $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$ estimated using (PX)-VBEM are well-estimated), we plug-in our estimates from the (PX)-VBEM algorithm into the marginal likelihood in Equation 13 to construct the test statistic. We term the resulting likelihood with plug-in estimates from the (PX)-VBEM algorithm a calibrated ELBO and provide more details in the next section. Briefly, the calibrated ELBO is used as a proxy to the marginal likelihood in the test statistics. Our numerical studies show that a test constructed using the calibrated ELBO works well.

3.2 Calibrated ELBO

We postulate that the ELBO can be calibrated using the form from the (PX)-EM algorithm by plugging the posterior mean estimates and parameter estimates from (PX)-VBEM, which can be used as a proxy to marginal log-likelihood. Here we describe the procedures to calibrate the ELBO in detail. The marginal log-likelihood from the PX-EM algorithm for model (8) can be written as follows

$$\begin{aligned}
&\widetilde{\mathcal{L}}(\boldsymbol{\theta}, \mathbf{u}_j) \\
=&-\frac{n_1}{2}\log(2\pi\sigma_e^2) - \frac{||\mathbf{y}^* - \tau\mathbf{W}_{1j}\mathbf{u}_j||^2 + \tau^2\mathrm{tr}(\mathbf{W}_{1j}\boldsymbol{\Sigma}_j\mathbf{W}_{1j}^T)}{2\sigma_e^2} \\
&+\alpha_j\mathbf{u}_j^T\widehat{\mathbf{S}}_j^{-2}\widehat{\boldsymbol{\gamma}}_j - \frac{1}{2}\alpha_j^2\mathbf{u}_j^T\widehat{\mathbf{S}}_j^{-1}\widehat{\mathbf{R}}_j\widehat{\mathbf{S}}_j^{-1}\mathbf{u}_j - \frac{\alpha_j^2}{2}\mathrm{tr}(\boldsymbol{\Sigma}_j\widehat{\mathbf{S}}_j^{-1}\widehat{\mathbf{R}}_j\widehat{\mathbf{S}}_j^{-1}) \\
&-\frac{m_j}{2}\log(2\pi\sigma_{\gamma_j}^2) - \frac{||\mathbf{u}_j||^2 + \mathrm{tr}(\boldsymbol{\Sigma}_j)}{2\sigma_{\gamma_j}^2} + \frac{1}{2}\log|2\pi e\boldsymbol{\Sigma}_j|
\end{aligned} \tag{14}$$

where $\mathbf{u}_j$ and $\boldsymbol{\Sigma}_j$ are the joint posterior mean and posterior variance for latent variable $\boldsymbol{\gamma}_j$, and $\boldsymbol{\Sigma}_j$ is expressed as

$$\boldsymbol{\Sigma}_j = \left(\tau^2\frac{\mathbf{W}_{1j}^T\mathbf{W}_{1j}}{\sigma_e^2} + \alpha_j^2\widehat{\mathbf{S}}_j^{-1}\widehat{\mathbf{R}}_j\widehat{\mathbf{S}}_j^{-1} + \frac{1}{\sigma_{\gamma_j}^2}\mathbf{I}_{m_j}\right)^{-1} \tag{15}$$

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Note that we explicitly express the marginal log-likelihood (14) from the PX-EM algorithm that depends on the posterior mean, $\mathbf{u}_j$, and parameter estimates, $\boldsymbol{\theta}$, as the posterior variance $\boldsymbol{\Sigma}_j$ is fully characterized by model parameters $\boldsymbol{\theta}$. Thus, we first fit the data using the PX-VBEM (Algorithm 1 in the supplementary document). Then, we re-evaluate the marginal log-likelihood from the (PX)-EM algorithm by plugging the posterior mean estimates and parameter estimates $\boldsymbol{\theta} = \{\sigma^2_{\gamma_j}, \sigma^2_e, \alpha_g, \boldsymbol{\beta}, \tau\}$ from the PX-VBEM algorithm as equation (14). Given the fact that the posterior means from (PX)-VBEM are accurate enough, the calibrated ELBO is close to the marginal log-likelihood. In Section 4, we conducted simulation studies to show that the marginal log-likelihood evaluated under the proposed calibrated procedure approximates well to that from (PX)-EM algorithms.

4 Simulations

4.1 Simulation settings

We conducted simulation studies to demonstrate that (a) CoMM-S^2 has comparable performance as CoMM (Section 4.2.1) and that (b) CoMM-S^2 generally performs as well or better than competing methods that also utilize summary statistics (Section 4.2.2). For (b), we compared the performance of CoMM-S^2 with S-PrediXcan, with both ridge regression and Enet [47], denoted as S-PrediXcan:Ridge and S-PrediXcan:Enet, respectively.

We considered the following simulation settings to evaluate the performance of CoMM-S^2. We assumed sample sizes of $n_1 = 400$, $n_2 = 5,000$, and $n_3 = 400$, which are the sample sizes for the transcriptome dataset, GWAS dataset and the reference panel dataset, respectively. To generate genotype data, we first generated a data matrix using a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\rho))$, where $\boldsymbol{\Sigma}(\rho)$ is an auto-regressive correlation structure with $\rho = 0.2$, 0.5 and 0.8, representing weak, moderate and strong LD, respectively. We then generated minor allele frequencies from a uniform distribution $\mathcal{U}(0.05, 0.5)$ and categorized the data

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

matrix into trinary variables taking values 0, 1, 2 using these minor allele frequencies and assuming Hardy-Weinberg equilibrium. All three genotype matrices, $\mathbf{W}_{1j}$, $\mathbf{W}_{2j}$ and $\mathbf{W}_j^{\text{ref}}$ (where $\mathbf{W}_{1j}$, $\mathbf{W}_{2j}$ are as defined in Section 2.1, and $\mathbf{W}_j^{\text{ref}}$ is a genotype data from a reference panel as described in Section 2.4) are generated in this manner.

To generate eQTL data from genotype data, we considered different cellular-level heritability levels (h_C^2) and sparsity levels, which are parameters that describe the genetic architecture of gene expression [41]. The cellular-level heritability (h_C^2) represents the proportion of variance of the eQTL that can be explained by genotype, while sparsity represents the proportion of genetic variants that are associated with the gene expression. For a given cellular-level heritability h_C^2, a larger number of genetic variants that are associated with gene expression levels implies a smaller genetic influence on gene expression, per genetic variant. We generated eQTL data assuming $\mathbf{y}_j = \mathbf{W}_{1j}\boldsymbol{\gamma}_j + \mathbf{e}_1$, where $\mathbf{e}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_{e_1}^2 \mathbf{I}_{n_1})$, and the non-zero $\boldsymbol{\gamma}_j$ were generated assuming $\gamma_{jk} \sim \pi\mathcal{N}(\mathbf{0}, \sigma_{\gamma_j}^2) + (1-\pi)\delta_0$, π is the eQTL sparsity level, δ_0 denotes a Dirac delta mass function at 0, and k is the index for genetic variants within gene j. $\sigma_{e_1}^2$ and $\sigma_{\gamma_j}^2$ were chosen to correspond to cellular-level heritability levels h_C^2 of 0.01, 0.05, or 0.09, which are close to the median gene expression heritability estimates computed across all genes [29]. We considered different eQTL sparsity levels of 0.1, 0.2 ,0.3, 0.4 and 0.5, where a sparsity level of 0.2 indicates that only 20% of the SNPs have non zero effects (i.e. 20% of the γ_j's are non-zero).

We generated a complex trait assuming $\mathbf{z} = \alpha_j \mathbf{W}_{2j}\boldsymbol{\gamma}_j + \mathbf{e}_2$. We assumed 100 local genetic variants (cis-SNPs). $\mathbf{e}_2$ was chosen such that the organismal-level heritability level, defined as $h_T^2 = \frac{\alpha_j^2 \text{Var}(\mathbf{W}_{2j}\boldsymbol{\gamma}_j)}{\text{Var}(\mathbf{z})}$ is controlled at $h_T^2 \in \{0, 0.001, 0.002, 0.003\}$. A organismal-level heritability $h_T^2 = 0$ corresponds to the null hypothesis that the gene has no association with the organismal-level trait.

Summary statistics were generated by applying a single-variant analysis to the GWAS

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dataset.

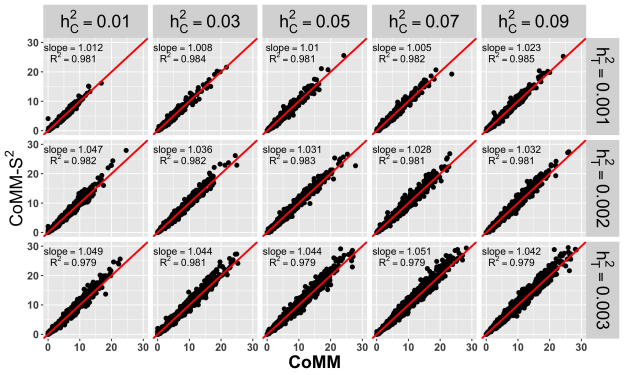


Figure 1: The scatter plot of test statistics from LRT for CoMM-S^2 vs CoMM with the setting $n_1 = 400$, $n_2 = 5{,}000$, $n_3 = 400$, $m_j = 100$, $\rho = 0.8$, $\pi = 1$. The number of replication is 2000. The reference panel is subsampled from GWAS dataset.

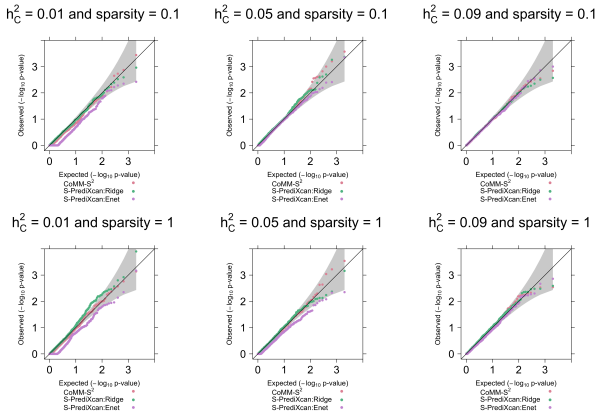


Figure 2: The qq-plot of p-values for each method (CoMM, PrediXcan:Ridge, PrediXcan:Enet, CoMM-S^2, S-PrediXcan:Ridge and S-PrediXcan:Enet) with the setting $n_1 = 400$, $n_2 = 5{,}000$, $n_3 = 400$, $\rho = 0.8$. The number of replication is 1000. The sparsity varies $\pi \in \{0.1, 1\}$ and the cellular-level heritability varies from 0.01 , 0.05, to 0.09.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

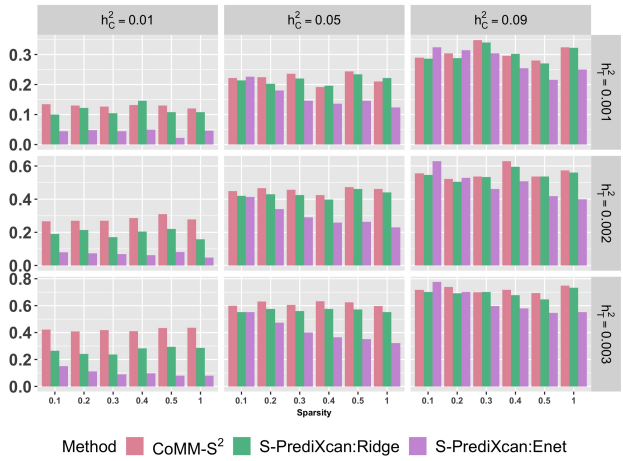


Figure 3: The comparison of power for CoMM, CoMM-S^2, PrediXcan:Ridge, PrediXcan:Enet, S-PrediXcan:Ridge and S-PrediXcan:Enet with the setting $n_1 = 400$, $n_2 = 5{,}000$, $n_3 = 400$, $\rho = 0.8$. The number of replication is 500. For each subplot, the x-axis stands for the sparsity of SNP and the y-axis stands for the proportion of significant genes within 500 replications.

4.2 Simulation results

4.2.1 CoMM-S^2 and CoMM have comparable performance

We first compared the LRT test statistics from both CoMM-S^2 and ComMM. We used 1,000 simulation replicates to compare the LRT test statistics of the two methods As shown in Figures 1 and S2 - S7, the LRT test statistics of CoMM and CoMM-S^2 are close to each other with a R^2 around 0.98.

Next, we compared the calibrated ELBO as described in Section 3.2 with the marginal log-likelihood evaluated the using EM algorithm. Here, we consider the reference panel to be the GWAS data itself, which enables the evaluation of the marginal log-likelihood from EM algorithm. As shown in Figure S1 demonstrates that the calibrated ELBO is very similar to the marginal log-likelihood.

In the real data analysis of the NFBC1966 dataset (Section 5.2), we observed that a small proportion of test statistics from CoMM are degenerate zero. To better understand

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

this phenomenon, we conducted additional simulations described in Supplementary Section 4.3. As shown in Figures S8 (a) and (b), while test statistics from CoMM degenerate to zero when $h^2_C = 0$, CoMM-S^2 performs adequately under this setting.

4.2.2 CoMM-S^2 generally has better or comparable performance compared with alternative methods that use summary statistics

We evaluated the performance of CoMM-S^2, S-PrediXcan:Ridge and S-PrediXcan:Enet under the null hypothesis $h^2_T = 0$, using 1,000 simulation replicates. The corresponding qq-plots are shown in Figures 2 and S9 - S11. The results show that CoMM-S^2 can effectively control the type-I error, while S-PrediXcan shows a deflation when cellular heritability is low (*e.g.*, $h^2_C = 0.01$). We also compared the power of the three methods. As shown in Figures 3 and S12, the power of all three methods increases as the cellular heritability (h^2_C) increases. CoMM-S^2 generally outperforms S-PrediXcan (both Enet and Ridge) at low or moderate levels of cellular heritability ($h^2_C = 0.01$ or 0.05). CoMM-S^2 and S-PrediXcan:Ridge have comparable performance for larger values of h^2_C. S-PrediXcan:Enet generally performs well at high levels of cellular heritability and when sparsity is low. The results also indicate that while the performance of CoMM-S^2 and S-PrediXcan:Ridge do not vary very much with the sparsity, the performance of S-PrediXcan:Enet depends on the sparsity levels.

5 Real Data Analysis

We applied CoMM-S^2 to two data sets, individual-level data NFBC1966 [30] and summary statistics from 14 traits (see Tables S1 and S2), with the transcriptome data from GEUVADIS Project [21] and Genotype-Tissue Expression (GTEx) Project [23], respectively. The NFBC dataset consists of information on ten quantitative traits. The ten quantitative traits include body mass index (BMI), systolic blood pressure (SysBP), diastolic blood pressure (DiaBP), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(LDL-C), triglycerides (TG), total cholesterol (TC), insulin levels, glucose levels and C-reactive protein (CRP). We also collected fourteen traits/diseases from multiple GWAS consortia including six diseases traits and eight BMI-related traits. Diseases traits are Alzheimer's disease (AD), coronary artery disease (CAD), autism spectrum disorder (ASD), schizophrenia (SCZ), type 2 diabetes (T2D), myocardial infarction (MI). There are eight summary statistics for BMI-related traits from 2017 GIANT Gene-Physical Activity Interaction Meta-analysis [12] including two sex-combined traits, BMI for physically active individuals (BMIPA), BMI for physically inactive individuals (BMIPI), and other six sex-specific traits, BMI for physically active individuals in men(BMIPAM), BMI for physically active individuals in women(BMIPAW), BMI for physically inactive individuals in men(BMIPIM), BMI for physically inactive individuals in women(BMIPIW), BMI adjusted for physical activity for individuals in men(BMIadjPAM) and BMI adjusted for physical activity for individuals in women(BMIadjPAW). For transcriptome data, GEUVADIS study contains 15,810 genes and GTEx project has gene expressions for 48 tissues, where the number of genes in each tissue ranges from 16,333 to 27,378.

5.1 Analysis of NFBC1966 datatset

As we have individual-level data for NFBC1966 dataset, we then applied both CoMM and CoMM-S^2 using individual-level data and summary statistics, respectively. We first analyzed the individual-level data from NFBC1966 with the transcriptome data from GEUVADIS using CoMM. The results from CoMM can be taken as a benchmark as it uses individual-level data. We then conducted single-variant analysis for NFBC1966 dataset to generate summary statistics. Finally, we applied CoMM-S^2 using summary statistics from NFBC1966 dataset. As CoMM-S^2 uses a reference panel to give estimates for LD, $\widehat{\mathbf{R}}_j$, we applied two different choices of reference panel, namely, 400 subsamples from NFBC1966 and European

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

samples from 1KG [7]. The scatter plots of LRT test statistics for CoMM against CoMM-S^2 using 400 subsamples from NFBC1966 are given in Figure 4 and the one using 1KG samples as reference data is shown in Figure S24 in the supplementary document. When we used the subsamples as the reference panel (in Figure 4), we notice that the test statistics from CoMM-S^2 are close to their counterpart with slope around 1 and R^2 ranging from 0.91 to 0.99. When the reference panel becomes 1KG as shown in Figure S13, one can observe that the test statistics in the null region ($\Lambda_g < 20.84 \equiv p$-value ¿ 5×10^{-6}) from both CoMM and CoMM-S^2 are roughly around the line with slope equals to 1. The test statistics in the non-null region ($\Lambda_g \geq 20.84 \equiv p$-value $\leq 5 \times 10^{-6}$) are inflated. This is primarily due to the reference panel we used to estimate correlations. When we applied sub-samples from NFBC1966 as reference panel data, this difference essentially disappeared. The reason for this phenomenon could be that despite that NFBC1966 dataset is a Finn's study from Europe, Finnish samples was shown its genetic distinctness in previous studies [31].Although this genetic discrepancy for Finnish population, we found that the inflation only appears in the non-null regions, which makes the use of 1KG as reference panel practically useful. Note that in these comparisons, we removed genes with cellular heritability less than 0.01 as the test statistics for tiny cellular heritability using CoMM is not reliable. The use of cutoff here is to make fair comparisons between CoMM and CoMM-S^2 as test statistics of CoMM degenerates to zero when cellular heritability is small (see Figure S8). In practice, we do not require this limit as CoMM-S^2 also works in tiny cellular heritability regions. To verify the phenomenon of degeneration, we compared CoMM-S^2 and CoMM with RL-SKAT [33] for genes with cellular heritability less than 0.01 as shown in Figure S14. The result is consistent with the simulation results from both Section 4.2 and our previous work [42].

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

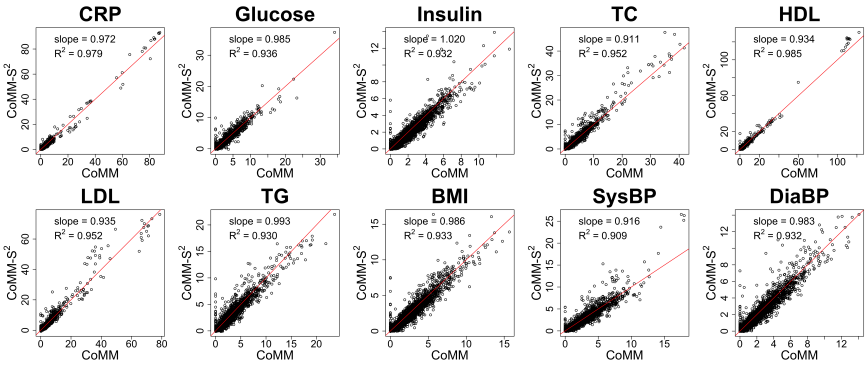


Figure 4: The scatter plot of LRT statistics for CoMM-IS vs CoMM, with transcriptiome data from GTEx tissue Adipose Subcutaneous, GWAS data from NFBC1966 dataset, and reference panel data from 400 subsamples of NFBC1966 dataset. We remove the points with cellular heritability less than 0.001.

5.2 Analysis of 14 traits

We performed the analysis for 14 GWAS summary statistics with transcriptome data from GTEx and the detailed information of these 14 traits are shown in Tables S1 and S2. Specifically, we applied CoMM-S^2 together with S-PrediXcan (both Enet and Ridge) to examine the associations between each pair of a gene and the complex trait. We display qq-plots for each trait across all tissues in Figure S15. After completing the analysis using three approaches, we conducted genomic control for each trait-tissue pair. The genome-wide significant threshold is set to be 5×10^{-6} based on Bonferroni correction.

The results indicate that CoMM-S^2 identified more significant associations than S-PrediXcan.The analysis for each individual summary statistics together with the transcriptome data for a tissue can be done around 20 min on a Linux platform with 2.6 GHz intel Xeon CPU E5-2690 with 30 720 KB cache and 96 GB RAM (0nly 6$\sim$7 GB RAM used) on 24 cores.

CoMM-S^2 is primarily developed to identify the associations between genes and complex traits using summary statistics. It is not only power in the region that cellular heritability is relatively large, but can also identify associations in the weak cellular heritability region.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We show both the number of unique genes passing genome-significance level across different tissues and the number of genes reported in the previous studies in Table 1 while the total number of genes identified across 48 tissues using CoMM-S^2 and S-PrediXcan (both Enet and Ridge) is shown in Table S3. For example, as shown in Table 1, CoMM-S^2 identified 199 unique genes across all tissues and 37 of them were previously reported in NHGRI-EBI GWAS Catalog [4] for AD while S-PrediXcan (both Enet and Ridge) identified 95 and 108 genes with 24 and 28 genes reported before, respectively. However, CoMM-S^2 identified 4,614 genes in total across 48 tissues while S-PrediXcan (both Enet and Ridge) identified only 398 and 664 genes in total, respectively, which indicates that there are more overlapped genes identified by CoMM-S^2 but the genes identified by S-PrediXcan (both Enet and Ridge) are more or less unique across tissues. Specifically, in GTEx adipose subcutaneous tissue, CoMM-S^2 identified seven genes in band 2q14.3, 12 genes in band 8p21.2, 14 genes in bands 11q12.1 and 11q12.2, 13 genes in bands 11q14.1 and 11q14.2, and 51 genes in band 19q13.32. Among them, 24 genes were reported to be associated with AD in previous studies [5, 6, 24, 9, 28, 19]. However, S-PrediXcan can only identify six of these 24 genes. Note that 18 out of these 24 genes have $h^2_C > 1\%$ and among the six genes having $h^2_C \leq 1\%$, S-PrediXcan (both Enet and Ridge) only identified one gene and none, respectively. In addition, gene *BIN1* was identified among all 48 tissues and cellular heritability for gene *BIN1* are larger than 10% in eight tissues, *e.g.*, $h^2_C = 13.5\%$ in brain cerebellum, $h^2_C = 13.7\%$ in brain cortex, $h^2_C = 23.3\%$ in esophagus muscularis mucosa, and highest in pancreas with $h^2_C = 30.6\%$. Hence, the associations between gene *BIN1* and AD in tissues with large h^2_C are more likely to be causal, where targeting *BIN1* might present novel AD therapy [35]. Gene *PTK2B* has largest cellular heritability in tissue brain cerebellum ($h^2_C = 32.4\%$), which was one of most significant genes associated with AD in a meta-analysis [20]. In a mouse model, overexpression of *PTK2B* improved the behavioral and molecular phenotype of a

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

strain of AD-linked mutated mice [11]. Gene *CLU* has the largest cellular heritability in the adrenal gland tissue($h^2_C = 22.7\%$), which has been found to be related to cholesterol synthesis, transport, uptake or metabolism in AD that links between cholesterol and AD pathogenesis [15].

For CAD, CoMM-S^2 identified 117 unique genes across all tissues and 15 of them were previously reported in NHGRI-EBI GWAS Catalog [4] while S-PrediXcan (both Enet and Ridge) identified 71 and 73 genes with 11 and 11 genes reported before, respectively. However, CoMM-S^2 identified 2,067 genes in total across all tissues while S-PrediXcan (both Enet and Ridge) identified only 109 and 184 genes in total, respectively. Specifically, in the artery aorta tissue, CoMM-S^2 identified two genes in band 2q33.2, one gene in band 3q22.3, three genes in band 6p24.1, two genes in band 6q25.3, eight genes in band 9p21.3, four genes in bands 12q24.11 and 12q24.12, six genes in band 13q34, seven genes in band 15q25.1, and five genes in band 19p13.2. Among them, nine genes were reported to be associated with CAD in previous studies [14, 25, 37]. However, S-PrediXcan can only identify two of these nine genes. Gene *NBEAL1* was identified to be genome-wide significant in eleven tissues. Among these eleven tissues, $h^2_C = 13.9\%$ in artery aorta, $h^2_C = 9.3\%$ in artery tibial, and $h^2_C = 10.2\%$ in pituitary are relative large, in which gene *NBEAL1* is likely to be causal for CAD.

The results of the identified genes for all 14 traits across tissues together with their corresponding test statistics and p-values for CoMM-S^2, S-PrediXcan (both Enet and Ridge) can be found in excel tables in the supplementary files.

6 Conclusion

In this article, we have developed a collaborative mixed model using summary statistics from GWAS to account for uncertainty in transcriptome imputation. We examined the

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

	CoMM-S^2	S-PrediXcan:Enet	S-PrediXcan:Ridge
AD	199(37)	105(28)	111(29)
ASD	2(0)	1(0)	5(0)
CAD	117(15)	71(11)	73(11)
MI	119(6)	81(7)	51(7)
SCZ	61(11)	57(17)	53(10)
T2D	160(27)	109(27)	110(22)
BMIPA	258(54)	137(35)	126(36)
BMIPI	53(11)	22(7)	28(7)
BMIPAM	31(4)	9(1)	17(4)
BMIPAW	123(31)	68(18)	67(21)
BMIPIM	12(2)	8(2)	7(4)
BMIPIW	24(4)	13(3)	11(2)
BMIadjPAM	104(21)	54(16)	76(21)
BMIadjPAW	208(45)	102(31)	107(30)

Table 1: The number of significant genes identified across the tissues at the significant level (5×10^{-6}) for each method. In this table, the same genes identified across the tissues are only counted once. The number within the parenthesis denoted the number of genes reported in NHGRI-EBI GWAS Catalog [4].

relationship between CoMM and CoMM-S^2. Our numerical results show that CoMM-S^2 has comparable performance as CoMM. CoMM-S^2 has several advantages over CoMM. First, CoMM-S^2 can be computationally more efficient than CoMM when applied to GWAS that have large sample sizes. This is because CoMM-S^2 is applied to summary statistics, which is faster to compute in large sample sizes, while CoMM is applied to individual-level data. Second, through empirical studies, we show that CoMM-S^2 has better performance when the cellular heritability is low. However, CoMM-S^2 is not without limitations. First, CoMM-S^2 cannot be utilized in a cross-tissue analysis, for example as illustrated in [17]. Furthermore, CoMM-S^2 cannot differentiate whether the identified genes are simply associated with the complex traits or if they are real causal effects. These are avenues for further research.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Funding

This work was supported in part by grant R-913-200-098-263 and R-913-200-127-263 from the Duke-NUS Medical School, AcRF Tier 2 (MOE2016-T2-2-029, MOE2018-T2-1-046 and MOE2018-T2-2-006) from the Ministry of Education, Singapore, grant No. 71501089, No. 11501579 and No. 71472023 from National Natural Science Foundation of China, and grant No. 22302815, No. 12316116 and No. 12301417 from the Hong Kong Research Grant Council.

References

[1] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications*, 9(1):1825, 2018.

[2] Christopher M Bishop. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[4] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The NHGRI-EBI gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2018.

[5] J Chapuis, F Hansmannel, Marc Gistelinck, A Mounier, C Van Cauwenberghe,

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

KV Kolen, F Geller, Y Sottejeau, D Harold, P Dourlen, et al. Increased expression of bin1 mediates alzheimer genetic risk by modulating tau pathology. *Molecular psychiatry*, 18(11):1225, 2013.

[6] Ganesh Chauhan, Hieab HH Adams, Joshua C Bis, Galit Weinstein, Lei Yu, Anna Maria Töglhofer, Albert Vernon Smith, Sven J Van Der Lee, Rebecca F Gottesman, Russell Thomson, et al. Association of Alzheimer's disease gwas loci with mri markers of brain aging. *Neurobiology of aging*, 36(4):1765–e7, 2015.

[7] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

[8] Mingwei Dai, Jingsi Ming, Mingxuan Cai, Jin Liu, Can Yang, Xiang Wan, and Zongben Xu. IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. *Bioinformatics*, 33(18):2882–2889, 2017.

[9] P Dourlen, FJ Fernandez-Gomez, C Dupont, B Grenier-Boley, C Bellenguez, H Obriot, R Caillierez, Y Sottejeau, J Chapuis, A Bretteville, et al. Functional screening of alzheimer risk loci identifies ptk2b as an in vivo modulator and early marker of tau pathology. *Molecular psychiatry*, 22(6):874, 2017.

[10] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

[11] Albert Giralt, Benoit de Pins, Carmen Cifuentes-Díaz, Laura López-Molina, Amel Thamila Farah, Marion Tible, Vincent Deramecourt, Stefan T Arold, Silvia

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ginés, Jacques Hugon, et al. Ptk2b/pyk2 overexpression improves a mouse model of Alzheimer's disease. *Experimental neurology*, 307:62–73, 2018.

[12] Mariaelisa Graff, Robert A Scott, Anne E Justice, Kristin L Young, Mary F Feitosa, Llilda Barata, Thomas W Winkler, Audrey Y Chu, Anubha Mahajan, David Hadley, et al. Genome-wide physical activity interactions in adiposity—a meta-analysis of 200,452 adults. *PLoS genetics*, 13(4):e1006528, 2017.

[13] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245, 2016.

[14] Jörg Hager, Yoichiro Kamatani, Jean-Baptiste Cazier, Sonia Youhanna, Michella Ghassibe-Sabbagh, Daniel E Platt, Antoine B Abchee, Jihane Romanos, Georges Khazen, Raed Othman, et al. Genome-wide association study in a lebanese cohort confirms phactr1 as a major determinant of coronary artery stenosis. *PloS one*, 7(6):e38663, 2012.

[15] J Hardy, N Bogdanovic, B Winblad, E Portelius, N Andreasen, A Cedazo-Minguez, and Henrik Zetterberg. Pathways to Alzheimer's disease. *Journal of internal medicine*, 275(3):296–303, 2014.

[16] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[17] Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M Zekavat, Zhao-

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

long Yu, Boyang Li, Jianlei Gu, Sydney Muchnik, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *bioRxiv*, page 286013, 2019.

[18] Jian Huang, Yuling Jiao, Jin Liu, and Can Yang. REMI: Regression with marginal information and its application in genome-wide association studies. *arXiv preprint arXiv:1805.01284*, 2018.

[19] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nature genetics*, page 1, 2019.

[20] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452, 2013.

[21] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506, 2013.

[22] Chuanhai Liu, Donald B Rubin, and Ying Nian Wu. Parameter expansion to accelerate em: the px-em algorithm. *Biometrika*, 85(4):755–770, 1998.

[23] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature genetics*, 45(6):580, 2013.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[24] Jing Ma, Jin-Tai Yu, and Lan Tan. Ms4a cluster in alzheimer's disease. *Molecular neurobiology*, 51(3):1240–1248, 2015.

[25] Christopher P Nelson, Anuj Goel, Adam S Butterworth, Stavroula Kanoni, Tom R Webb, Eirini Marouli, Lingyao Zeng, Ioanna Ntalla, Florence Y Lai, Jemma C Hopewell, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature genetics*, 49(9):1385, 2017.

[26] Alexandra C Nica, Stephen B Montgomery, Antigone S Dimas, Barbara E Stranger, Claude Beazley, Inês Barroso, and Emmanouil T Dermitzakis. Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLoS genetics*, 6(4):e1000895, 2010.

[27] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[28] Biswajit Padhy, Bushra Hayat, Gargi Gouranga Nanda, Pranjya Paramita Mohanty, and Debasmita Pankaj Alone. Pseudoexfoliation and Alzheimer's associated clu risk variant, rs2279590, lies within an enhancer element and regulates clu, ephx2 and ptk2b gene expression. *Human molecular genetics*, 26(22):4519–4529, 2017.

[29] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS genetics*, 7(2):e1001317, 2011.

[30] Chiara Sabatti, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, et al.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35, 2009.

[31] Elina Salmela et al. Genetic structure in finland and sweden: aspects of population history and gene mapping. 2012.

[32] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[33] Regev Schweiger, Omer Weissbrod, Elior Rahmani, Martina Müller-Nurasyid, Sonja Kunze, Christian Gieger, Melanie Waldenberger, Saharon Rosset, and Eran Halperin. RL-SKAT: an exact and efficient score test for heritability and set tests. *Genetics*, 207(4):1275–1283, 2017.

[34] Xingjie Shi, Yuling Jiao, Yi Yang, Ching-Yu Cheng, Can Yang, Xinyi Lin, and Jin Liu. VIMCO: Variational inference for multiple correlated outcomes in genome-wide association studies. *Bioinformatics*, page accepted, 2019.

[35] Meng-Shan Tan, Jin-Tai Yu, and Lan Tan. Bridging integrator 1 (bin1): form, function, and alzheimer's disease. *Trends in molecular medicine*, 19(10):594–603, 2013.

[36] Richard Eric Turner and Maneesh Sahani. *Two problems with variational expectation maximisation for time series models*, page 104–124. Cambridge University Press, 2011.

[37] Pim van der Harst and Niek Verweij. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research*, 122(3):433–443, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[38] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[39] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[40] Bo Wang and DM Titterington. Inadequacy of interval estimates corresponding to variational bayesian approximations. In *AISTATS*. Barbados, 2005.

[41] Heather E Wheeler, Kaanan P Shah, Jonathon Brenner, Tzintzuni Garcia, Keston Aquino-Michaels, Nancy J Cox, Dan L Nicolae, Hae Kyung Im, GTEx Consortium, et al. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS genetics*, 12(11):e1006423, 2016.

[42] Can Yang, Xiang Wan, Xinyi Lin, Mengjie Chen, Xiang Zhou, and Jin Liu. CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*, 2018.

[43] Yi Yang, Mingwei Dai, Jian Huang, Xinyi Lin, Can Yang, Min Chen, and Jin Liu. LPG: A four-group probabilistic approach to leveraging pleiotropy in genome-wide association studies. *BMC genomics*, 19(1):503, 2018.

[44] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[45] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The annals of applied statistics*, 11(3):1561, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/652263; this version posted June 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[46] Xiang Zhu and Matthew Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications*, 9(1):4361, 2018.

[47] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.