CoMM: A Collaborative Mixed Model That Integrates GWAS and eQTL Data Sets to Investigate the Genetic Architecture of Complex Traits

Bioinformatics and Biology Insights
Volume 13: 1–6

DOI: 10.1177/1177932219881435

Kar-Fu Yeung[1], Yi Yang[1], Can Yang[2] and Jin Liu[1]

[1]Centre for Quantitative Medicine, Programme in Health Services and System Research, Duke-NUS Medical School, Singapore. [2]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China.

ABSTRACT: Genome-wide association study (GWAS) analyses have identified thousands of associations between genetic variants and complex traits. However, it is still a challenge to uncover the mechanisms underlying the association. With the growing availability of transcriptome data sets, it has become possible to perform statistical analyses targeted at identifying influential genes whose expression levels correlate with the phenotype. Methods such as PrediXcan and transcriptome-wide association study (TWAS) use the transcriptome data set to fit a predictive model for gene expression, with genetic variants as covariates. The gene expression levels for the GWAS data set are then 'imputed' using the prediction model, and the imputed expression levels are tested for their association with the phenotype. These methods fail to account for the uncertainty in the GWAS imputation step, and we propose a collaborative mixed model (CoMM) that addresses this limitation by jointly modelling the multiple analysis steps. We illustrate CoMM's ability to identify relevant genes in the Northern Finland Birth Cohort 1966 data set and extend the model to handle the more widely available GWAS summary statistics.

KEYWORDS: Transcriptome-wide association studies, Linear mixed model, Probabilistic model, EM algorithm

RECEIVED: September 13, 2019. **ACCEPTED:** September 18, 2019.

TYPE: Commentary

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Duke-NUS Medical School (grant no. R-913-200-098-263) and Singapore's Ministry of Education Academic Research Fund (AcRF) Tier 2 (grant nos. MOE2016-T2-2-029, MOE2018-T2-1-046, and MOE2018-T2-2-006).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Jin Liu, Centre for Quantitative Medicine, Programme in Health Services and System Research, Duke-NUS Medical School 169857, Singapore. Email: jin.liu@duke-nus.edu.sg

COMMENT ON: Yang C, Wan X, Lin X, Chen M, Zhou X, Liu J. CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*. 2018;35:1644-1652. doi:10.1093/bioinformatics/bty865. PubMed PMID: 30295737. https://www.ncbi.nlm.nih.gov/pubmed/30295737

Introduction

In the last decade, genome-wide association studies (GWASs) have identified thousands of genetic variants associated with complex traits. However, an understanding of the mechanisms linking a genetic variant to a complex trait is relatively limited. As many GWAS loci are located outside coding regions and regulatory variation plays an important role in shaping observed traits, gene expression has been proposed as an informative intermediate phenotype.[1] Incorporating regulatory information into statistical analyses provides us with a principled approach to study the genetic contribution to complex traits through the regulation of gene expression.

One approach to incorporate functional information is to do so without explicitly modelling the relationship between gene expression and phenotype. Sequence Kernel Association Tests (SKATs),[2] for instance, can be used to prioritize genetic variants based on functional annotation. It is based on the idea that genetic variants with known biological functions are more likely to be associated with a trait. Hence, when testing for association between genetic variants and a trait, the genetic variants are prioritized by placing larger weights on those with known functions.[2]

Recently, the growing availability of transcriptome data has given rise to methods that evaluate genetically regulated gene expression using both GWAS and transcriptome data sets. Large-scale transcriptome data sets, which contain information on genotypes and gene expression levels, include the Genotype-Tissue Expression Consortium (GTEx),[3] the Genetic European in Health and Disease (GEUVADIS) Project,[4] Braineac,[5] Depression Genes and Networks[6] and eQTLGen.[7]

Methods that have leveraged on both GWAS and transcriptome data include PrediXcan[8,9] and transcriptome-wide association study (TWAS)[10] and generally proceed in 3 steps. First, using transcriptome data, they fit predictive models for gene expression with genetic variants near a gene as covariates. PrediXcan proposed the use of elastic net regression or ridge regression to build a predictive model, while TWAS proposed the use of a linear mixed model. The fitted models are then used to predict the gene expression levels for individuals in the GWAS data set. Finally, a simple linear regression is used to examine the association between the predicted expression levels and the complex trait in the GWAS data set.

Methods that proceed in such a stage-wise manner do not account for the uncertainty that arises when imputing the gene expression levels in the GWAS data set, which may lead to a loss in statistical power. To address this limitation, we proposed the collaborative mixed model (CoMM),[11] which accounts for the uncertainty in the 'imputation' model by jointly fitting the imputation and the association analysis models.

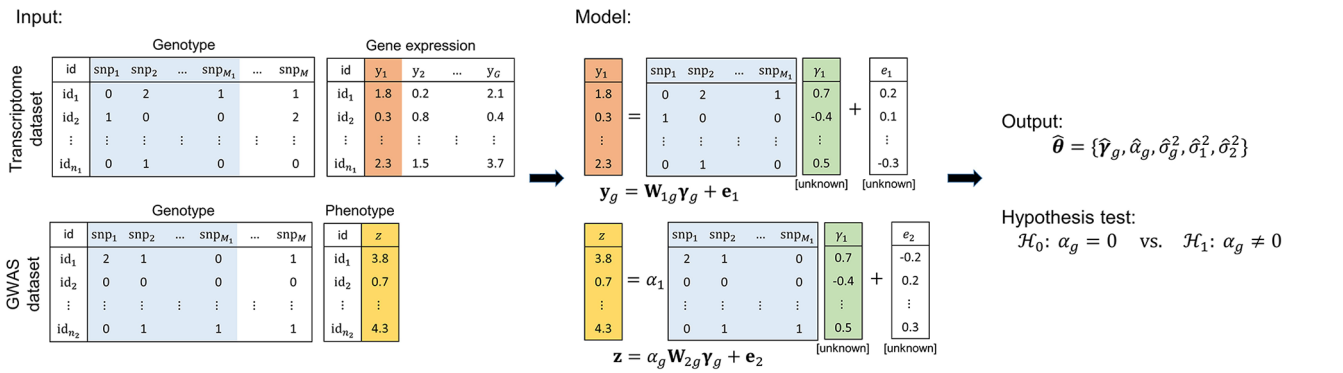

Figure 1. Schematic of CoMM. The transcriptome and GWAS data sets are used to fit the parameters in the model given by equations (1) and (2). The parameter estimates are used to evaluate the likelihood ratio test statistic, which tests for association between the phenotype and genetically regulated gene expression. CoMM indicates collaborative mixed model; GWAS, genome-wide association studies.

Method

Suppose that the gene expression levels for G genes and the allele counts for M single-nucleotide polymorphisms (SNPs) are measured for n_1 samples in the transcriptome data set and that the phenotype values and the allele counts for the same M SNPs are measured for n_2 samples in the GWAS data set. The transcriptome data set consists of the n_1 by G gene expression data matrix $\mathbf{Y}$ and the n_1 by M genotype data matrix $\mathbf{W}_1$. The GWAS data set consists of the phenotype vector $\mathbf{z}$ and the n_2 by M genotype data matrix $\mathbf{W}_2$. We predict the gene expression levels one gene at a time and denote the gene expression levels at the gth gene by $\mathbf{y}_g$. As we are interested in the variation in gene expression attributable to variation in its cis-SNPs, we model the gene expression levels and phenotype value using only nearby SNPs. Let $\mathbf{W}_{1g}$ and $\mathbf{W}_{2g}$ denote the genotype matrix corresponding to the gene's nearby SNPs, in the transcriptome data set and the GWAS data set, respectively. Let M_g denote the number of SNPs corresponding to the gth gene. We assume that $\mathbf{y}_g$ is mean centred, and $\mathbf{W}_{1g}$ and $\mathbf{W}_{2g}$ are standardized (columns have zero mean and unit variance).

In the CoMM, we first model the relationship between gene expression $\mathbf{y}_g$ and genotype $\mathbf{W}_{1g}$ in the transcriptome data set

$$\mathbf{y}_g = \mathbf{W}_{1g}\gamma_g + \mathbf{e}_1 \tag{1}$$

where γ_g is an M_g vector of SNP effects on the gene expression level, and $\mathbf{e}_1 \sim N(0, \sigma_1^2 \mathbf{I}_{n_1})$ is an n_1 vector representing the error associated with the expression level. Next, we model the relationship between phenotype $\mathbf{z}$ and genotype $\mathbf{W}_{2g}$ in the GWAS data set as

$$\mathbf{z} = X\beta + \alpha_g \mathbf{W}_{2g}\gamma_g + \mathbf{e}_2 \tag{2}$$

where $\mathbf{X}$ contains covariates that control for population stratification and other confounding variables, β is a vector of fixed-effects coefficients, and $\mathbf{e}_2 \sim N(0, \sigma_2^2 \mathbf{I}_{n_2})$ is an n_2 vector representing the error associated with the phenotype.

The quantity of interest is α_g, the effect of gene g's expression level on the phenotype. In addition, we assume the prior distribution on γ_g

$$\gamma_g \sim N\left(0, \sigma_g^2 \mathbf{I}_{M_g}\right) \tag{3}$$

effectively treating the effects of genotype on gene expression as random. An accelerated expectation–maximization (EM) algorithm using parameter expansion[12] is used to estimate all parameters in the joint model given by equations (1) and (2). Figure 1 summarizes the data input, the joint model, and output for CoMM.

As our objective is to evaluate whether a gene is associated with the phenotype via gene expression, we perform the hypothesis test

$$\mathcal{H}_0 : \alpha_g = 0, \text{v.s.} \mathcal{H}_1 : \alpha_g \neq 0 \tag{4}$$

The likelihood ratio test is used, and the test statistic is given by

$$\Lambda_g = 2\left(\log \Pr\left(\mathbf{y}_g, \mathbf{z}, \gamma_g \mid \hat{\theta}\right) - \log \Pr\left(\mathbf{y}_g, \mathbf{z}, \gamma_g \mid \hat{\theta}_0\right)\right)$$

where $\hat{\theta}$ and $\hat{\theta}_0$ contain the parameter estimates obtained under the full model and $\mathcal{H}_0$, respectively. The test statistic Λ_g is asymptotically distributed as $\chi^2_{df=1}$ under the null hypothesis. The key thing to note is that the likelihood reflects the uncertainty in both the imputation and association analysis models (equations (1) and (2)). As such, the CoMM test statistic for expression-trait association takes into account the uncertainty in the imputation model.

Results in the Northern Finland Birth Cohort 1966

We analysed the GWAS data set from the Northern Finland Birth Cohort 1966 (NFBC1966)[13] with the aid of transcriptome data from GTEx (tissue: subcutaneous adipose).[3] The NFBC1966 data set records traits such as body mass index (BMI), low-density lipoprotein cholesterol (LDL), triglycerides (TGs), total cholesterol (TC), systolic blood pressure (SysBP), and diastolic blood pressure (DiaBP).

The CoMM returns a larger number of significant findings than PrediXcan and SKAT, as indicated by the QQ-plots of the P values (Figure 2). In particular, CoMM reported 12 significant genes associated with triglyceride (TG) levels, whereas

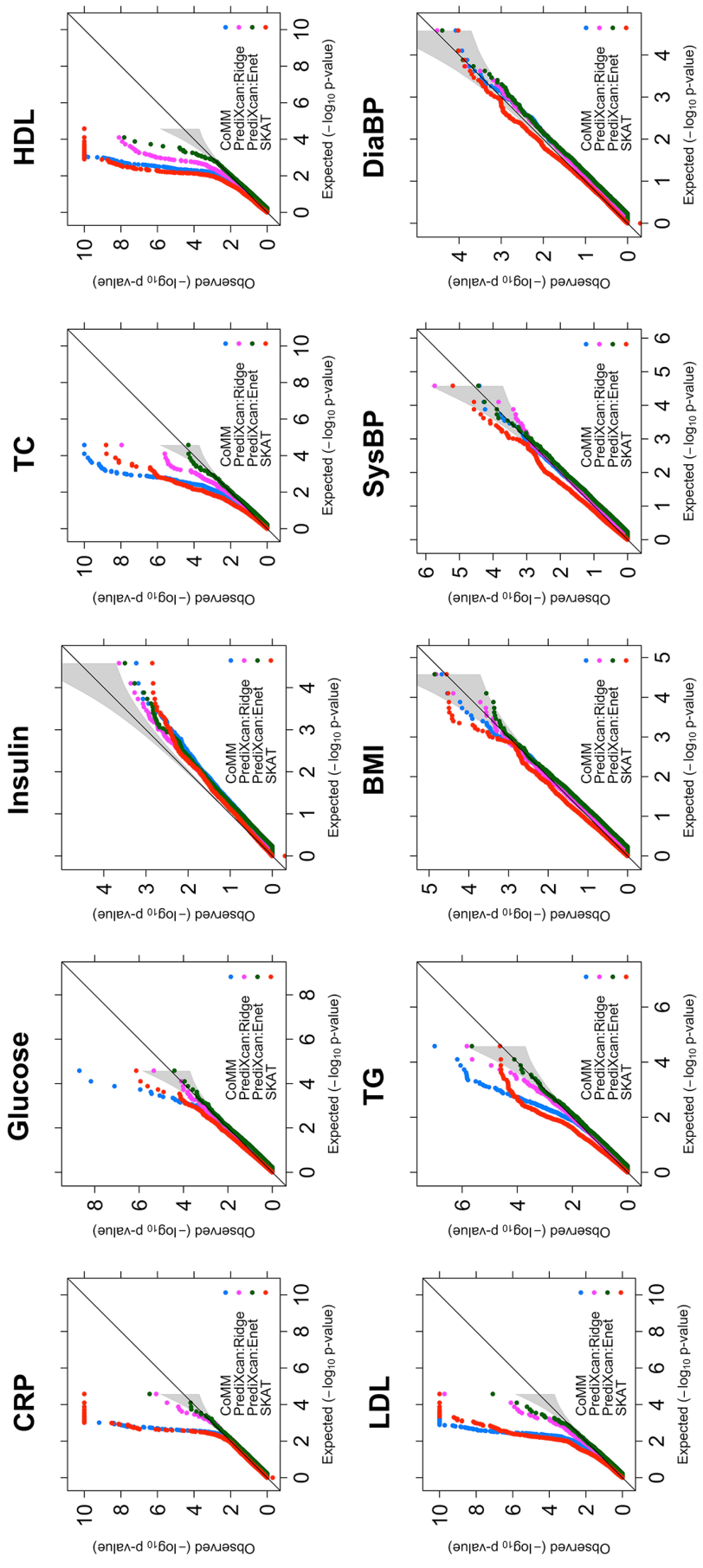

Figure 2. The QQ-plots of *P*-values for the quantitative traits in NFBC1966.
CoMM shows a larger statistical power than PrediXcan:Ridge and PrediXcan:Enet across all traits. For glucose, total cholesterol and triglyceride levels, CoMM also outperforms SKAT. NFBC1966 indicates Northern Finland Birth Cohort 1966; CoMM, collaborative mixed model; CRP, C-reactive protein; TC, total cholesterol; HDL, high-density lipoprotein cholesterol; LDL, low-density lipoprotein cholesterol; TG, triglyceride; BMI, body mass index; SysBP, systolic blood pressure; DiaBP, diastolic blood pressure.

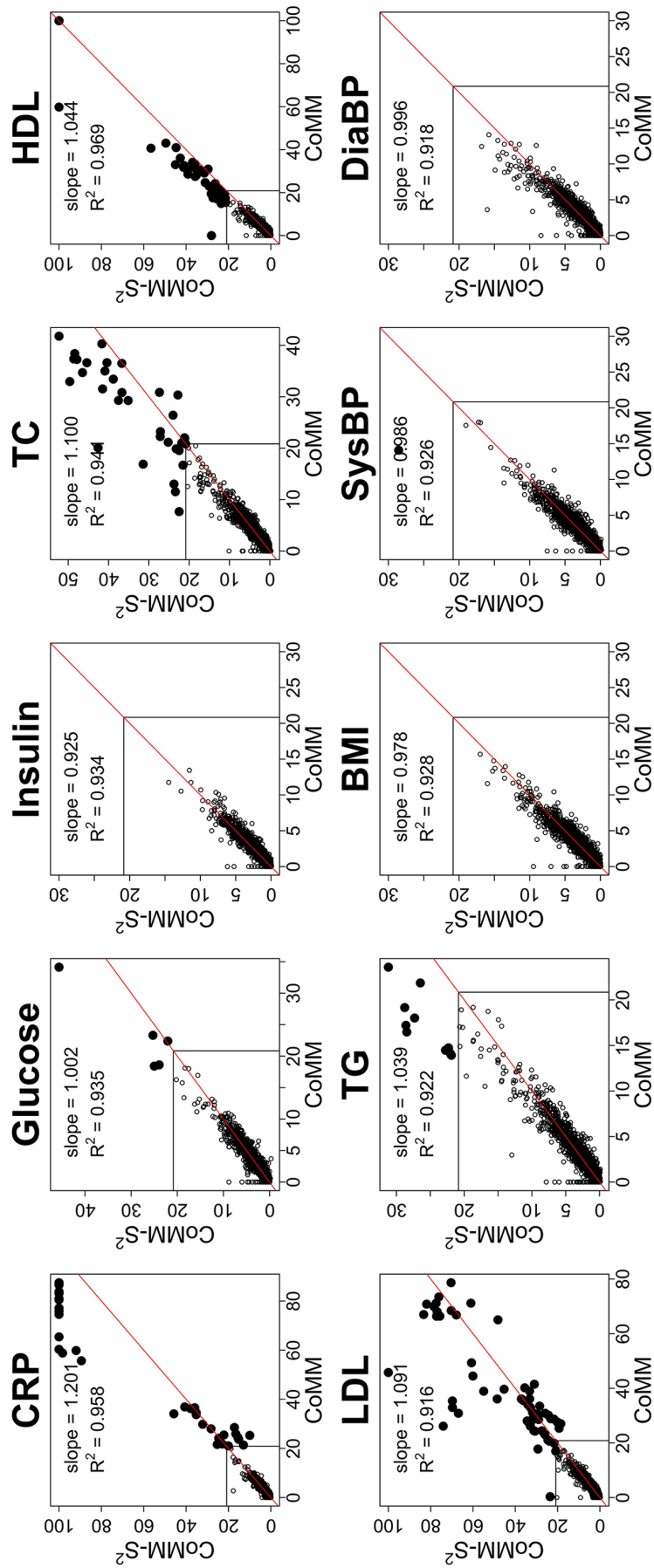

Figure 3. Scatter plot of test statistics from the likelihood ratio test for CoMM-S^2 versus CoMM, using GTEx (tissue: subcutaneous adipose) transcriptome data, NFBC1966 GWAS data, and 1000 Genomes Project reference panel data.
In the null region indicated by the bottom-left boxed region of the plot, the test statistics for CoMM and CoMM-S^2 are close to the 45° line. CoMM indicates collaborative mixed model; NFBC1966, Northern Finland Birth Cohort 1966; CRP, C-reactive protein; TC, total cholesterol; HDL, high-density lipoprotein cholesterol; LDL, low-density lipoprotein cholesterol; TG, triglyceride; BMI, body mass index; SysBP, systolic blood pressure; DiaBP, diastolic blood pressure.

PrediXcan:Enet, PrediXcan:Ridge, and SKAT reported 2, 1, and 0 significant genes, respectively. Among the 12 identified genes, 2 (*OST4* and *EIF2B4*) have nonnegligible cellular heritability ($h_C^2 = 2.03\%$ and 1.33%, respectively) and have reported associations with TG in previous studies.[14,15] In this instance, CoMM performs better than SKAT due to the use of gene regulation information in the GTEx data, and outperforms PrediXcan by taking into account the uncertainty in the imputation model.

Extension of CoMM to Analyse GWAS Summary Data

A limitation of CoMM is that it requires individual-level data and is unable to make use of large-scale GWAS that provide only summary statistics. To capitalize on these GWAS, we extend CoMM so that summary statistics, in the form of estimated SNP effect sizes and their variances, can replace the role of individual-level GWAS data.

We adapt the method of Zhu and Stephens,[16] which made use of summary statistics by introducing a regression with summary statistics (RSS) likelihood in a Bayesian framework. As gene expression levels are modelled using multiple SNPs, we additionally require information on the correlations among SNPs (linkage disequilibrium). Fortunately, such information is available in public data sets such as the 1000 Genomes Project Consortium.[17] In CoMMs for GWAS summary statistics (CoMM-S^2),[18] the association between phenotype and genetically regulated gene expression is evaluated by combining the distribution for individual-level transcriptome data with an RSS distribution for GWAS summary statistics, while taking into account linkage disequilibrium as estimated from a reference panel.

Even though CoMM-S^2 utilizes GWAS summary statistics, it has comparable performance as CoMM. To illustrate the performance of CoMM-S^2 relative to CoMM, we use NFBC1966 as the GWAS data set, GTEx as the transcriptome data set and the 1000 Genomes Project as a reference panel to estimate linkage disequilibrium. In general, the test statistic values from CoMM-S^2 are close to their corresponding values from CoMM: the regression slope is around 1 and R^2 ranges from 0.91 to 0.99 (Figure 3). The close correspondence in test statistic values is most apparent in the null region. In the nonnull region, the test statistics for CoMM-S^2 may be inflated (Figure 3). One possible reason for this is linkage disequilibrium misspecification,[18] due to the genetic differences between the Finnish cohort in NFBC1966 and the European sample in the 1000 Genomes Project.[19] Nonetheless, as a strong inflation occurs only when the CoMM statistics is large, CoMM-S^2 maintains a reasonable false-positive rate in the presence of misspecified linkage disequilibrium.

Finally, we note that both CoMM and CoMM-S^2 are designed for single-tissue analysis. When a multi-tissue transcriptome data set is available, an approach that takes into account genetic correlation across tissues may be preferable. Such an approach would be better equipped to identify biologically relevant tissues for each gene, and may also provide an increase in statistical power for tissues that are difficult to obtain.[20] Recently, 2 multi-tissue approaches, UTMOST[20] and MultiXcan,[21] have been proposed. They are more powerful than single-tissue approaches in expression-trait association analyses. However, they ignore the uncertainty due to the imputation step, and similar to what has been proposed for CoMM and CoMM-S^2, the ability to detect relevant genes can be further improved by combining the imputation model and association analysis model via a unified likelihood framework. This remains a promising avenue for further research.

Acknowledgement

The authors thank the National Supercomputing Centre, Singapore, for providing computational resources for the project.

Author Contributions

JL and CY conceived this commentary. YY performed the analysis and computation. K-FY wrote the manuscript in consultation with JL and CY.

Data Availability and Implementation

The developed R package is available at https://github.com/gordonliu810822/CoMM.

ORCID iDs

Can Yang https://orcid.org/0000-0002-4407-3055
Jin Liu https://orcid.org/0000-0002-5707-2078

REFERENCES

1. Gallagher MD, Chen-Plotkin AS. The post-GWAS era: from association to function. *Am J Hum Genet*. 2018;102:717-730. doi:10.1016/j.ajhg.2018.04.002.
2. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the Sequence Kernel Association Test. *Am J Hum Genet*. 2011;89:82-93. doi:10.1016/j.ajhg.2011.05.029.
3. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017;550:204-213. doi:10.1038/nature24277.
4. Lappalainen T, Sammeth M, Friedländer MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501:506-511. doi:10.1038/nature12531.
5. Ramasamy A, Trabzuni D, Guelfi S, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat Neurosci*. 2014;17:1418-1428. doi:10.1038/nn.3801.
6. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res*. 2014;24:14-24. doi:10.1101/gr.155192.113.
7. Võsa U, Claringbould A, Westra HJ, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*. 2018:447367. doi:10.1101/447367.
8. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47:1091-1098. doi:10.1038/ng.3367.
9. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*. 2018;9:1825. doi:10.1038/s41467-018-03621-1.
10. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48:245-252. doi:10.1038/ng.3506.

11. Yang C, Wan X, Lin X, Chen M, Zhou X, Liu J. CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*. 2018;35:1644-1652. doi:10.1093/bioinformatics/bty865.
12. Liu C, Rubin DB, Wu YN. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*. 1998;85:755-770. doi:10.1093/biomet/85.4.755.
13. Sabatti C, Service SK, Hartikainen AL, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*. 2009;41:35-46. doi:10.1038/ng.271.
14. Rundblad A, Larsen SV, Myhrstad MC, et al. Differences in peripheral blood mononuclear cell gene expression and triglyceride composition in lipoprotein subclasses in plasma triglyceride responders and non-responders to omega-3 supplementation. *Genes Nutr*. 2019;14:10. doi:10.1186/s12263-019-0633-y.
15. Chen YC, Xu C, Zhang JG, et al. Multivariate analysis of genomics data to identify potential pleiotropic genes for type 2 diabetes, obesity and dyslipidemia using meta-CCA and gene-based approach. *PLoS ONE*. 2018;13:e0201173. doi:10.1371/journal.pone.0201173.
16. Zhu X, Stephens M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann Appl Stat*. 2017;11:1561-1592.
17. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56-65. doi:10.1038/nature11632.
18. Yang Y, Shi X, Jiao Y, et al. CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. *bioRxiv*. 2019:652263. doi:10.1101/652263.
19. Salmela E. *Genetic Structure in Finland and Sweden: Aspects of Population History and Gene Mapping* [dissertation]. Helsinki, Finland: University of Helsinki; 2012.
20. Hu Y, Li M, Lu Q , et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet*. 2019;51:568-576. doi:10.1038/s41588-019-0345-7.
21. Barbeira AN, Pividori MD, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet*. 2019;15:e1007889. doi:10.1371/journal.pgen.1007889.