STATISTICAL APPROACHES IN OMICS DATA ASSOCIATION STUDIES

EDITED BY: Qi Yan, Jiebiao Wang and Zhonghua Liu
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-362-7
DOI 10.3389/978-2-88976-362-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

STATISTICAL APPROACHES IN OMICS DATA ASSOCIATION STUDIES

Topic Editors:
Qi Yan, Columbia University, United States
Jiebiao Wang, University of Pittsburgh, United States
Zhonghua Liu, The University of Hong Kong, Hong Kong, SAR China

Citation: Yan, Q., Wang, J., Liu, Z., eds. (2022). Statistical Approaches in Omics Data Association Studies. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-362-7

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 December 2020
doi: 10.3389/fgene.2020.608630

Exploring the Relationship Between Psychiatric Traits and the Risk of Mouth Ulcers Using Bi-Directional Mendelian Randomization

Kai Wang[1], Lin Ding[1], Can Yang[2], Xingjie Hao[1] and Chaolong Wang[1]**

[1] Key Laboratory for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2] Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

Background: Although the association between mouth ulcers and psychiatric traits has been reported by observational studies, their causal relationship remains unclear. Mendelian randomization (MR), powered by large-scale genome-wide association studies (GWAS), provides an opportunity to clarify the causality between mouth ulcers and psychiatric traits.

Methods: We collected summary statistics of mouth ulcers (sample size n = 461,106) and 10 psychiatric traits from the largest publicly available GWAS on Europeans, including anxiety disorder (n = 83,566), attention deficit/hyperactivity disorder (n = 53,293), autism spectrum disorder (n = 46,350), bipolar disorder (n = 51,710), insomnia (n = 1,331,010), major depressive disorder (n = 480,359), mood instability (n = 363,705), neuroticism (n = 168,105), schizophrenia (n = 105,318), and subjective wellbeing (n = 388,538). We applied three two-sample bi-directional MR analysis methods, namely the Inverse Variance Weighted (IVW) method, the MR pleiotropy residual sum and outlier (MR-PRESSO) method, and the weighted median method, to assess the causal relationship between each psychiatric trait and mouth ulcers.

Results: We found significant effects of autism spectrum disorder, insomnia, major depressive disorder, and subjective wellbeing on mouth ulcers, with the corresponding odds ratio (*OR*) from the IVW method being 1.160 [95% confidence interval (*CI*): 1.066–1.261, $P = 5.39 \times 10^{-4}$], 1.092 (1.062–1.122, $P = 3.37 \times 10^{-10}$), 1.234 (1.134–1.342, $P = 1.03 \times 10^{-6}$), and 0.703 (0.571–0.865, $P = 8.97 \times 10^{-4}$), respectively. We also observed suggestive evidence for mood instability to cause mouth ulcers [IVW, OR = 1.662 (1.059–2.609), P = 0.027]. These results were robust to weak instrument bias and heterogeneity. We found no evidence on causal effects between other psychiatric traits and mouth ulcers, in either direction.

Conclusion: Our findings suggest a protective effect of subjective wellbeing and risk effects of autism spectrum disorder, insomnia, major depressive disorder, and mood instability on mouth ulcers. These results clarify the causal relationship between psychiatric traits and the development of mouth ulcers.

Keywords: psychiatric traits, mouth ulcers, Mendelian randomization, causality, GWAS summary statistics

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong, Hong Kong

Reviewed by:
Xihao Li,
Harvard University, United States
Maggie Haitian Wang,
The Chinese University of Hong Kong, China

****Correspondence:***
Xingjie Hao
xingjie@hust.edu.cn
orcid.org/0000-0003-1535-9860
Chaolong Wang
chaolong@hust.edu.cn
orcid.org/0000-0003-3945-1012

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 September 2020*
Accepted: *09 November 2020*
Published: *16 December 2020*

Citation:
Wang K, Ding L, Yang C, Hao X and Wang C (2020) Exploring the Relationship Between Psychiatric Traits and the Risk of Mouth Ulcers Using Bi-Directional Mendelian Randomization. Front. Genet. 11:608630. doi: 10.3389/fgene.2020.608630

INTRODUCTION

A mouth ulcer (also termed oral ulceration) is an ulcer that occurs on the mucous membrane of the oral cavity, involving damage to both epithelium and lamina propria (Scully, 2008; Tugrul et al., 2016). Mouth ulcers are prevalent worldwide, affecting nearly 25% of young adults and a higher proportion of children (Scully, 2006; Paleri et al., 2010; Tugrul et al., 2016; Dudding et al., 2019). Although mouth ulcers do not pose a substantial health burden, they can interfere with daily activities (such as speaking or swallowing) and have detrimental effects on individual quality of life, overall wellbeing, and social interaction (Huling et al., 2012; Almoznino et al., 2014; Al-Omiri et al., 2015). Furthermore, mouth ulcers are one of the common clinical signals of several serious diseases, such as oral cancer, gastrointestinal diseases, and human immunodeficiency virus infection (Paleri et al., 2010; Bilodeau and Lalla, 2019). Besides, mouth ulcers have been reported to associate with head and neck cancer, pancreatic cancer, breast cancer, and prostate cancer by a recent epidemiology study (Qin et al., 2018).

The high prevalence of mouth ulcers and its undesired impact on life quality have motivated numerous studies on the etiology and efficient therapy of this disease. Recurrent aphthous stomatitis (RAS) is the most common cause, followed by local trauma, malignancy, and infection (Paleri et al., 2010; Gavic et al., 2014; Al-Omiri et al., 2015; Bilodeau and Lalla, 2019). Nevertheless, the pathogenesis of mouth ulcers is still poorly understood. Psychiatric disorders are potential risk factors for mouth ulcers, as suggested by observational studies. For example, patients with depression and anxiety are more likely to develop mouth ulcers according to a series of observational studies (Huling et al., 2012; Alshahrani and Baccaglini, 2014; Ma et al., 2015; Ge, 2018); high levels of psychological stress were found in mouth ulcers patients (Gallo Cde et al., 2009); depression and neuroticism were genetically correlated with mouth ulcers (Dudding et al., 2019); and a transitory rise in salivary cortisol and/or changes in immunoregulatory activity caused by psychiatric disorders were linked to mouth ulcers (MacGregor et al., 1969; Redwine et al., 2003; Slebioda and Dorocka-Bobkowska, 2019). These observations together lead to a hypothesis that psychiatric disorders may trigger mouth ulcers. Nevertheless, the causal relationship between psychiatric traits and mouth ulcers remains largely unclear.

Abbreviations: GWAS, genome-wide association studies; LD, linkage disequilibrium; SNP, single nucleotide polymorphisms; MR, Mendelian randomization; IV, instrumental variable; InSIDE, Instrument Strength Independent of Direct Effect; IVW, Inverse Variance Weighted; MR-PRESSO, MR pleiotropy residual sum and outlier; OR, odds ratio; CI, confidence interval; SD, standard deviation; UKB, UK Biobank; 23andMe, 23andMe company; PGC29, the Psychiatric Genomics Consortium, 29 European samples; deCODE, deCODE Genetics company; GenScot, Generation Scotland: Scottish Family Health Study; GERA, Genetic Epidemiology Research on Adult Health and Aging Study; iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research; SSGAC, Social Science Genetics Association Consortium; PGC, the Psychiatric Genomics Consortium; GPC, the Genetics of Personality Consortium; RAS, Recurrent aphthous stomatitis; ADHD, attention deficit/hyperactivity disorder; ASD, autism spectrum disorder; BIP, bipolar disorder; MDD, major depressive disorder; SCZ, schizophrenia.

With the development of large-scale GWAS and Mendelian randomization (MR), causal inference between complex traits and diseases has become possible (Lawlor et al., 2008; Hartwig et al., 2017). The MR approach uses genetic variants, such as single nucleotide polymorphisms (SNPs), associated with a modifiable exposure (e.g., a psychiatric trait) as the instrumental variables (IVs) to estimate the causality between this exposure and an outcome of interest (e.g., mouth ulcers) (Lawlor et al., 2008). The basic idea is that SNPs associated with the exposure, which were randomly passed from parents to offsprings during meiosis irrespective of confounders, would also be associated with the outcome if the exposure is causally associated with the outcome. To ensure the validity of MR for causal inference, the IVs need to satisfy three model assumptions: (a) associated with the exposure (the relevance assumption); (b) independent of any confounder of the exposure-outcome association (the independence assumption); and (c) only affect the outcome through the exposure (the exclusion restriction assumption) (Lawlor et al., 2008; Hartwig et al., 2016). Recent studies have found that the exclusion restriction assumption may be too strong given the polygenic architecture of complex traits/disease and the ubiquity of pleiotropy (Zhao et al., 2019). Instead, an alternative weaker assumption named Instrument Strength Independent of Direct Effect (InSIDE) has been proposed (Bowden et al., 2015). This assumption allows for the direct effects of IVs on the outcome, assuming that genetic associations with the exposure are independent of the direct effects (Bowden et al., 2015). Two-sample MR refers to the application of MR on GWAS summary statistics of the exposure and the outcome from two independent samples, which can overcome the winner's curse and maximize the statistical power (Burgess et al., 2016). Further information about the assumptions and interpretations of MR can be found elsewhere (Haycock et al., 2016; Zheng et al., 2017; Davies et al., 2018).

As the pathogenesis of mouth ulcers is complicated, identification of causal risk factors will be useful to facilitate both the prevention and treatment of the disease. In this study, we aim to systematically investigate the causal relationship between mouth ulcers and psychiatric traits. We conducted two-sample bi-directional MR analyses using publicly available GWAS summary statistics of mouth ulcers and 10 psychiatric traits, including anxiety disorder, attention deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), bipolar disorder (BIP), insomnia, major depressive disorder (MDD), mood instability, schizophrenia (SCZ), neuroticism, and subjective wellbeing (Pardiñas et al., 2018; Turley et al., 2018; Wray et al., 2018; Demontis et al., 2019; Grove et al., 2019; Jansen et al., 2019; Purves et al., 2019; Stahl et al., 2019; Ward et al., 2020).

MATERIALS AND METHODS

Data Collection

We collected GWAS summary statistics of eight psychiatric traits and mouth ulcers from published studies with the largest sample sizes of European ancestry (**Table 1**). In addition, we only

TABLE 1 | Description of GWAS consortiums used for each trait.

Trait	Population	Sample size (cases/controls)	Sample overlap[$]	Data source	References
Anxiety disorders	Europeans	25,453/58,113	18.2%	UKB	Purves et al., 2019
ADHD	96% Europeans	19,099/34,194	0	PGC	Demontis et al., 2019
ASD	Europeans	18,381/27,969	0	PGC	Grove et al., 2019
BIP	Europeans	20,352/31,358	0	PGC	Stahl et al., 2019
Insomnia	Europeans	397,972/933,038	29.0%	UKB, 23andMe	Jansen et al., 2019
MDD	Europeans	135,458/344,901	6.5%	UKB, 23andMe, PGC29, deCODE, GenScot, GERA, iPSYCH	Wray et al., 2018
Mood instability	Europeans	157,039/206,666	78.9%	UKB	Ward et al., 2020
Neuroticism	Europeans	168,105	22.5%	UKB, GPC	Turley et al., 2018
SCZ	Europeans	40,675/64,643	0	CLOZUK, PGC	Pardiñas et al., 2018
Subjective wellbeing	Europeans	388,538	8.8%	UKB, 23andMe, SSGAC	Turley et al., 2018
Mouth ulcers	Europeans	47,079/414,027	–	UKB	Dudding et al., 2019

ADHD, attention deficit/hyperactivity disorder; ASD, autism spectrum disorder; BIP, bipolar disorder; MDD, major depressive disorder; SCZ, Schizophrenia; UKB, the UK Biobank; 23andMe, 23andMe company; PGC29, the Psychiatric Genomics Consortium, 29 European samples; deCODE, deCODE Genetics company; GenScot, Generation Scotland: Scottish Family Health Study; GERA, Genetic Epidemiology Research on Adult Health and Aging Study; iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research; SSGAC, Social Science Genetics Association Consortium; PGC, the Psychiatric Genomics Consortium; GPC, the Genetics of Personality Consortium. [$]The overlapping sample size divided by the larger sample size of the corresponding psychiatric trait and mouth ulcers. For the two quantitative traits, neuroticism and subjective wellbeing, the total sample size was present.

obtained summary statistics of significant associated SNPs for anxiety disorders and mood instability due to restricted access to the GWAS summary statistics of these two traits. GWAS on mouth ulcers were based on the UK Biobank (UKB), in which all participants were asked about their oral health in the baseline questionnaire. "Mouth ulcers (yes/no)" was defined as having mouth ulcers within the last year. **Supplementary Table S1** lays out the definitions of 10 psychiatric traits.

Patient and Public Involvement

Because this study used published GWAS summary statistics available in the public domain, specific ethical review or consent from study participants was not sought.

Statistical Analyses

The overall workflow of our analyses was summarized in **Figure 1**. We took the following steps to choose valid instrumental SNPs given the assumptions of MR. Firstly, candidate IVs were restricted to those of genome-wide significant association ($P < 5 \times 10^{-8}$) with the exposure (e.g., a psychiatric trait). Secondly, we pruned the candidate IVs to independent SNPs ($r^2 < 0.05$, window size = 1 Mb), keeping those with the smallest *P*-values, based on linkage disequilibrium (LD) calculated from the 1000 Genomes Project Phase 3 European dataset using PLINK v.1.90 (Chang et al., 2015; Consortium, 2015). Thirdly, because bi-directional MR assumes no overlap or LD between the IVs for the exposure and the outcome (Davey Smith and Hemani, 2014), we excluded SNPs in LD ($r^2 > 0.05$) with the significant SNPs for mouth ulcers. Finally, we removed potential pleiotropic SNPs by excluding SNPs of suggestive association ($P < 10^{-5}$) with mouth ulcers (Au Yeung et al., 2017; Zeng and Zhou, 2019). The remaining SNPs were used as valid IVs to conduct MR analyses. Valid IVs for all exposure-outcome pairs are listed in **Supplementary Tables S2–S19**. For each IV, we computed the *F* statistic to quantify whether it was strongly associated with the exposure (Lawlor et al., 2008). For multiple IVs, we computed the *F* statistic as the mean of the *F* statistics of individual IVs and the 95% confidence interval (*CI*) by 10,000 bootstraps (Burgess et al., 2016).

Given an IV, the causal effect of exposure (X) on outcome (Y), β_{XY} can be estimated as $\hat{\beta}_{IV} = \hat{\beta}_{ZY}/\hat{\beta}_{ZX}$ where $\hat{\beta}_{ZY}$ represents the effect of the IV (Z) on the outcome (Y), $\hat{\beta}_{ZX}$ represents the effect of the IV (Z) on the exposure (X), and the variance of $\hat{\beta}_{IV}$ can be estimated by the delta method (Thomas et al., 2007). In the presence of multiple IVs (e.g., multiple instrumental SNPs), several methods have been proposed to estimate β_{XY} under different assumptions. In this study, we used three different methods, namely the Inverse Variance Weighted (IVW) method (Burgess et al., 2013), the MR pleiotropy residual sum and outlier (MR-PRESSO) method (Verbanck et al., 2018), and the weighted median method (Bowden et al., 2016a).

Briefly, the IVW estimate is the IVW average of $\hat{\beta}_{IV}$, assuming all SNPs are valid IVs or the overall bias is zero (balanced pleiotropy) (Bowden et al., 2016b). We performed multiplicative random effects IVW to account for potential heterogeneity, which is measured by the Cochran's Q statistic (Hemani et al., 2018a). The IVW method is equivalent to fitting a weighted linear regression with no intercept of $\hat{\beta}_{ZY}$ on $\hat{\beta}_{ZX}$ where the weight is the inverse variance of $\hat{\beta}_{ZY}$ and the estimated regression slope is the estimated causal effect of the exposure on the outcome (β_{XY}).

The MR-PRESSO method is designed to correct for horizontal pleiotropy, in which the IV acts on the outcome via a pathway other than through the exposure. MR-PRESSO is based on the IVW regression framework and detects IVs of horizontal pleiotropy as outliers in the regression. In particular, MR-PRESSO implements a global test based on the leave-one-out approach to test for the existence of horizontal pleiotropy and

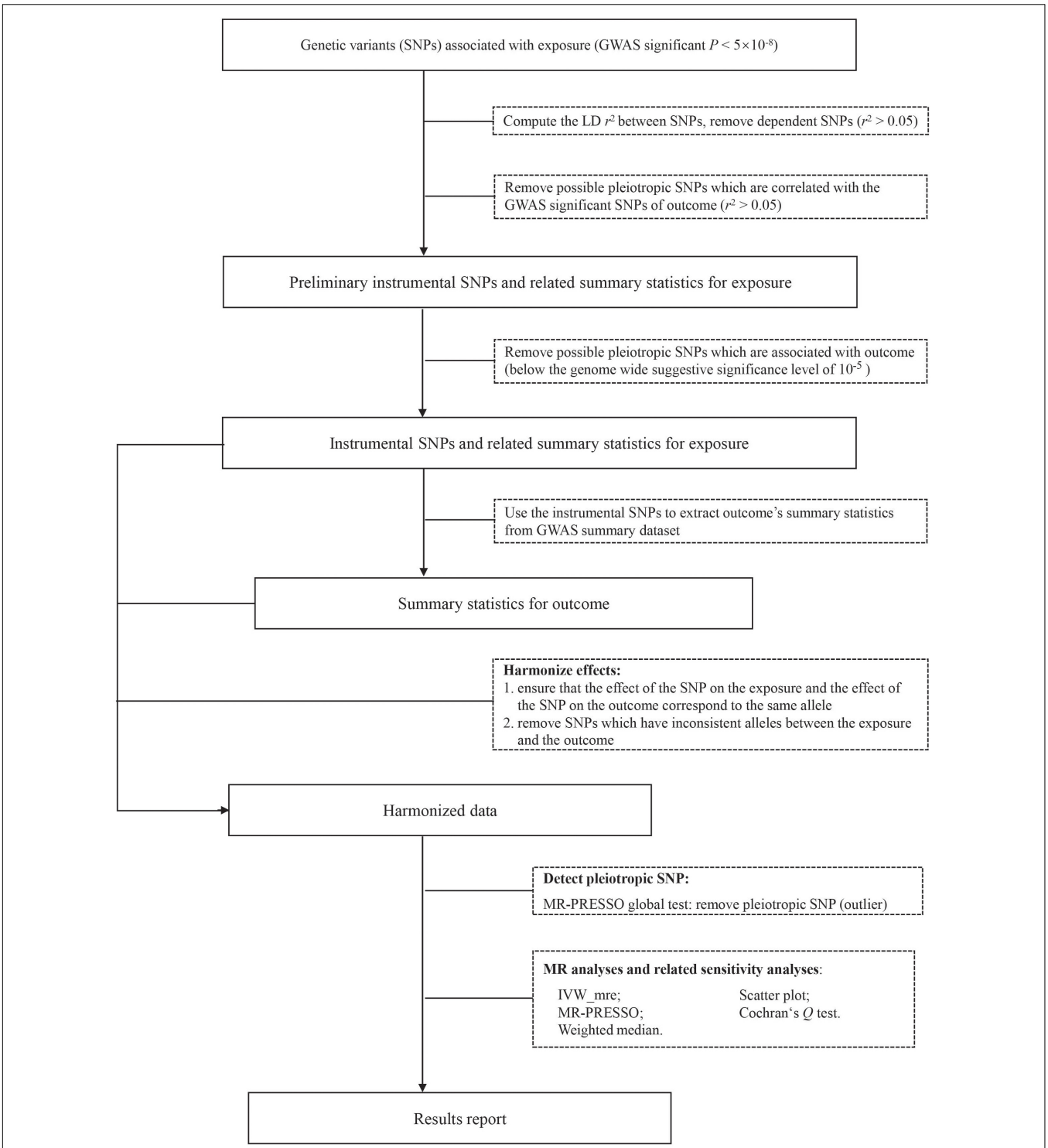


FIGURE 1 | The analysis flowchart of this study. IVW_mre, inverse variance weighted multiplicative random effects; MR-PRESSO, Mendelian randomization pleiotropy residual sum and outlier.

an outlier test to detect specific SNPs with horizontal pleiotropy. MR-PRESSO provides the final IVW estimate after removing outlier IVs (Verbanck et al., 2018).

Finally, the weighted median method uses the inverse variance of $\hat{\beta}_{IV}$ as weight to construct the empirical distribution of $\hat{\beta}_{IV}$, and derives the final estimate by taking the median (Bowden et al.,

2016a). The confidence interval of the weighted median estimate is obtained by a parametric bootstrap method. This method can provide a consistent estimate as long as at least 50% of the weight comes from the valid IVs.

We displayed the scatter plot of genetic effect on the outcome ($\hat{\beta}_{ZY}$) vs. genetic effect on the exposure ($\hat{\beta}_{ZX}$) for each IV to facilitate the identification of possible heterogeneity and the illustration of causal effects. We used mRnd[1] to calculate *post hoc* statistical power. With a Bonferroni-corrected significance level of 2.8×10^{-3} ($\alpha = 0.05/18$, correcting 18 exposure-outcome paired tests), we estimated the required OR of exposure on outcome (in the unit of per standard deviation increment in exposure) to achieve 80% statistical power given the summary statistics (Brion et al., 2013). A causal effect of an exposure on the outcome is concluded if the effect estimates agree in direction and magnitude among MR methods, pass the Bonferroni-corrected significance threshold of 2.8×10^{-3} in the IVW method, and show no evidence of heterogeneity in the Cochran's *Q*-test and MR-PRESSO global test. Findings with *P*-values between 0.05 and 2.8×10^{-3} were deemed suggestive evidence of causality. Analyses were performed with TwoSampleMR and MR-PRESSO packages in R version 3.5.3 (Hemani et al., 2018b; Verbanck et al., 2018; Team RC, 2019).

RESULTS

Psychiatric Traits Predicting Mouth Ulcers

After the IV selection process, we displayed the genetic associations with mouth ulcers over genetic associations with psychiatric traits for the valid IVs (**Figure 2**). By the MR-PRESSO outlier test, we detected two outlier SNPs (solid red dots in **Figures 2C,I**): one each for ASD and SCZ. After removing these two SNPs, all three MR methods agreed well in fitting the linear relationship between the genetic effect sizes on mouth ulcers and each of the psychiatric traits (colored solid lines in **Figure 2**). Estimates of the causal effects of 10 psychiatric traits on mouth ulcers were presented in **Figure 3**. We found that ASD, insomnia, and MDD have significant risk effects and subjective wellbeing has significant protective effect on mouth ulcers. The corresponding effect sizes from the IVW method were OR = 1.160 (95% CI: 1.066–1.261, $P = 5.39 \times 10^{-4}$), 1.092 (1.062–1.122, $P = 3.37 \times 10^{-10}$), 1.234 (1.134–1.342, $P = 1.03 \times 10^{-6}$), and 0.703 (0.571–0.865, $P = 8.97 \times 10^{-4}$) for ASD, insomnia, MDD, and subjective wellbeing, respectively. There was suggestive evidence for risk effect of mood instability on mouth ulcers (IVW, OR = 1.662, 95% CI: 1.059–2.609, $P = 0.027$). All the *F* statistics were greater than 32, indicating robust causal estimates against the weak instrument bias (**Figure 3**). We confirmed that these estimated effect sizes were close to or above the threshold to achieve 80% statistical power given the available summary statistics (**Supplementary Table S20**). Importantly, the MR-PRESSO global test and Cochran *Q*-test suggested no heterogeneity or pleiotropic effect (**Supplementary Table S22**). We found no evidence of causal effects on mouth ulcers from all three MR methods for the remaining five psychiatric traits (anxiety disorder, ADHD, BIP, neuroticism, and SCZ, **Figures 2**, **3** and **Supplementary Tables S20**, **S22**).

Mouth Ulcers Predicting Psychiatric Traits

Because the summary statistics of anxiety disorders and mood instability were only available for significantly associated SNPs, we could not perform MR analyses of mouth ulcers on these two traits. For the remaining eight psychiatric traits, we displayed their genetic effect sizes vs. the genetic effect sizes on mouth ulcers for the valid IVs (**Figure 4**). Two outlier SNPs (solid red dots in **Figures 4D,G**) were detected by the MR-PRESSO outlier test and excluded subsequently. Although the instrumental SNPs could explain a substantial amount of phenotypic variance ($\geq$ 0.9% for all eight traits, **Supplementary Table S21**) and the *F* statistics indicated strong instrumental effects (all $F > 52$), we found no significant evidence of causal effects of mouth ulcers on these psychiatric traits (**Figures 4**, **5**). The only suggestive evidence was given by the MR-PRESSO method for mouth ulcers on ASD (*OR* = 1.065, 95% CI: 1.002–1.132, $P = 0.046$). The effect estimates for mouth ulcers on ASD were 1.065 (0.994–1.141, $P = 0.071$) and 1.092 (0.986–1.209, $P = 0.094$) by the IVW method and the weighted median method, respectively (**Figure 5**). Furthermore, these effect estimates were below the threshold to achieve 80% statistical power, suggesting a high potential for false discoveries (**Supplementary Table S21**). No heterogeneity or directional pleiotropy was indicated by the MR-PRESSO global test and Cochran *Q*-test (**Supplementary Table S23**).

DISCUSSION

Psychiatric disorders have been suggested to associate with mouth ulcers by observational studies. We performed two-sample bi-directional MR analyses to explore the causality between 10 psychiatric traits (anxiety disorder, ADHD, ASD, BIP, insomnia, MDD, mood instability, neuroticism, SCZ, and subjective wellbeing) and mouth ulcers based on summary statistics of the largest available GWAS to date. Our analyses suggested that ASD, insomnia, MDD, and mood instability have risk effects and subjective wellbeing has a protective effect on mouth ulcers, whereas mouth ulcers have no significant effect on any of these psychiatric traits. Our analyses were well-powered and did not suffer from weak instrumental bias according to the *F* statistics. The MR-PRESSO global test, Cochran's *Q*-test, and scatter plots indicated no directional pleiotropy or heterogeneity.

It has been pointed out that stress, depression, and anxiety are associated with mouth ulcers by a cross-sectional study (Alshahrani and Baccaglini, 2014). A recent study, based on linkage disequilibrium score regression analysis, also found a significant genetic correlation (correlation coefficient = 0.24, $P = 5.73 \times 10^{-7}$) between depression and mouth ulcers in Europeans (Dudding et al., 2019). Using bi-directional MR analyses, we tested these observational results and confirmed that

[1]http://cnsgenomics.com/shiny/mRnd/

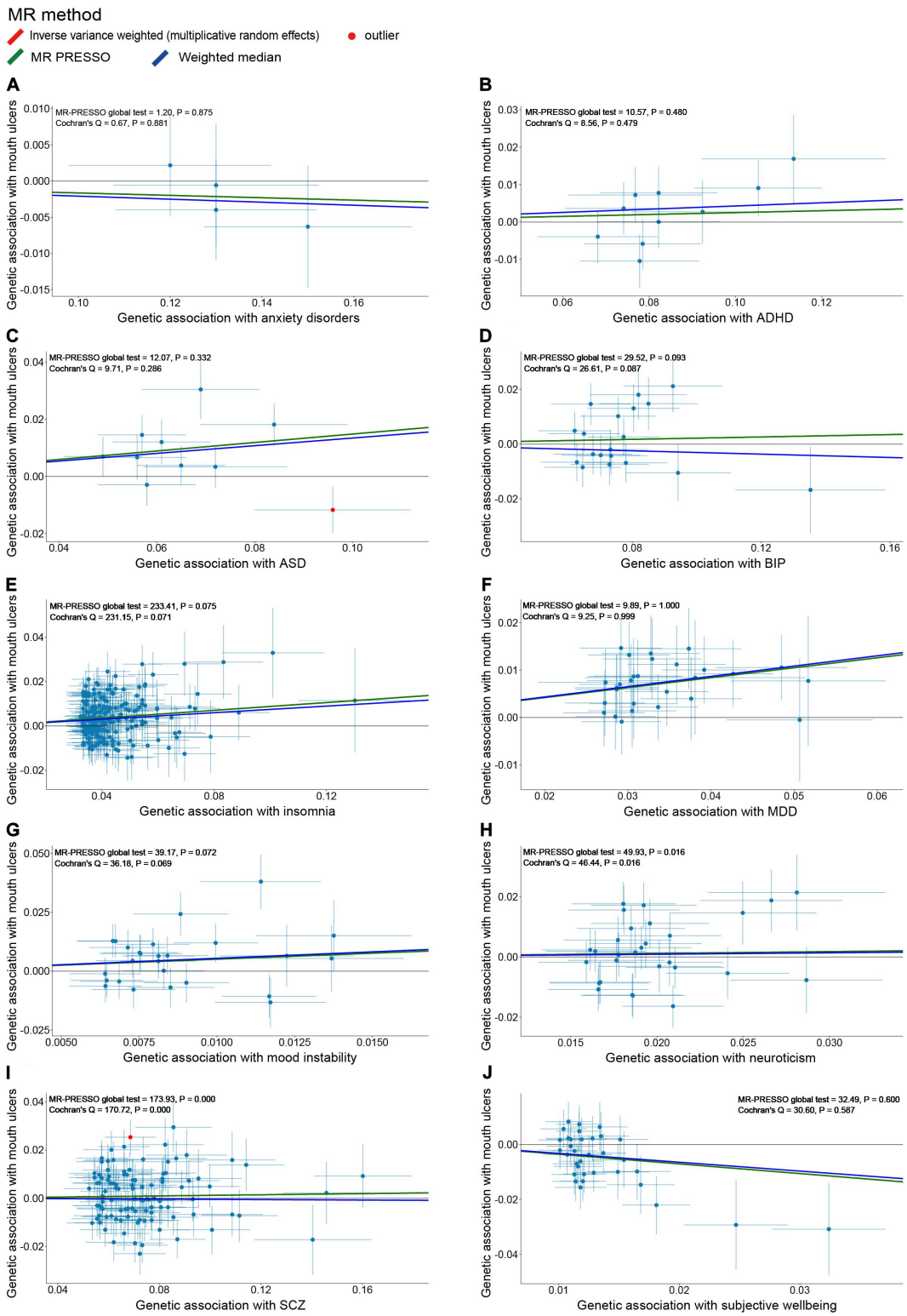
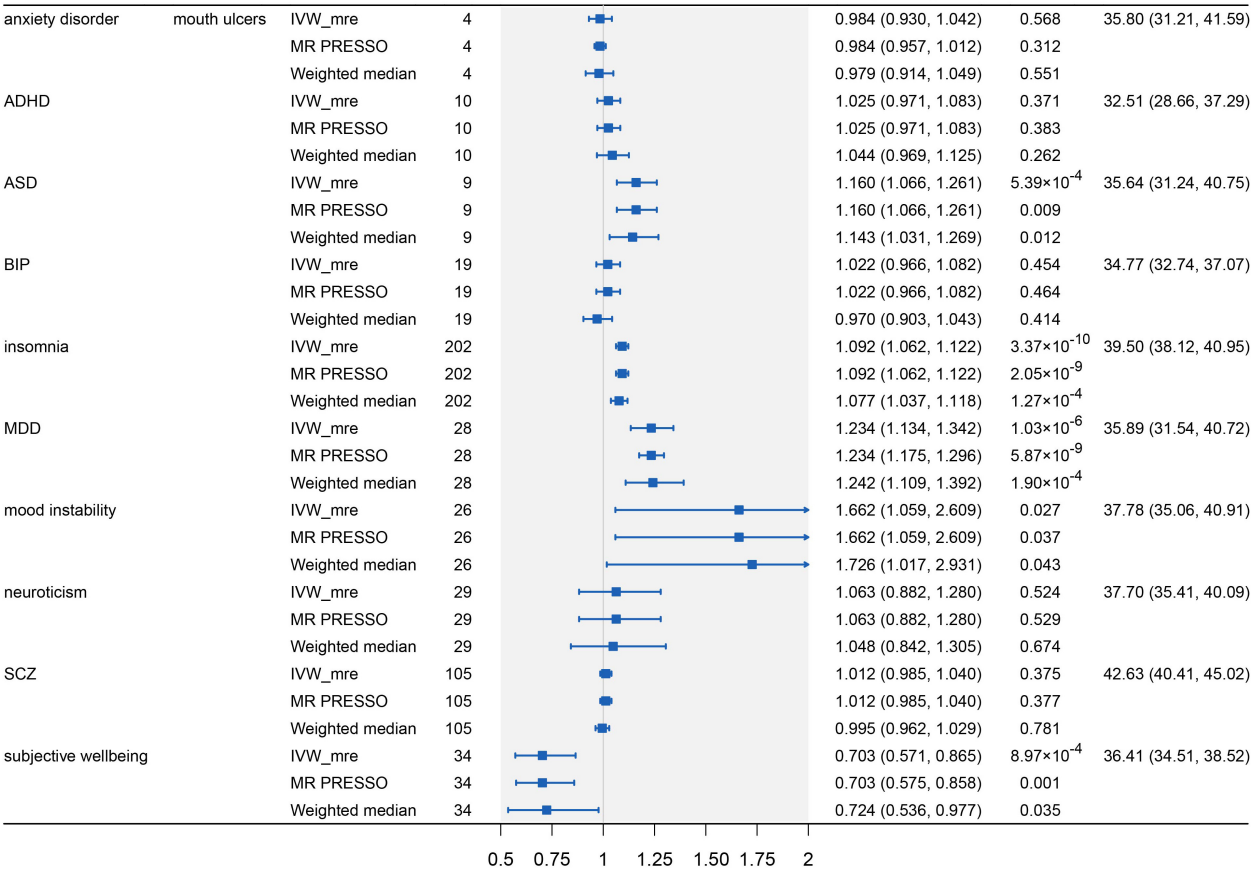


FIGURE 2 | Scatter plots of genetic associations with mouth ulcers (outcome) vs. genetic associations with 10 psychiatric traits (exposure) for all the valid IVs. **(A)** anxiety disorders; **(B)** ADHD; **(C)** ASD; **(D)** BIP; **(E)** insomnia; **(F)** MDD; **(G)** mood instability; **(H)** neuroticism; **(I)** SCZ; **(J)** subjective wellbeing. Each dot corresponds to one genetic variant, with corresponding standard error bars of its association with psychiatric trait (horizontal) and mouth ulcers (vertical); solid red dot represents the pleiotropic SNP (outlier) identified by MR-PRESSO global test; the solid lines illustrate estimations of the causal effect after excluding outlier SNP, colored by different colors with different MR methods. The horizontal gray solid line indicates no effect. In this study, the causal effect estimations from the IVW_mre and MR-PRESSO are consistent, such that the red line (IVW_mre) is covered by the green line (MR-PRESSO), and no red line could be observed.



Exposure	Outcome	Method	N SNPs	*OR* (95% *CI*)	*P*	*F* statistics
anxiety disorder	mouth ulcers	IVW_mre	4	0.984 (0.930, 1.042)	0.568	35.80 (31.21, 41.59)
		MR PRESSO	4	0.984 (0.957, 1.012)	0.312	
		Weighted median	4	0.979 (0.914, 1.049)	0.551	
ADHD		IVW_mre	10	1.025 (0.971, 1.083)	0.371	32.51 (28.66, 37.29)
		MR PRESSO	10	1.025 (0.971, 1.083)	0.383	
		Weighted median	10	1.044 (0.969, 1.125)	0.262	
ASD		IVW_mre	9	1.160 (1.066, 1.261)	5.39×10^{-4}	35.64 (31.24, 40.75)
		MR PRESSO	9	1.160 (1.066, 1.261)	0.009	
		Weighted median	9	1.143 (1.031, 1.269)	0.012	
BIP		IVW_mre	19	1.022 (0.966, 1.082)	0.454	34.77 (32.74, 37.07)
		MR PRESSO	19	1.022 (0.966, 1.082)	0.464	
		Weighted median	19	0.970 (0.903, 1.043)	0.414	
insomnia		IVW_mre	202	1.092 (1.062, 1.122)	3.37×10^{-10}	39.50 (38.12, 40.95)
		MR PRESSO	202	1.092 (1.062, 1.122)	2.05×10^{-9}	
		Weighted median	202	1.077 (1.037, 1.118)	1.27×10^{-4}	
MDD		IVW_mre	28	1.234 (1.134, 1.342)	1.03×10^{-6}	35.89 (31.54, 40.72)
		MR PRESSO	28	1.234 (1.175, 1.296)	5.87×10^{-9}	
		Weighted median	28	1.242 (1.109, 1.392)	1.90×10^{-4}	
mood instability		IVW_mre	26	1.662 (1.059, 2.609)	0.027	37.78 (35.06, 40.91)
		MR PRESSO	26	1.662 (1.059, 2.609)	0.037	
		Weighted median	26	1.726 (1.017, 2.931)	0.043	
neuroticism		IVW_mre	29	1.063 (0.882, 1.280)	0.524	37.70 (35.41, 40.09)
		MR PRESSO	29	1.063 (0.882, 1.280)	0.529	
		Weighted median	29	1.048 (0.842, 1.305)	0.674	
SCZ		IVW_mre	105	1.012 (0.985, 1.040)	0.375	42.63 (40.41, 45.02)
		MR PRESSO	105	1.012 (0.985, 1.040)	0.377	
		Weighted median	105	0.995 (0.962, 1.029)	0.781	
subjective wellbeing		IVW_mre	34	0.703 (0.571, 0.865)	8.97×10^{-4}	36.41 (34.51, 38.52)
		MR PRESSO	34	0.703 (0.575, 0.858)	0.001	
		Weighted median	34	0.724 (0.536, 0.977)	0.035	

FIGURE 3 | Two-sample Mendelian randomization analyses showing the effect estimates of 10 psychiatric traits on mouth ulcers. ADHD, attention deficit/hyperactivity disorder; ASD, autism spectrum disorder; BIP, bipolar disorder; MDD, major depressive disorder; SCZ, schizophrenia; IVW_mre, inverse variance weighted with multiplicative random effects; MR-PRESSO, Mendelian randomization pleiotropy residual sum and outlier; N SNP, number of the instrumental SNPs used to conduct MR analyses; Effect estimates express the change in odds ratio (*OR*) per standard deviation (SD) increment in psychiatric traits, error bars indicate 95% confidence intervals.

MDD has a causal effect on mouth ulcers. However, inconsistent findings were observed for anxiety; our analyses did not support a causal relationship between anxiety and mouth ulcers. The relatively small sample size (n = 83,566) and only 0.2% of phenotype variation explained by the four IVs of anxiety may explain the null finding. While the precise mechanism linking depression to mouth ulcers is not well understood, the immune system or inflammatory response is suggested to be involved (Al-Omiri et al., 2012; Huling et al., 2012; Alshahrani and Baccaglini, 2014). Depression can increase the number of leukocytes, which exhibit increased motility and enhanced adhesion to endothelial cells and thus induce endothelial dysfunction and mouth ulcers ultimately (Gavic et al., 2014; Demir et al., 2015; Qin et al., 2018). Besides, a serotonin transporter gene polymorphism (*5-HTTLPR*), which is commonly found in depressed patients, is also significantly enriched in patients with mouth ulcers (Victoria et al., 2005). Further functional experiments are required to clarify the mechanistic link between MDD and mouth ulcers.

Many observational studies have reported positive associations between stress and mouth ulcers (Al-Omiri et al., 2012; Huling et al., 2012; Ma et al., 2015; Ge, 2018). For example, ulceration is exacerbated during examination periods and lessened during periods of vacation for students (Scully, 2013). Meanwhile, stress is well known to correlate with mood instability and subjective wellbeing (Schneiderman et al., 2005; Atanes et al., 2015; Berrios et al., 2016; Gillett and Crisp, 2017; Faurholt-Jepsen et al., 2019). Our study suggested that mood instability and subjective wellbeing are causally associated with mouth ulcers using several MR methods. Stress is thought to affect multiple immune system components including the distribution and proliferation of lymphocytes and natural killer cells and production of cytokines and antibodies (Huling et al., 2012). Stressful situations can cause a transitory increase of salivary cortisol and stimulate immunoregulatory activity by increasing the number of leukocytes in inflammatory sites, which are often observed during the pathogenesis of mouth ulcers (Albanidou-Farmaki et al., 2008; Gallo Cde et al., 2009; Al-Omiri et al., 2015). However, the exact mechanism about how stress-related mood instability and subjective wellbeing trigger mouth ulcers remains to be elucidated.

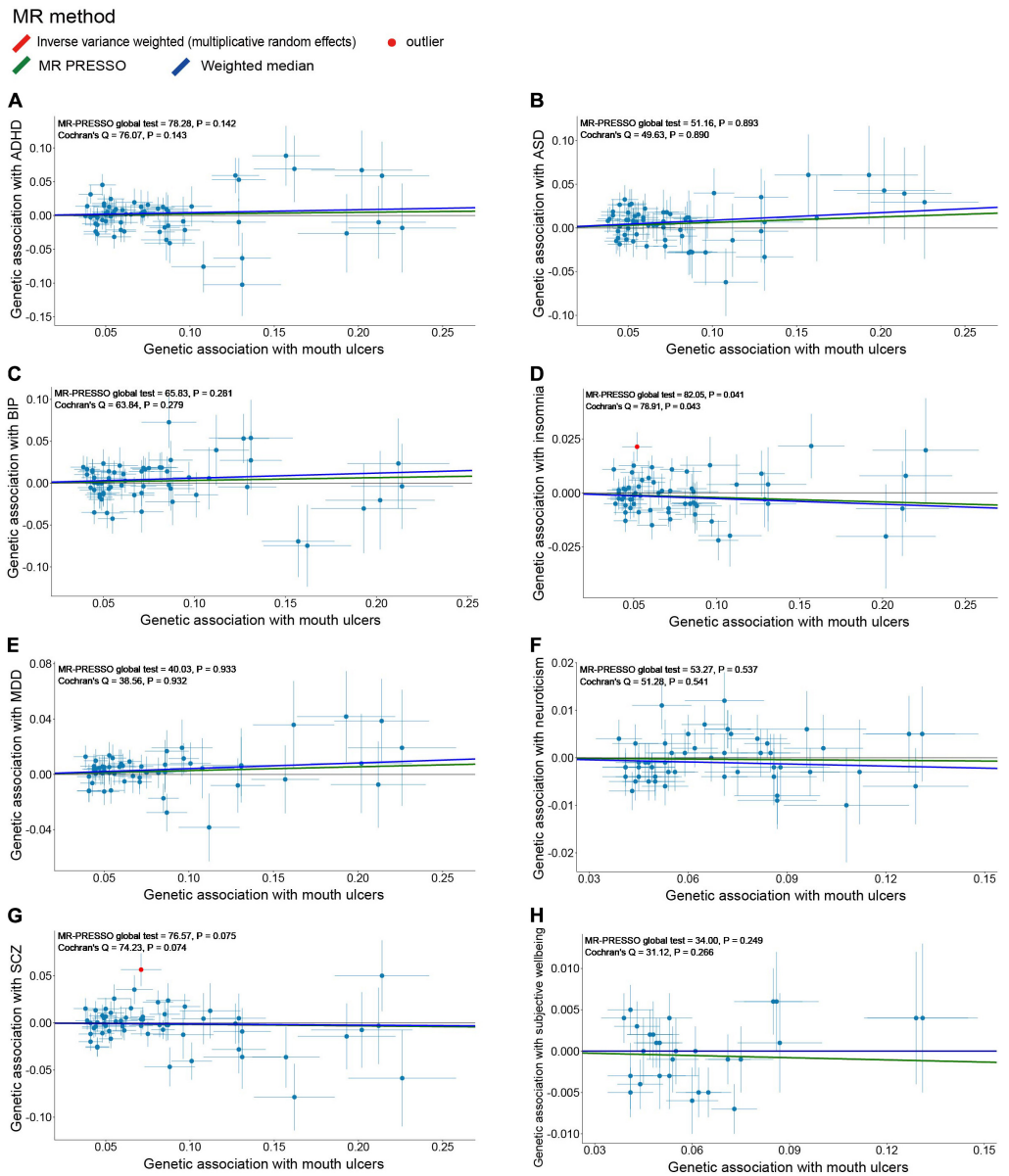


FIGURE 4 | Scatter plots of genetic associations with eight psychiatric traits (outcome) vs. genetic associations with mouth ulcers (exposure) for all the valid IVs. **(A)** ADHD; **(B)** ASD; **(C)** BIP; **(D)** insomnia; **(E)** MDD; **(F)** neuroticism; **(G)** SCZ; **(H)** subjective wellbeing. Each dot corresponds to one genetic variant, with corresponding standard error bars of its association with mouth ulcers (horizontal) and psychiatric trait (vertical); solid red dot represents the pleiotropic SNP (outlier) identified by MR-PRESSO global test; the solid lines illustrate estimations of the causal effect after excluding outlier SNP, colored by different colors with different MR methods. The horizontal gray solid line indicates no effect. In this study, the causal effect estimations from the IVW_mre and MR-PRESSO are consistent, such that the red line (IVW_mre) is covered by the green line (MR-PRESSO), and no red line could be observed.

Consistent with previous observational studies (Ma et al., 2015; Du et al., 2018), we also found that insomnia has a causal effect to increase the risk of mouth ulcers. Insomnia will lead to late bedtime, which can disturb the secretion of hormones, such as growth hormone, cortisol, and adrenocorticotropic hormone (Ma et al., 2015). The reduced secretion of growth hormone can promote the occurrence of mouth ulcers and delay healing (Brandenberger, 2004; Dioufa et al., 2010; Lee et al., 2010; Smaniotto et al., 2011). Insufficient secretion of cortisol and adrenocorticotropic hormone may also increase inflammation and allergic reactions and facilitate the occurrence of mouth ulcers (MacGregor et al., 1969; Bierwolf et al., 2000; Sakamoto et al., 2013; Gavic et al., 2014). Hormonal factors are capable of altering the thickness of the mucosa, which is an important factor in mouth ulcers (Neville et al., 2008; Scully, 2013).

It is worth noting that MR uses genetic variants as the IVs such that its causal effect estimate represents the average effect of lifetime exposure on the outcome (Holmes et al., 2017). Most

Exposure	Outcome	Method	N SNPs	OR (95% CI)	P	F statistics
mouth ulcers	ADHD	IVW_mre	65	1.024 (0.953, 1.101)	0.512	54.23 (48.55, 60.40)
		MR PRESSO	65	1.024 (0.953, 1.101)	0.514	
		Weighted median	65	1.044 (0.946, 1.151)	0.384	
	ASD	IVW_mre	64	1.065 (0.994, 1.141)	0.071	54.67 (48.75, 61.24)
		MR PRESSO	64	1.065 (1.002, 1.132)	0.046	
		Weighted median	64	1.092 (0.986, 1.209)	0.094	
	BIP	IVW_mre	59	1.034 (0.961, 1.111)	0.377	54.36 (48.71, 60.39)
		MR PRESSO	59	1.034 (0.961, 1.111)	0.381	
		Weighted median	59	1.062 (0.957, 1.178)	0.260	
	insomnia	IVW_mre	60	0.979 (0.951, 1.008)	0.171	54.58 (48.42, 61.31)
		MR PRESSO	60	0.979 (0.951, 1.008)	0.177	
		Weighted median	60	0.974 (0.935, 1.015)	0.205	
	MDD	IVW_mre	54	1.027 (0.984, 1.073)	0.219	52.70 (47.27, 58.57)
		MR PRESSO	54	1.027 (0.990, 1.066)	0.155	
		Weighted median	54	1.042 (0.984, 1.103)	0.160	
	neuroticism	IVW_mre	54	0.995 (0.978, 1.013)	0.626	56.39 (49.40, 63.96)
		MR PRESSO	54	0.995 (0.978, 1.013)	0.622	
		Weighted median	54	0.985 (0.958, 1.013)	0.284	
	SCZ	IVW_mre	59	0.984 (0.932, 1.040)	0.584	55.79 (49.68, 62.44)
		MR PRESSO	59	0.984 (0.932, 1.040)	0.586	
		Weighted median	59	0.988 (0.919, 1.062)	0.756	
	subjective wellbeing	IVW_mre	28	0.991 (0.968, 1.015)	0.475	54.05 (46.64, 62.41)
		MR PRESSO	28	0.991 (0.968, 1.015)	0.481	
		Weighted median	28	1.000 (0.967, 1.034)	1.000	

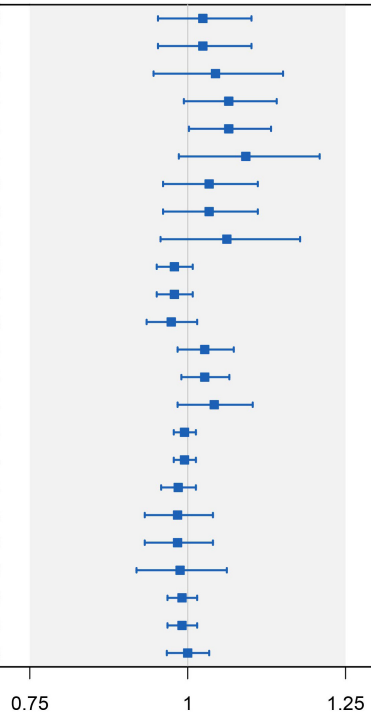


FIGURE 5 | Two-sample Mendelian randomization analysis showing the effect of mouth ulcers on eight psychiatric traits. ADHD, attention deficit/hyperactivity disorder; ASD, autism spectrum disorder; BIP, bipolar disorder; MDD, major depressive disorder; SCZ, schizophrenia; IVW_mre, inverse variance weighted with multiplicative random effects; MR-PRESSO, Mendelian randomization pleiotropy residual sum and outlier; N SNP, number of the instrumental SNPs used to conduct MR analyses; Effect estimates express the change in odds ratio (*OR*) per standard deviation (*SD*) increment in mouth ulcers, error bars indicate 95% confidence intervals.

of the psychiatric traits we studied were clinically diagnosed long-term disorders (**Supplementary Table S1**), but their clinical symptoms might be time-dependent. For example, patients with anxiety disorders might present different levels of anxiety across time periods. Hence, the risk of developing mouth ulcers is also likely to be time-dependent if the anxiety symptom is causal. More caution needs to be taken when interpreting causal effect sizes derived from MR analysis in clinical practice.

Our bi-directional MR analyses had important strengths. Firstly, using randomly allocated genetic variants as IVs, we could reduce the potential impacts of conventional confounders and reverse causality, which are common in observational studies. Secondly, the SNP-exposure and SNP-outcome estimates we used were derived from studies of the largest sample sizes to date (ranging from 46,350 to 1,331,010 individuals), allowing credible causal inference between psychiatric traits and mouth ulcers in the European population. Thirdly, by utilizing a bi-directional MR design, we evaluated the causal relationship between two traits simultaneously and could assess the causal direction more confidently. Finally, our conclusions were drawn based on comprehensive analyses involving 10 psychiatric traits, three credible MR methods, and several heterogeneity tests to prevent possible pleiotropic bias.

There were also some limitations in our study. First, our analysis did not distinguish different types of mouth ulcers, because mouth ulcers in UKB were inferred from the questionnaire rather than clinical examination. Given that most of the significant variants from UKB have been validated in independent samples, including three specific to RAS, the major type of mouth ulcers (Bilodeau and Lalla, 2019), while other types of ulcers, such as traumatic mouth ulcers, are less likely to be genetic (Dudding et al., 2019), our findings are expected to largely reflect the causality between psychiatric traits and RAS. Second, the sample overlapping between GWAS of mood instability and mouth ulcers was as large as 78.9%, which violated the assumption of two-sample MR. Nevertheless, the *F* statistic was large enough (F = 37.78, 95% CI: 35.06–40.91), suggesting that the sample overlapping would not materially affect the causal inference (Burgess et al., 2016). Third, consistent with findings in other MR studies involving of psychiatric traits, the effect sizes of genetic variants on psychiatric traits were estimated with large standard errors (**Figure 2**), indicating difficulty to accurately measure these traits (Wootton et al., 2018; Vermeulen et al., 2019). For this reason, we did not use the MR-Egger method, because it assumes that the associations between IVs and the exposure are precisely estimated or have a wide spread (Bowden et al., 2016b; Burgess and Thompson, 2017). Fourth, cautions are needed when generalizing our findings, which were derived from data of European population, to non-European populations, because different environmental factors might have

substantial impacts on psychiatric traits and mouth ulcers. Lastly, we did not consider sex-specific effects, which might differ for psychiatric traits and mouth ulcers due to differences in hormone levels. Because GWAS summary statistics we collected were not stratified by sex, we could not perform sex-specific analyses to validate different sex-specific causal effects of psychiatric traits on mouth ulcers observed in epidemiological studies (Huling et al., 2012; Slebioda and Dorocka-Bobkowska, 2019).

CONCLUSION

In conclusion, utilizing large-scale GWAS summary statistics and two-sample bi-directional MR analyses, our study provides causal evidence on the risk role of ASD, insomnia, MDD, and mood instability, and the protective role of subjective wellbeing on mouth ulcers in the European population. Future work is needed to understand the biological pathways from psychiatric traits to mouth ulcers.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

XH and CW conceived and supervised the study. KW and LD collected and analyzed the data. KW, XH, and CW wrote the manuscript with inputs from CY. All authors have reviewed and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (NSFC 81973148 and 82003561) and Hong Kong Research Grant Council (16307818, 16301419, and 16308120).

ACKNOWLEDGMENTS

We thank participants and investigators who contributed to the GWASs included in our analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.608630/full#supplementary-material

REFERENCES

Albanidou-Farmaki, E. P. A., Epivatianos, A., Farmakis, K., Karamouzis, M., and Antoniades, D. (2008). Increased anxiety level and high salivary and serum cortisol concentrations in patients with recurrent aphthous stomatitis. *Tohoku J. Exp. Med.* 214, 291–296.

Almoznino, G., Zini, A., Mizrahi, Y., and Aframian, D. J. (2014). Elevated serum IgE in recurrent aphthous stomatitis and associations with disease characteristics. *Oral Dis.* 20, 386–394. doi: 10.1111/odi.12131

Al-Omiri, M. K., Karasneh, J., Alhijawi, M. M., Zwiri, A. M., Scully, C., and Lynch, E. (2015). Recurrent aphthous stomatitis (RAS): a preliminary within-subject study of quality of life, oral health impacts and personality profiles. *J. Oral Pathol. Med.* 44, 278–283. doi: 10.1111/jop.12232

Al-Omiri, M. K., Karasneh, J., and Lynch, E. (2012). Psychological profiles in patients with recurrent aphthous ulcers. *Int. J. Oral Maxillofaci. Surg.* 41, 384–388. doi: 10.1016/j.ijom.2011.12.024

Alshahrani, S., and Baccaglini, L. (2014). Psychological screening test results for stress, depression, and anxiety are variably associated with clinical severity of recurrent aphthous stomatitis and oral lichen planus. *J. Evid. Based Dent. Pract.* 14, 206–208. doi: 10.1016/j.jebdp.2014.10.004

Atanes, A. C., Andreoni, S., Hirayama, M. S., Montero-Marin, J., Barros, V. V., Ronzani, T. M., et al. (2015). Mindfulness, perceived stress, and subjective well-being: a correlational study in primary care health professionals. *BMC Complement. Alternat. Med.* 15:303. doi: 10.1186/s12906-015-0823-0

Au Yeung, S. L., Lam, H., and Schooling, C. M. (2017). Vascular Endothelial Growth Factor and Ischemic Heart Disease Risk: A Mendelian Randomization Study. *J. Am. Heart Assoc.* 6:5619. doi: 10.1161/JAHA.117.005619

Berrios, M. P., Extremera, N., and Nieto-Flores, M. P. (2016). Exploring the socio-emotional factors associated with subjective well-being in the unemployed. *PeerJ.* 4:e2506. doi: 10.7717/peerj.2506

Bierwolf, C., Kern, W., Mölle, M., Born, J., and Fehm, H. L. (2000). Rhythms of pituitary-adrenal activity during sleep in patients with Cushing's disease. *Exp. Clinic. Endocrinol. Diabetes* 108, 470–479. doi: 10.1055/s-2000-8143

Bilodeau, E. A., and Lalla, R. V. (2019). Recurrent oral ulceration: Etiology, classification, management, and diagnostic algorithm. *Periodontol* 80, 49–60. doi: 10.1111/prd.12262

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525. doi: 10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016a). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Gen. Epidemiol.* 40, 304–314. doi: 10.1002/gepi.21965

Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016b). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int. J. Epidemiol.* 45, 1961–1974. doi: 10.1093/ije/dyw220

Brandenberger, G. W. L. (2004). The 24-h growth hormone rhythm in men Sleep and circadian influences questioned. *J. Sleep Res.* 13, 251–255.

Brion, M. J., Shakhbazov, K., and Visscher, P. M. (2013). Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* 42, 1497–1501. doi: 10.1093/ije/dyt179

Burgess, S., and Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* 32, 377–389. doi: 10.1007/s10654-017-0255-x

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Gen. Epidemiol.* 37, 658–665. doi: 10.1002/gepi.21758

Burgess, S., Davies, N. M., and Thompson, S. G. (2016). Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol.* 40, 597–608. doi: 10.1002/gepi.21998

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. doi: 10.1186/s13742-015-0047-8

Consortium, G. P. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Gen.* 23, R89–R98. doi: 10.1093/hmg/ddu328

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ.* 362:k601. doi: 10.1136/bmj.k601

Demir, S., Atli, A., Bulut, M., İbiloğlu, A. O., Güneş, M., Kaya, M. C., et al. (2015). Neutrophil-lymphocyte ratio in patients with major depressive disorder undergoing no pharmacological therapy. *Neuropsychiatr. Dis. Treat.* 11, 2253–2258. doi: 10.2147/NDT.S89470

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Gen.* 51, 63–75. doi: 10.1038/s41588-018-0269-7

Dioufa, N., Schally, A. V., Chatzistamou, I., Moustou, E., Block, N. L., Owens, G. K., et al. (2010). Acceleration of wound healing by growth hormone-releasing hormone and its agonists. *Proc. Natl. Acad. Sci. U S A.* 107, 18611–18615. doi: 10.1073/pnas.1013942107

Du, Q., Ni, S., Fu, Y., and Liu, S. (2018). Analysis of Dietary Related Factors of Recurrent Aphthous Stomatitis among College Students. *Evid. Based Complement. Alternat. Med.* 2018:2907812. doi: 10.1155/2018/2907812

Dudding, T., Haworth, S., Lind, P. A., Sathirapongsasuti, J. F., Tung, J. Y., Mitchell, R., et al. (2019). Genome wide analysis for mouth ulcers identifies associations at immune regulatory loci. *Nat. Commun.* 10:1052. doi: 10.1038/s41467-019-08923-6

Faurholt-Jepsen, M., Frost, M., Busk, J., Christensen, E. M., Bardram, J. E., Vinberg, M., et al. (2019). Is smartphone-based mood instability associated with stress, quality of life, and functioning in bipolar disorder? *Bipol. Disord.* 21, 611–620. doi: 10.1111/bdi.12796

Gallo Cde, B., Mimura, M. A., and Sugaya, N. N. (2009). Psychological stress and recurrent aphthous stomatitis. *Clinics* 64, 645–648. doi: 10.1590/S1807-59322009000700007

Gavic, L., Cigic, L., Biocina Lukenda, D., Gruden, V., and Gruden Pokupec, J. S. (2014). The role of anxiety, depression, and psychological stress on the clinical status of recurrent aphthous stomatitis and oral lichen planus. *J. Oral Pathol. Med.* 43, 410–417. doi: 10.1111/jop.12148

Ge, L. (2018). Healthy lifestyle habits benefit remission of recurrent aphthous stomatitis and RAS type ulceration. *Br. Dent. J.* 224, 70–71. doi: 10.1038/sj.bdj.2018.38

Gillett, J. E., and Crisp, D. A. (2017). Examining coping style and the relationship between stress and subjective well-being in Australia's 'sandwich generation'. *Aus. J. Ageing.* 36, 222–227. doi: 10.1111/ajag.12439

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nat. Gen.* 51, 431–444. doi: 10.1038/s41588-019-0344-8

Hartwig, F. P., Borges, M. C., Horta, B. L., Bowden, J., and Davey Smith, G. (2017). Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study. *JAMA Psychiatry* 74, 1226–1233. doi: 10.1001/jamapsychiatry.2017.3191

Hartwig, F. P., Davies, N. M., Hemani, G., and Davey Smith, G. (2016). Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int. J. Epidemiol.* 45, 1717–1726. doi: 10.1093/ije/dyx028

Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C., and Davey Smith, G. (2016). Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am. J. Clinic. Nutr.* 103, 965–978. doi: 10.3945/ajcn.115.118216

Hemani, G., Bowden, J., and Davey Smith, G. (2018a). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* 27, R195–R208. doi: 10.1093/hmg/ddy163

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018b). The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7:e34408. doi: 10.7554/eLife.34408

Holmes, M. V., Ala-Korpela, M., and Smith, G. D. (2017). Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* 14, 577–590. doi: 10.1038/nrcardio.2017.78

Huling, L. B., Baccaglini, L., Choquette, L., Feinn, R. S., and Lalla, R. V. (2012). Effect of stressful life events on the onset and duration of recurrent aphthous stomatitis. *J. Oral Pathol. Med.* 41, 149–152. doi: 10.1111/j.1600-0714.2011.01102.x

Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., et al. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Gen.* 51, 394–403. doi: 10.1038/s41588-018-0333-3

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statist. Med.* 27, 1133–1163. doi: 10.1002/sim.3034

Lee, S. W., Kim, S. H., Kim, J. Y., and Lee, Y. (2010). The effect of growth hormone on fibroblast proliferation and keratinocyte migration. *J. Plastic Reconstruct. Aesthetic Surg.* 63, e364–e369. doi: 10.1016/j.bjps.2009.10.027

Ma, R., Chen, H., Zhou, T., Chen, X., Wang, C., Chen, Y., et al. (2015). Effect of bedtime on recurrent aphthous stomatitis in college students. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* 119, 196–201. doi: 10.1016/j.oooo.2014.10.014

MacGregor, R. R., Sheagren, J. N., Lipsett, M. B., and Wolff, S. M. (1969). Alternate-Day Prednisone Therapy. *New Engl. J. Med.* 280, 1427–1431. doi: 10.1056/NEJM196906262802601

Neville, B. W. D. D., Allen, C. M., and Bouquot, J. E. (2008). *Oral and Maxillofacial Pathology*, 3rd Edn. Philadelphia: W.B. Saunders.

Paleri, V., Staines, K., Sloan, P., Douglas, A., and Wilson, J. (2010). Evaluation of oral ulceration in primary care. *BMJ.* 340:c2639. doi: 10.1136/bmj.c2639

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Gen.* 50, 381–389. doi: 10.1038/s41588-018-0059-2

Purves, K. L., Coleman, J. R. I., Meier, S. M., Rayner, C., Davis, K. A. S., Cheesman, R., et al. (2019). A major role for common genetic variation in anxiety disorders. *Mol. Psychiatry* 2019:31748690. doi: 10.1038/s41380-019-0559-1

Qin, L., Kao, Y. W., Lin, Y. L., Peng, B. Y., Deng, W. P., Chen, T. M., et al. (2018). Recurrent aphthous stomatitis may be a precursor or risk factor for specific cancers: A case-control frequency-matched study. *Cancer Med.* 7, 4104–4114. doi: 10.1002/cam4.1685

Redwine, L., Snow, S., Mills, P., and Irwin, M. (2003). Acute psychological stress: effects on chemotaxis and cellular adhesion molecule expression. *Psychos. Med.* 65, 598–603. doi: 10.1097/01.psy.0000079377.86193.a8

Sakamoto, N. N. A., Kochi, T., Tsuruoka, H., Pham, N. M., Kabe, I., Matsuda, S., et al. (2013). Bedtime and sleep duration in relation to depressive symptoms among Japanese workers. *J. Occupat. Health* 55, 479–486.

Schneiderman, N., Ironson, G., and Siegel, S. D. (2005). Stress and health: psychological, behavioral, and biological determinants. *Annu. Rev. Clinic. Psychol.* 1, 607–628. doi: 10.1146/annurev.clinpsy.1.102803.144141

Scully, C. (2006). Clinical practice. Aphthous ulceration. *New Engl. J. Med.* 355, 165–172. doi: 10.1056/NEJMcp054630

Scully, C. (2008). *Oral and Maxillofacial Medicine : The Basis of Diagnosis and Treatment*, 2nd Edn. Edinburgh: Churchill Livingstone.

Scully, C. (2013). *Oral and Maxillofacial Medicine*, 3rd Edn. Edinburgh: Churchill Livingstone.

Slebioda, Z., and Dorocka-Bobkowska, B. (2019). Systemic and environmental risk factors for recurrent aphthous stomatitis in a Polish cohort of patients. *Postepy Dermatol. Alergol.* 36, 196–201. doi: 10.5114/ada.2018.74638

Smaniotto, S., Martins-Neto, A. A., Dardenne, M., and Savino, W. (2011). Growth hormone is a modulator of lymphocyte migration. *Neuroimmunomodulation* 18, 309–313. doi: 10.1159/000329497

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., et al. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Gen.* 51, 793–803. doi: 10.1038/s41588-019-0397-8

Team RC (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Thomas, D. C., Lawlor, D. A., and Thompson, J. R. (2007). Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. *Ann. Epidemiol.* 17, 511–513. doi: 10.1016/j.annepidem.2006.12.005

Tugrul, S., Koçyiğit, A., Doğan, R., Eren, S. B., Senturk, E., Ozturan, O., et al. (2016). Total antioxidant status and oxidative stress in recurrent aphthous stomatitis. *Int. J. Dermatol.* 55, e130–e135. doi: 10.1111/ijd.13101

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Gen.* 50, 229–237. doi: 10.1038/s41588-017-0009-4

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Gen.* 50, 693–698. doi: 10.1038/s41588-018-0099-7

Vermeulen, J. M., Wootton, R. E., Treur, J. L., Sallis, H. M., Jones, H. J., Zammit, S., et al. (2019). Smoking and the risk for bipolar disorder: evidence from a bidirectional Mendelian randomisation study. *Br. J. Psychiatry* 2019, 1–7. doi: 10.1192/bjp.2019.202

Victoria, J. M. C.-S. J. F., Pimenta, F. J., Kalapothakis, E., and Gomez, R. S. (2005). Serotonin transporter gene polymorphism (5-HTTLPR) in patients with recurrent aphthous stomatitis. *J. Oral Pathol. Med.* 8, 494–497. doi: 10.1111/j.1600-0714.2005.00344.x

Ward, J., Tunbridge, E. M., Sandor, C., Lyall, L. M., Ferguson, A., Strawbridge, R. J., et al. (2020). The genomic basis of mood instability: identification of 46 loci in 363,705 UK Biobank participants, genetic correlation with psychiatric disorders, and association with gene expression and function. *Mol. Psychiatry* 25, 3091–3099. doi: 10.1038/s41380-019-0439-8

Wootton, R. E., Lawn, R. B., Millard, L. A. C., Davies, N. M., Taylor, A. E., Munafò, M. R., et al. (2018). Evaluation of the causal effects between subjective wellbeing and cardiometabolic health: mendelian randomisation study. *BMJ.* 362:k3788. doi: 10.1136/bmj.k3788

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Gen.* 50, 668–681. doi: 10.1038/s41588-018-0090-3

Zeng, P., and Zhou, X. (2019). Causal effects of blood lipids on amyotrophic lateral sclerosis: a Mendelian randomization study. *Hum. Mol. Gen.* 28, 688–697. doi: 10.1093/hmg/ddy384

Zhao, J., Ming, J., Hu, X., Chen, G., Liu, J., and Yang, C. (2019). Bayesian weighted Mendelian randomization for causal inference based on summary statistics. *Bioinformatics* 36, 1501–1508. doi: 10.1093/bioinformatics/btz749

Zheng, J., Baird, D., Borges, M. C., Bowden, J., Hemani, G., Haycock, P., et al. (2017). Recent Developments in Mendelian Randomization Studies. *Curr. Aller. Asthma Rep.* 4, 330–345. doi: 10.1007/s40471-017-0128-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 11 February 2021
doi: 10.3389/fgene.2021.632385

An Improved Genome-Wide Polygenic Score Model for Predicting the Risk of Type 2 Diabetes

Wei Liu[1], Zhenhuang Zhuang[2], Wenxiu Wang[2], Tao Huang[2,3,4] and Zhonghua Liu[1]**

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China, [2] Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China, [3] Center for Intelligent Public Health, Institute for Artificial Intelligence, Peking University, Beijing, China, [4] Key Laboratory of Molecular Cardiovascular Diseases, Ministry of Education, Peking University, Beijing, China

Polygenic risk score (PRS) has been shown to be predictive of disease risk such as type 2 diabetes (T2D). However, the existing studies on genetic prediction for T2D only had limited predictive power. To further improve the predictive capability of the PRS model in identifying individuals at high T2D risk, we proposed a new three-step filtering procedure, which aimed to include truly predictive single-nucleotide polymorphisms (SNPs) and avoid unpredictive ones into PRS model. First, we filtered SNPs according to the marginal association *p*-values ($p \leq 5 \times 10^{-2}$) from large-scale genome-wide association studies. Second, we set linkage disequilibrium (LD) pruning thresholds (r^2) as 0.2, 0.4, 0.6, and 0.8. Third, we set *p*-value thresholds as 5×10^{-2}, 5×10^{-4}, 5×10^{-6}, and 5×10^{-8}. Then, we constructed and tested multiple candidate PRS models obtained by the PRSice-2 software among 182,422 individuals in the UK Biobank (UKB) testing dataset. We validated the predictive capability of the optimal PRS model that was chosen from the testing process in identifying individuals at high T2D risk based on the UKB validation dataset (n = 274,029). The prediction accuracy of the PRS model evaluated by the adjusted area under the receiver operating characteristics curve (AUC) showed that our PRS model had good prediction performance [AUC = 0.795, 95% confidence interval (CI): (0.790, 0.800)]. Specifically, our PRS model identified 30, 12, and 7% of the population at greater than five-, six-, and seven-fold risk for T2D, respectively. After adjusting for sex, age, physical measurements, and clinical factors, the AUC increased to 0.901 [95% CI: (0.897, 0.904)]. Therefore, our PRS model could be useful for population-level preventive T2D screening.

Keywords: type 2 diabetes, UK Biobank, screening, prediction model, polygenic risk score

OPEN ACCESS

Edited by:
Guolian Kang,
St. Jude Children's Research Hospital, United States

Reviewed by:
Yufang Pei,
Soochow University Medical College, China
Lei Sun,
University of Toronto, Canada

****Correspondence:***
Zhonghua Liu
zhhliu@hku.hk
Tao Huang
huangtaotao@pku.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 November 2020*
Accepted: *11 January 2021*
Published: *11 February 2021*

Citation:
Liu W, Zhuang Z, Wang W, Huang T and Liu Z (2021) An Improved Genome-Wide Polygenic Score Model for Predicting the Risk of Type 2 Diabetes. Front. Genet. 12:632385. doi: 10.3389/fgene.2021.632385

INTRODUCTION

Type 2 diabetes (T2D) is a global public health problem. Identifying individuals at high risk for T2D for early targeted detection, prevention, and intervention is of great public health significance. Besides the well-known behavioral and environmental factors, T2D has a strong genetic component (Zimmet et al., 2014). Genome-wide association studies (GWASs) have successfully identified many common genetic variants that confer T2D susceptibility (Burton et al., 2007;

Scott et al., 2007; Palmer et al., 2012; Visscher et al., 2017; Pärna et al., 2020). However, all of these common genetic variants discovered by GWAS can only be able to account for a small proportion of the total heritability (McCarthy, 2010; Herder and Roden, 2011; Prasad and Groop, 2015) and thus lead to low predictive power. Polygenic risk score (PRS) that aggregates the information of many common single-nucleotide polymorphisms (SNPs) weighted by the effect size obtained from large-scale discovery GWAS has been used to predict T2D risk. PRS is expected to have better predictive power and the potential to improve the performance in T2D risk assessment (Wray et al., 2013; Khera et al., 2019).

The most commonly used method for constructing PRS is called clumping and thresholding (C + T) [or pruning and thresholding (P + T)] method, which applies two filtering steps. To retain SNPs that weakly correlated with each other, it first forms clumps around SNPs by using linkage disequilibrium (LD)-driven clumping procedure (Privé et al., 2019). Each clumping contains all SNPs within 250 kb of the index SNPs, and the degree of LD is determined by a provided pairwise correlation (r^2). Then, it removes SNPs with p-values obtained from a disease-related GWAS larger than a given threshold. C+T is regarded as the most intuitive and easiest method to generate PRS. There are two common software programs (i.e., PLINK and PRSice) that can be used to implement C + T method. Recently, Choi et al. developed a new software PRSice-2 from https://www.prsice.info (Choi and O'Reilly, 2019), which is demonstrated to be more computationally efficient and scalable than alternative PRS software while maintaining comparable predictive power.

Several researchers have tried to construct PRS models based on the C + T method for predicting T2D risk by PLINK or PRSice software. The earliest PRS model assessed the combined risk of only three variants that had been published to predispose to T2D in 6,078 individuals. The area under the receiver operating characteristics curve (AUC) of their PRS model was 0.571 (Weedon et al., 2006). Thereafter, other researchers have attempted various strategies to improve the predictive ability of the PRS model, including increasing the number of SNPs, adjusting for sex and age, some physical measurements [e.g., body mass index (BMI), diastolic blood pressure (DBP), and systolic blood pressure (SBP)] (Lango et al., 2008) and clinical factors [e.g., triglyceride level (TL), glucose level (GL), and cholesterol level (CL)] (Lyssenko et al., 2008; Meigs et al., 2008; Vassy et al., 2014). The AUC of those improved PRS models increased to some extent (range from 0.600 to 0.800). However, there are still several limitations. First, their sample sizes are not large (range from 2,776 to 39,117). Second, they only take a small number of SNPs (range from 3 to 1,000) that passed the "GWAS significant variant" derivation strategy ($p \leq 1 \times 10^{-8}$ and $r^2 < 0.2$) into account, which is too strict and might miss predictive SNPs. Amit et al. (Khera et al., 2018) constructed the PRS model across the whole genome and finally included a total of 409,258 individuals with 6,917,436 SNPs from the UK Biobank (UKB) project. The AUC was 0.730 after adjusting for age, sex, and the first four principal components for ancestry. This strategy has a slight improvement in prediction accuracy; however, the computational burden is relatively large.

To further explore the prediction capability of the PRS model in identifying high-risk individuals for T2D, we proposed a new strategy to construct PRS model by the following three-step filtering procedure to consider a statistical compromise between signal and noise. First, rather than including SNPs across the whole genome, we selected a subset of SNPs by a lenient significance threshold ($p \leq 5 \times 10^{-2}$) from a huge number of SNPs included in large-scale GWASs. Second, we set r^2 equal to 0.2, 0.4, 0.6, and 0.8 as candidate LD pruning thresholds according to Khera et al. (2018). Third, we set p-value thresholds as $5 \times 10^{-2}, 5 \times 10^{-4}, 5 \times 10^{-6}$, and 5×10^{-8} . After applying the above thresholds to the GWAS summary data, a total of 16 candidate PRS models were then generated based on the PRSice-2 software in the target samples. We conducted testing using the UKB testing dataset (n = 182,422) to avoid the model overfitting issue. Finally, we chose the best predictive PRS model among a set of candidate PRS models and evaluated it in the UKB validation dataset (n = 262,751). We also considered non-genetic risk factors, including sex, age, physical measurements, and clinical factors to further increase prediction accuracy. Real data analysis showed that our PRS model outperforms previous prediction models for T2D.

MATERIALS AND METHODS

Study Design and Population

Our study was conducted based on the UKB project[1], one of the largest prospective cohort studies (Conroy et al., 2019). Nearly half a million participants aged 40–69 years were enrolled from the United Kingdom at the time of their baseline assessment visited from 2006 to 2010 (Sudlow et al., 2015). A wide kind of physical measures (e.g., height, weight, blood pressure, and spirometry) and biological samples (e.g., blood, urine, and saliva) were collected. It then converted the limited information contained in the biological samples into widely shared cohort-wide genotyping (Bycroft et al., 2018) and whole-exome sequencing data (Khera et al., 2019). More details about the study design, method, and participants of the UKB project have been provided elsewhere (Sudlow et al., 2015).

A total of 487,409 individuals with available genotyping array and altogether 625,394 variants were originally collected from UKB. We conducted strict quality control (QC) steps described by Marees et al. (2018) based on PLINK 2.0 from https://www.cog-genomics.org/plink2. Specifically, we first filtered out SNPs and individuals with very high levels of missingness. Based on a relaxed threshold of 0.2 (>20%), we removed 89,752 variants and 30,855 subjects. There were also 262,751 SNPs removed with minor allele frequency <0.03 and 1,204 SNPs removed with a p-value of Hardy–Weinberg equilibrium Fisher's exact test $< 1 \times 10^{-6}$. Finally, 456,451 individuals and 271,687 variants passed QC and were considered in the following analysis.

[1]http://biobank.ctsu.ox.ac.uk/crystal/

The ascertainment of T2D was based on a composite of self-report, the International Classification of Diseases, Ninth Revision (ICD-9) codes of 25000 and 25010, and the International Classification of Diseases, Tenth Revision (ICD-10) code of E11. The individual-level data of T2D-related risk factors, including sex, age, physical measures [e.g., BMI, waist circumference (WC), DBP, and SBP] and clinical factors [e.g., GL, CL, TL, high-density lipoprotein (HDL), low-density lipoprotein (LDL)] were also collected from the UKB project. We further imputed the inevitably missing values of these factors by their means. To analyze individuals with a relatively homogeneous ancestry, the population was constructed centrally based on a combination of self-reported ancestry and genetically confirmed ancestry using the first 10 principal components (i.e., PC_1, ..., PC_{10}). To construct, test, and further validate the robustness of the polygenic predictor of T2D, we randomly divided the overall data into two parts, i.e., the testing and validation dataset. We assigned 40% of all individuals as the UKB testing dataset (n = 182,422) and the remaining 60% as the UKB validation dataset (n = 274,029). Other ratios were also tried to divide the testing and validation datasets, i.e., 30–70%, 50–50%, 60–40%, and 70–30%. Individuals in the UKB validation dataset were distinct from those in the UKB testing dataset. The detail of the study design is described in **Figure 1**.

Genome-Wide Polygenic Score Construction, Testing, and Validation

The PRS model provides a quantitative metric of an individual's inherited risk based on the cumulative impact of many SNPs. Generally, the PRS model can be unweighted or weighted. Suppose that we have n subjects and K SNPs that passed the first-step filtering procedure. The unweighted PRS model is defined as,

$$\mathrm{PRS}_u = G_1 + ..., G_K,$$

where $G_k (k = 1, ..., K)$ denotes the number of risk alleles for each genetic variant coded as 0, 1, or 2 under the additive genetic model. For the weighted PRS model, weights are generally assigned to each genetic variant according to the strength of association with a given disease. The weighted PRS model can be written as,

$$\mathrm{PRS}_w = \hat{\beta}_1 G_1 + \ldots, \hat{\beta}_K G_K,$$

where $\hat{\beta}_k \left(k = 1, \ldots, K\right)$ is the estimate of marginal genetic effect in the external large-scale GWAS. Both unweighted or weighted PRS models can be implemented by the PRSice-2 software (Choi and O'Reilly, 2019).

For PRS model construction, we used summary statistics from a T2D GWAS conducted among 60,786 participants with 12,056,346 SNPs of European ancestry[2] (Morris et al., 2012). Note that the UKB samples did not overlap with the samples from discovery GWAS. We first selected SNPs according to their association p-values ($p \leq 5 \times 10^{-2}$) obtained from the above GWAS, and 50,224 SNPs remained. We then considered multiple r^2 thresholds (0.2, 0.4, 0.6, and 0.8) according to Khera et al. (2018) and p-value thresholds (5×10^{-2},5×10^{-4},5×10^{-6}, and 5×10^{-8}) to conduct the second and third filtering procedures also on the DIAGRAM summary dataset. A total of 16 candidate PRS models were created for T2D based on the UKB testing dataset with 182,422 participants.

The PRS model with the best discriminative accuracy was determined based on the maximal AUC in the following logistic regression model adjusting for sex, age, and the first 10 principal components of ancestry. We use X_1, X_2 and $\boldsymbol{PC} = (\mathrm{PC}_1, \ldots, \mathrm{PC}_{10})^T$ to represent the value of sex, age, and the first 10 principal components of ancestry, respectively, where T denotes the transpose of a vector or matrix. Let Y be the T2D status with 0 and 1 representing control and case. The predictive model for T2D can be represented as,

$$\begin{aligned}\text{Logit} \quad & [P(Y = 1 \,|X_1,\, X_2,\, X_3\,,\, \mathrm{PRS}_w)] \\ & = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boldsymbol{\beta}_{\mathrm{PC}} \boldsymbol{PC} + \beta_g \mathrm{PRS}_w,\end{aligned}$$

where β_0 is the intercept, and β_1, β_2, $\boldsymbol{\beta}_{\mathrm{PC}} = (\beta_{\mathrm{PC1}}, \ldots, \beta_{\mathrm{PC10}})$, and β_g are the regression coefficients for X_1, X_2, $\boldsymbol{PC}$, and PRS_w. Then, the AUCs could be calculated with trapezoids (Fawcett, 2006), and their 95% confidence intervals (CI) could be computed by Delong's method (DeLong et al., 1988). Both AUC and their CI could be implemented directly by the "pROC" package[3] within R 3.6.3[4]. More details about this package are provided elsewhere (Robin et al., 2011). The best score created in the testing dataset carried forward into subsequent validation step.

Statistical Analysis in Validation Dataset

Baseline characteristics of the study population were described as means ± standard deviations (M ± SD) or percentages. Two independent sample t-test or chi-square test was used to compare the baseline characteristics between the UKB testing and validation datasets. Wilcoxon signed-rank test was applied to give more information about the difference of PRSs between the individuals with T2D and individuals without T2D. The relationship between PRS and T2D was determined in the UKB validation dataset based on logistic regression model adjusting for sex, age, and the first 10 principal components of ancestry ($model_1$), which can be represented as,

$$\mathrm{T2D} \sim \mathrm{PRS} + \mathrm{sex} + \mathrm{age} + \boldsymbol{PC}.$$

We stratified 274,029 participants in the UKB validation dataset as 100 groups according to the percentiles of the PRS, and then, the prevalence of T2D could be determined within each group.

To further observe the contribution of PRS, sex, age, physical measurements, and other clinical risk factors to T2D, we provided other four types of prediction models:

$$model_2 : \mathrm{T2D} \sim \mathrm{sex} + \mathrm{age} + \boldsymbol{PC}; \quad (1)$$

$$model_3 : \mathrm{T2D} \sim \mathrm{PRS}; \quad (2)$$

[2]http://diagram-consortium.org/

[3]https://cran.r-project.org/web/packages/pROC/index.html

[4]https://cran.r-project.org/bin/macosx/

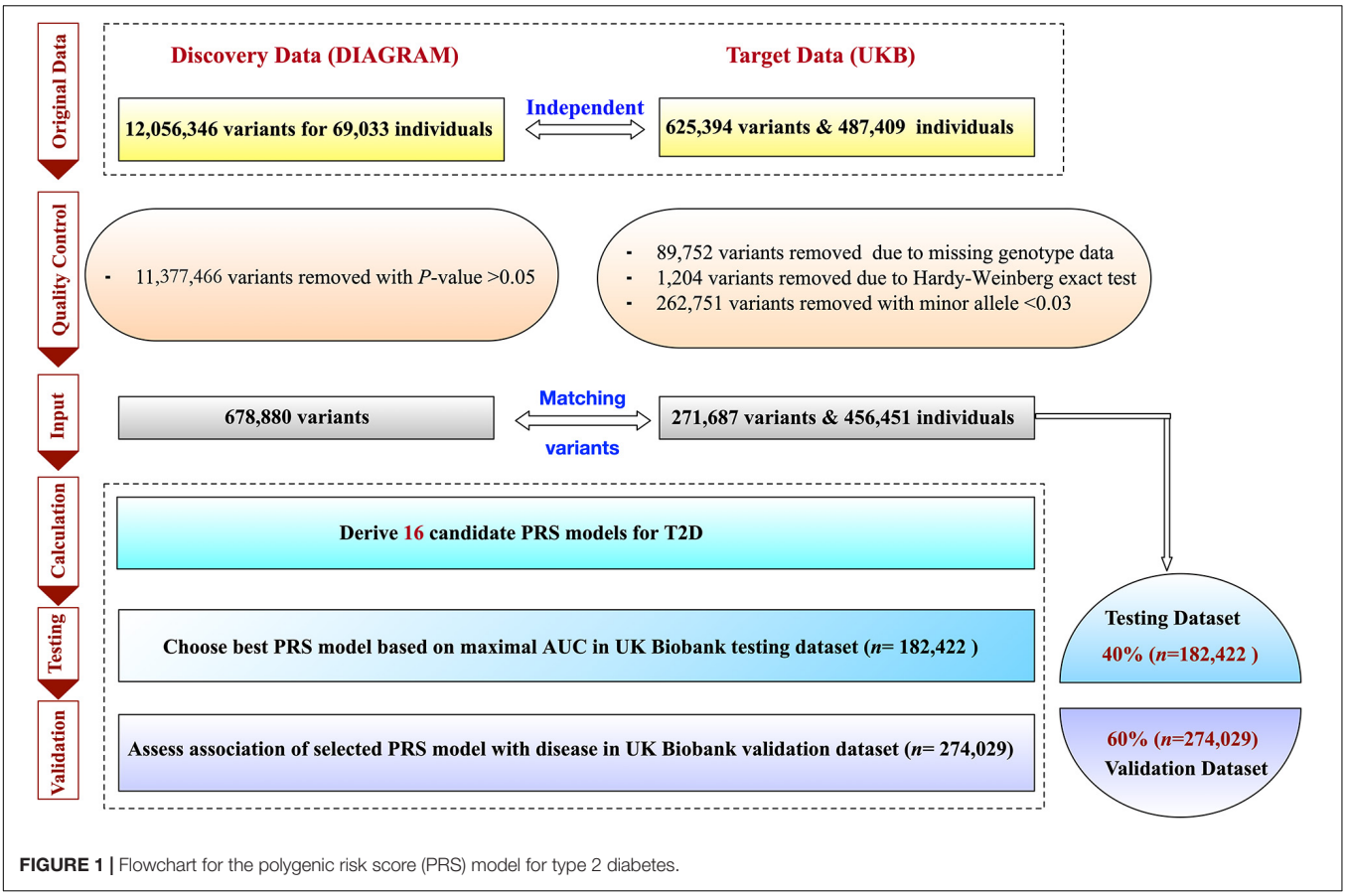


FIGURE 1 | Flowchart for the polygenic risk score (PRS) model for type 2 diabetes.

$$model_4 : \text{T2D} \sim \text{sex} + \text{age} + \boldsymbol{PC} + \text{BMI} + \text{GL} + \text{CL} + \text{HDL} + \text{LDL} + \text{TL} + \text{WC} + \text{DBP} + \text{SBP}; \quad (3)$$

$$model_5 : \text{T2D} \sim \text{PRS} + \text{sex} + \text{age} + \boldsymbol{PC} + \text{BMI} + \text{GL} + \text{CL} + \text{HDL} + \text{LDL} + \text{TL} + \text{WC} + \text{DBP} + \text{SBP}. \quad (4)$$

We have checked and did not find the presence of collinearity among the above variables. All of the above statistical analyses were conducted using R version 3.6.3 software.

RESULTS

A total of 456,451 participants collected in UKB were divided into the UKB testing dataset (n = 182,422) and the validation dataset (n = 274,029) randomly. The mean ages of participants were 57 years old, and 54% were female in both testing and validation datasets. There were nearly 5.494% (n = 10,023) participants who were cases in the testing dataset and 5.575% (n = 15,277) in the validation dataset. All of these factors were comparable at baseline. The details of baseline characteristics are shown in **Table 1**.

To obtain an optimal PRS model, we generated a total of 16 candidate PRS models implemented by PRSice-2 software. We evaluated the performance of these 16 PRS models in the UKB testing dataset and chose the best one for further validation analysis. The AUCs of these 16 candidate PRS models ranged from 0.691 to 0.792 (**Table 2**). We selected the best PRS model with the highest AUC [AUC = 0.792, 95% CI: (0.787, 0.796)] based on 25,454 SNPs when $p \leq 5 \times 10^{-2}$ and $r^2 < 0.2$. The AUCs of different ratios of the testing and validation datasets are shown in **Table 3**. We can see that the AUCs of different ratios were very close to each other, which ranged from 0.791 to 0.795. The AUC of the 40–60% ratio had the best performance in the validation dataset [AUC = 0.795, 95% CI: (0.790, 0.800)]. Additional details of PRS model construction, testing, and validation are provided in **Figure 1**.

To facilitate interpretation, we scaled PRS to have zero mean and one standard deviation. We investigated whether our PRS model could identify individuals at high T2D risk. **Figure 2** showed that the median of the standardized PRS was 0.941 for individuals with T2D versus −0.056 for individuals without T2D, a difference of 0.997 ($p < 0.00001$). From **Figure 3A**, we found that the standardized PRS approximated a normal distribution across the population with the empirical risk of T2D rising sharply in the right tail of the distribution. The PRS model identified nearly 30% of the population at greater than or equal to fivefold risk, 12% of the population at greater than or equal to sixfold risk, and the top 7% of the population at greater than or equal to sevenfold increased risk for T2D shown in **Figure 3A**. Then, we stratified the population according to the percentiles of

TABLE 1 | Baseline characteristics of the UK Biobank (UKB) testing dataset and the UKB validation dataset ($M \pm SD$ or %).

Variable	UKB testing (n = 182,422)	UKB validation (n = 274,029)	Statistics and p-value
Sex			
Male (%)	83,200 (45.609)	125,670 (45.860)	$x^2 = 2.783$, $p = 0.095$
Female (%)	99,222 (54.391)	148,359 (54.140)	
Age (years)	56.777 ± 8.020	56.809 ± 8.009	$t = -1.341$, $p = 0.179$
Physical measurements			
BMI (kg/m^2)	27.388 ± 4.758	27.404 ± 4.765	$t = -1.087$, $p = 0.277$
WC (cm)	90.250 ± 13.485	90.306 ± 13.505	$t = -1.135$, $p = 0.175$
DBP (mmHg)	82.174 ± 10.311	82.171 ± 10.313	$t = -0.118$, $p = 0.906$
SBP (mmHg)	139.924 ± 19.000	139.917 ± 19.000	$t = -0.116$, $p = 0.908$
Clinical factors			
CL (mmol/L)	5.711 ± 1.115	5.710 ± 1.117	$t = -0.314$, $p = 0.753$
GL (mmol/L)	5.119 ± 1.134	5.118 ± 1.132	$t = 0.150$, $p = 0.881$
TL (mmol/L)	1.753 ± 1.002	1.753 ± 1.000	$t = -0.010$, $p = 0.992$
HDL (mmol/L)	1.452 ± 0.357	1.453 ± 0.358	$t = -0.625$, $p = 0.532$
LDL (mmol/L)	3.556 ± 0.839	3.556 ± 0.841	$t = -0.083$, $p = 0.934$
Type 2 diabetes			
Case (%)	10,023 (5.494)	15,277 (5.575)	$x^2 = 1.342$, $p = 0.247$
Control (%)	172,399 (94.506)	258,752 (94.425)	

BMI, body mass index; CL, cholesterol level; DBP, diastolic blood pressure; GL, glucose level; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SBP, systolic blood pressure; TL, triglyceride level; WC, waist circumference.

the PRS and defined the top 10 percentiles as "high risk" group while the bottom 10 percentiles as "low risk" group. **Figure 3B** showed the prevalence of T2D increases with the percentiles of the PRS model. There were 5,642 (18.698%) cases in "high risk" group among 30,174 individuals, while only 282 (0.935%) cases in the "low risk" group, corresponding to a nearly 20-fold increase in the risk of T2D comparing the top 10 percentiles versus the bottom 10 percentiles.

TABLE 2 | The predictive power of candidate polygenic risk score (PRS) models for type 2 diabetes (T2D).

Tuning parameter	SNP number	AUC (95% CI)
$p \leq 5 \times 10^{-8}$ and $r^2 < 0.2$	363	0.706 (0.701–0.711)
$p \leq 5 \times 10^{-8}$ and $r^2 < 0.4$	486	0.702 (0.697–0.707)
$p \leq 5 \times 10^{-8}$ and $r^2 < 0.6$	670	0.696 (0.691–0.701)
$p \leq 5 \times 10^{-8}$ and $r^2 < 0.8$	957	0.691 (0.686–0.697)
$p \leq 5 \times 10^{-6}$ and $r^2 < 0.2$	750	0.715 (0.710–0.720)
$p \leq 5 \times 10^{-6}$ and $r^2 < 0.4$	1,013	0.709 (0.704–0.714)
$p \leq 5 \times 10^{-6}$ and $r^2 < 0.6$	1,335	0.701 (0.696–0.706)
$p \leq 5 \times 10^{-6}$ and $r^2 < 0.8$	1,853	0.696 (0.691–0.701)
$p \leq 5 \times 10^{-4}$ and $r^2 < 0.2$	2,616	0.736 (0.732–0.741)
$p \leq 5 \times 10^{-4}$ and $r^2 < 0.4$	3,394	0.726 (0.721–0.731)
$p \leq 5 \times 10^{-4}$ and $r^2 < 0.6$	4,299	0.715 (0.710–0.720)
$p \leq 5 \times 10^{-4}$ and $r^2 < 0.8$	5,690	0.708 (0.703–0.713)
$p \leq 5 \times 10^{-2}$ and $r^2 < 0.2$	**25,454**	**0.792 (0.787–0.796)**
$p \leq 5 \times 10^{-2}$ and $r^2 < 0.4$	32,600	0.782 (0.777–0.787)
$p \leq 5 \times 10^{-2}$ and $r^2 < 0.6$	40,001	0.771 (0.766–0.776)
$p \leq 5 \times 10^{-2}$ and $r^2 < 0.8$	50,224	0.760 (0.755–0.765)

AUC was determined using a logistic regression model adjusted for sex, age, and the first 10 principal components of ancestry. The highest AUC is denoted by the bold values.

We further investigated the contribution of polygenic predictor, sex, age, physical measurements, and clinical factors in identifying individuals at high risk of T2D. **Table 4** showed that the AUCs of $model_3$, which only included PRS into the prediction model without adjusting for any other covariates, was 0.749 [95% CI: (0.744,0.754)] in the testing dataset and 0.755 [95% CI: (0.752, 0.755)] in the validation dataset. Interestingly, if only considering sex, age, and the first 10 principal components of ancestry into the model, the AUC was 0.667 [95% CI: (0.663, 0.672)]. After adding PRS, the AUC reached 0.795 [95% CI: (0.790, 0.800)], which increased about 13% than $model_2$. The AUC of $model_4$ (i.e., considering sex, age, ***PC***, BMI, WC, DBP, SBP, GL, CL, HDL, LDL, and TL simultaneously) was 0.880 [95% CI: (0.878, 0.888)] and raised to 0.901 [95% CI: (0.897, 0.904)] in the validation dataset when adding PRS into the model. In brief, the polygenic score indeed helps to identify high-risk individuals for T2D, while the role of T2D-related covariates could also help increase prediction accuracy. As showed in **Table 5**, PRS, sex, age, physical measurements, and most clinical factors were all significantly associated with T2D ($p < 0.0001$).

DISCUSSION

Our results showed that the AUC of the best PRS model was 0.795 after adjusting for sex, age, and the first 10 principal components of ancestry. It demonstrated that the PRS was really helpful for identifying individuals at high risk of developing T2D. Meanwhile, the distributions of the PRS in cases and controls were substantially different from each other, i.e., the median PRS of cases (0.941) was much higher than that of the controls (−0.056). Moreover, about 30% of participants were at greater than or equal to fivefold increased risk of developing T2D, 12%

TABLE 3 | Area under the receiver operating characteristics curves (AUCs) of different ratios of the testing and validation dataset when $p \leq 5 \times 10^{-2}$ and $r^2 < 0.2$.

Dataset	30–70%	40–60%	50–50%	60–40%	70–30%
Testing	0.791	0.792	0.794	0.795	0.794
	(0.781–0.791)	(0.787–0.796)	(0.790–0.800)	(0.791–0.799)	(0.790–0.799)
Validation	0.794	0.795	0.793	0.792	0.791
	(0.790–0.799)	(0.790–0.800)	(0.789–0.797)	(0.787–0.796)	(0.781–0.791)

AUC was determined using a logistic regression model adjusted for sex, age, and first 10 principal components of ancestry.

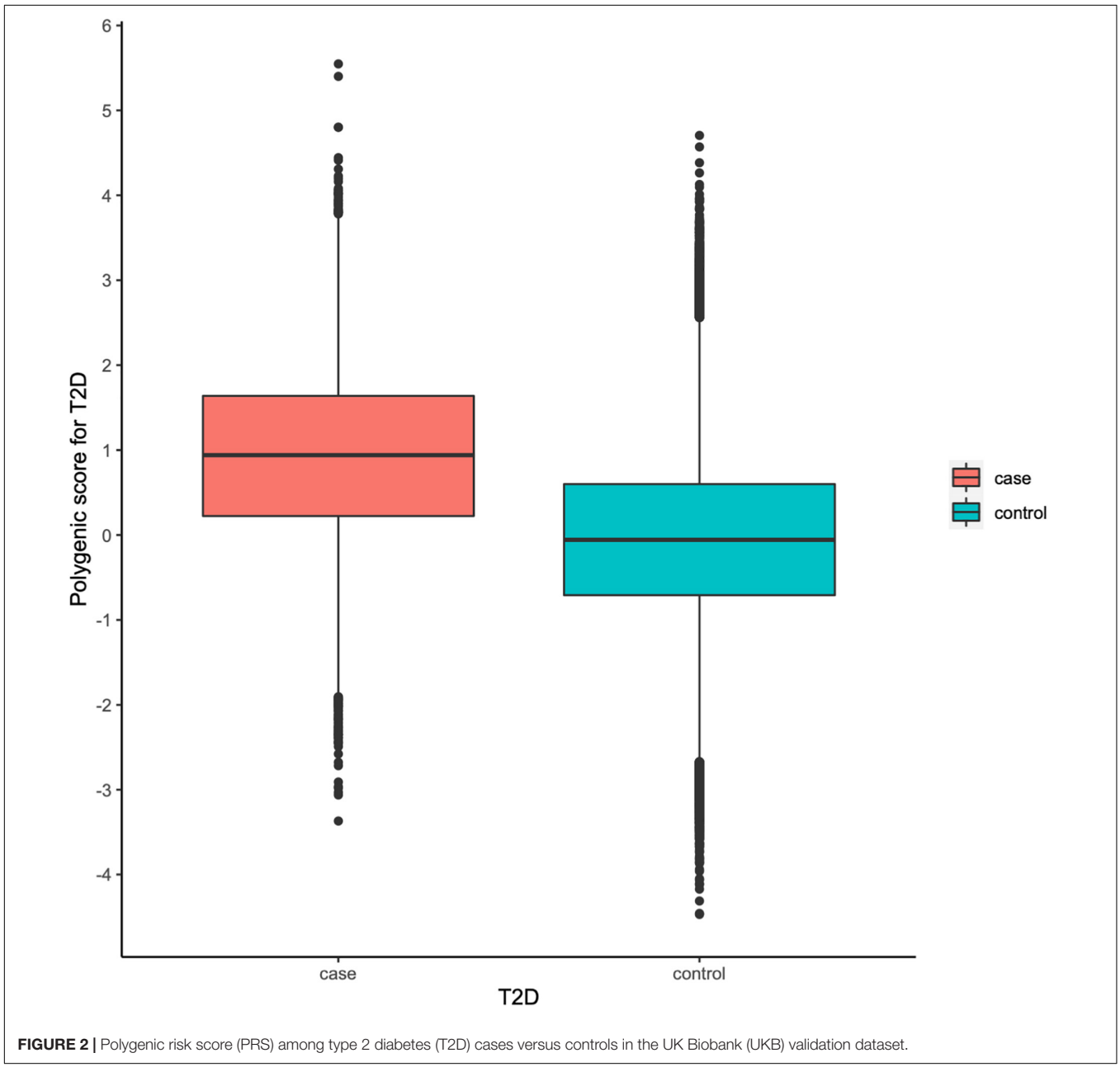


FIGURE 2 | Polygenic risk score (PRS) among type 2 diabetes (T2D) cases versus controls in the UK Biobank (UKB) validation dataset.

were at greater than or equal to sixfold risk, and the top 7% were at greater than or equal to sevenfold increased risk. Particularly, the stratified PRS according to their percentiles showed that the "high-risk" group is strongly associated with the risk of T2D.

The above results suggest that our PRS model can be used as a powerful tool in identifying individuals at high risk of T2D; improved previous studies that summarized in **Table 6**. The AUC of the PRS model assessed with only three SNPs that had been

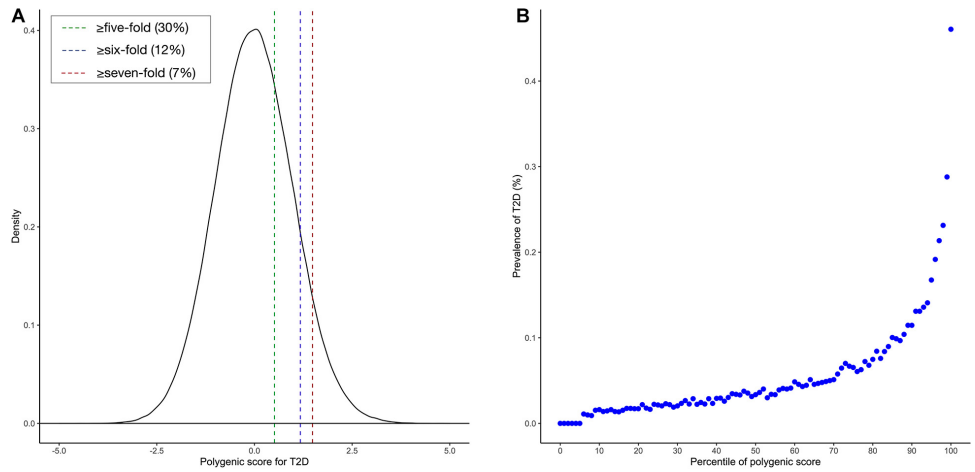


FIGURE 3 | Risk for type 2 diabetes (T2D) according to polygenic risk score (PRS). **(A)** Distribution of PRS for T2D in the UK Biobank (UKB) validation dataset (n = 301,736). The x-axis represents PRS for T2D, which was scaled to have zero mean and one standard deviation. Dotted lines reflect the proportion of the population with five-, six-, and seven-fold increased risk versus the remainder of the population, respectively. The odds ratio was assessed in a logistic regression model adjusting for sex, age, and the first 10 principal components of ancestry. **(B)** Prevalence of T2D according to 100 groups of the UKB validation dataset stratified according to the percentile of the PRS for T2D.

TABLE 4 | Area under the receiver operating characteristics curve (AUC) of different models in the testing and validation dataset.

Dataset	Mean	$model_2$	$model_3$	$model_1$	$model_4$	$model_5$
Testing	−0.003	0.671 (0.666–0.676)	0.749 (0.744–0.754)	0.792 (0.787–0.796)	0.886 (0.882–0.889)	0.902 (0.899–0.905)
Validation	−0.003	0.667 (0.663–0.672)	0.755 (0.752–0.755)	0.795 (0.790–0.800)	0.882 (0.878–0.888)	0.901 (0.897–0.904)

$model_1$: AUC was determined using a logistic regression model adjusted for sex, age, and the first 10 principal components of ancestry. $model_2$: AUC was determined using a logistic regression model only considering sex and age. $model_3$: AUC was determined using a logistic regression model only considering genome-wide polygenic score. $model_4$: AUC was determined using a logistic regression model considering demographic factors, physical measurements, and clinical factors. $model_5$: AUC was determined using a logistic regression model adjusted for sex, age, body mass index, waist circumference, diastolic blood pressure, systolic blood pressure, glucose level, cholesterol level, high-density lipoprotein, low-density lipoprotein, triglyceride level, and the first 10 principal components of ancestry.

published to predispose to T2D in 6,078 individuals was 0.571 (Weedon et al., 2006). After including more SNPs, Lango et al. (2008) constructed a PRS model with 18 SNPs and obtained an AUC of 0.600 (Lango et al., 2008). A later study with 22 SNPs had an AUC of 0.570 (Chatterjee et al., 2013) and allowed for the identification of 3.0% of the population at twofold or higher than average risk for T2D. Notably, the above three studies with smaller sample sizes (range from 4,907 to 39,117), and a smaller number of SNPs (range from 3 to 22) had relatively poor predictive performance compared to our study (AUC = 0.755) with 25,454 SNPs among 274,029 individuals.

TABLE 5 | Parameter estimations under $model_5$ in validation dataset.

Variables	Estimate beta	Stand error	Z	p-value
(Intercept)	24.500	0.495	49.474	$< 2 \times 10^{-16}$
PRS	12370.000	167.400	73.943	$< 2 \times 10^{-16}$
CL	−0.591	0.057	−10.377	$< 2 \times 10^{-16}$
HDL	0.051	0.063	0.876	0.381
LDL	0.010	0.068	0.140	0.888
TL	0.285	0.013	21.826	$< 2 \times 10^{-16}$
Sex	−0.214	0.028	−7.731	1.070×10^{-14}
WC	0.045	0.002	28.356	$< 2 \times 10^{-16}$
BMI	0.036	0.004	9.325	$< 2 \times 10^{-16}$
Age	0.060	0.002	38.401	$< 2 \times 10^{-16}$
DBP	−0.018	0.001	−13.928	$< 2 \times 10^{-16}$
SBP	0.005	0.001	7.626	2.410×10^{-16}
GL	0.449	0.006	69.917	$< 2 \times 10^{-16}$
PC10	0.020	0.004	4.726	2.280×10^{-16}

BMI, body mass index; CL, cholesterol level; DBP, diastolic blood pressure; GL, glucose level; PRS, genome-wide polygenic score; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SBP, systolic blood pressure; TL, triglyceride level; WC, waist circumference.

In addition, we highlight the role of non-genetic risk factors, i.e., sex, age, physical measurements, and clinical factors. When adjusting for sex and age, Meigs et al. (2008) obtained an AUC of 0.581 among 2,776 individuals, Vassy et al. (2014) provided an AUC of 0.726 among 11,883 people, and the AUC of Läll et al. (2017) reached 0.740. Interestingly, the study that handled nearly 7 million variants in 288,978 individuals only generated an AUC of 0.730 after adding sex and age, which was smaller than ours (0.795) including only 25,454 SNPs (Khera et al., 2018). They further reported that 3.5% of the population had inherited a genetic predisposition that conferred greater than or equal to threefold increased risk for T2D, 0.2% of the population greater than or equal to fourfold, and 0.05 of the population greater than or equal to fivefold. Their study differs from ours in four aspects. First, our study has larger sample size (456,451 versus 409,258). Second, we first perform SNP selection based on genome-wide

TABLE 6 | A comprehensive comparison with other researches.

Year	SNP	N	Case\Control	Case/N (%)	Dataset	AUC	Ethnicity	Covariates
Weedon et al., 2006	3	6,078	2,409\3,669	39	UKCS	0.571	British	–
Lango et al., 2008	18	4,907	2,309\2598	47	GoDARTS	0.600	Scotland	–
Lango et al., 2008	18	4,907	2,309\2598	47	GoDARTS	0.800	Scotland	Age, BMI, and sex
Lyssenko et al., 2008	16	18,831	2,201\16,630	11.68	MPP and BS	0.750	Finland	Sex, age, family history, BMI, BP, TL, and GL
Chatterjee et al., 2013	22	39,117	130\38,987	0.3	–	0.570	Caucasian	–
Chatterjee et al., 2013	22	39,117	130\38,987	0.3	–	0.740	Caucasian	Sex, age, and family history
Läll et al., 2017	1,000	10,273	1,181\9,092	11.5	EBC	0.74	Estonia	Sex and age
Läll et al., 2017	1,000	10,273	1,181\9,092	11.5	EBC	0.767	Estonia	Sex, age, and BMI
Läll et al., 2017	1,000	10,273	1,181\9,092	11.5	EBC	0.790	Estonia	Sex, age, BMI, BP, GL, physical activity, smoking, and food consumption
Khera et al., 2018	6,917,436	288,978	5,853\283,125	2	UKB	0.730	British	Sex and age
–	25,454	274,029	18,176\283,560	6	UKB	0.755	British	–
–	25,454	274,029	18,176\283,560	6	UKB	0.795	British	Sex and age
–	25,454	274,029	18,176\283,560	6	UKB	0.901	British	Sex, age, WC, BMI, SBP, DBP, GL, CL, TL, HDL, and LDL

association p-values ($p \leq 5 \times 10^{-2}$) so that we included more predictive SNPs (25,454) and avoided spurious SNPs into our PRS model. Third, they used the first 4 principal components of ancestry, while we used the first 10 principal components of ancestry for a better control of population stratification. Fourth, we generated PRS based on the more computationally efficient and scalable PRSice-2 software, while they used LDpred program (Ripke et al., 2015), which is much slower than PRSice-2. Those differences explain why our PRS model has better predictive power. Certainly, we also tried to incorporate more non-genetic risk factors, and the AUC increased from 0.755 to 0.901. Our study is thus more accurate in identifying individuals at low and high risk of developing T2D.

Our study has multiple strengths. First, we construct the PRS model based on the UKB dataset, which is one of the largest prospective cohort studies with comprehensive and abundant personal information, as well as high-quality genotyping data in the world. Second, we choose SNPs into our PRS model based on our proposed three-step filtering procedure. This approach is simple to implement and has a very good prediction performance. Third, we include new physical measurements and clinical factors (i.e., WC, DBP, HDL, and LDL) in our predictive model to increase prediction accuracy. Fourth, we adopted a new PRS software PRSice-2, which has been shown to outperform other competing methods and software in terms of prediction accuracy and computational speeds (Choi and O'Reilly, 2019).

Although the present study has made important contributions in identifying individuals with increased risk of developing T2D; however, there exists one major limitation. Individuals in the UKB dataset are primarily European ancestry; the specific PRS calculated here may not have optimal predictive power for other ethnic groups because the allele frequencies, LD patterns, and effect sizes of common SNPs may be different across populations with different ethnic backgrounds.

In conclusion, our findings show that the PRS model is highly predictive of T2D risk even based on genetic data only, and the prediction accuracy improves after including non-genetic risk factors, suggesting that our PRS model can be used as a powerful tool for preventive T2D screening.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZL and TH initiated the study. WL developed the strategy, performed the data analysis, and completed the manuscript writing. ZZ and WW contributed to the data collection and manuscript reservation. All authors contributed to the article and approved the submitted version.

FUNDING

The work was supported by grants from the start-up research fund at University of Hong Kong and the National Key Research and Development Project (2019YFC2003400).

ACKNOWLEDGMENTS

The authors thank the UK Biobank project for providing the individual-level data to support out analysis and consortium DIAGRAM for sharing their summary-level data for type 2 diabetes freely. The authors also thank the associate editor and reviewers for their constructive comments.

REFERENCES

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* 45, 400–405. doi: 10.1038/ng.2579

Choi, S. W., and O'Reilly, P. F. (2019). PRSice-2: polygenic risk score software for biobank-scale data. *GigaScience* 8:giz082. doi: 10.1093/gigascience/giz082

Conroy, M., Sellors, J., Effingham, M., Littlejohns, T. J., Boultwood, C., Gillions, L., et al. (2019). The advantages of UK Biobank's open-access strategy for health research. *J. Intern. Med.* 286, 389–397. doi: 10.1111/joim.12955

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Herder, C., and Roden, M. (2011). Genetics of type 2 diabetes: pathophysiologic and clinical relevance. *Eur. J. Clin. Invest.* 41, 679–692. doi: 10.1111/j.1365-2362.2010.02454.x

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50, 1219–1224. doi: 10.1038/s41588-018-0183-z

Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., et al. (2019). Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* 177, 587.e9–596.e9.

Läll, K., Mägi, R., Morris, A., Metspalu, A., and Fischer, K. (2017). Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet. Med.* 19, 322–329. doi: 10.1038/gim.2016.103

Lango, H., Palmer, C. N., Morris, A. D., Zeggini, E., Hattersley, A. T., McCarthy, M. I., et al. (2008). Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes Metab. Res. Rev.* 57, 3129–3135. doi: 10.2337/db08-0504

Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., et al. (2008). Clinical risk factors, DNA variants, and the development of type 2 diabetes. *New Engl. J. Med.* 359, 2220–2232. doi: 10.1056/nejmoa0801869

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et al. (2018). A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 27:e1608. doi: 10.1002/mpr.1608

McCarthy, M. I. (2010). Genomics, type 2 diabetes, and obesity. *N. Engl. J. Med.* 363, 2339–2350.

Meigs, J. B., Shrader, P., Sullivan, L. M., McAteer, J. B., Fox, C. S., Dupuis, J., et al. (2008). Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* 359, 2208–2219. doi: 10.1056/nejmoa0804742

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* 44:981. doi: 10.1038/ng.2383

Palmer, N. D., McDonough, C. W., Hicks, P. J., Roh, B. H., Wing, M. R., An, S. S., et al. (2012). A genome-wide association search for type 2 diabetes genes in African Americans. *PLoS One* 7:e29202. doi: 10.1371/journal.pone.0029202

Pärna, K., Snieder, H., Läll, K., Fischer, K., and Nolte, I. (2020). Validating the doubly weighted genetic risk score for the prediction of type 2 diabetes in the lifelines and estonian biobank cohorts. *Genet. Epidemiol.* 44, 589–600. doi: 10.1002/gepi.22327

Prasad, R. B., and Groop, L. (2015). Genetics of type 2 diabetes—pitfalls and possibilities. *Genes* 6, 87–123. doi: 10.3390/genes6010087

Privé, F., Vilhjálmsson, B. J., Aschard, H., and Blum, M. G. B. (2019). Making the most of clumping and thresholding for polygenic scores. *Am. J. Hum. Genet.* 105, 1213–1221. doi: 10.1016/j.ajhg.2019.11.001

Ripke, S., Neale, B., Corvin, A., Walters, J. R., Farh, K. H., Holmans, P., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316, 1341–1345.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12:e1001779. doi: 10.1371/journal.pmed.1001779

Vassy, J. L., Hivert, M.-F., Porneala, B., Dauriz, M., Florez, J. C., Dupuis, J., et al. (2014). Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes Metab. Res. Rev.* 63, 2172–2182. doi: 10.2337/db13-1663

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Weedon, M. N., McCarthy, M. I., Hitman, G., Walker, M., Groves, C. J., Zeggini, E., et al. (2006). Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med.* 3:e374. doi: 10.1371/journal.pmed.0030374

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14, 507–515. doi: 10.1038/nrg3457

Zimmet, P. Z., Magliano, D. J., Herman, W. H., and Shaw, J. E. (2014). Diabetes: a 21st century challenge. *Lancet Diabetes Endocrinol.* 2, 56–64. doi: 10.1016/s2213-8587(13)70112-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 04 March 2021
doi: 10.3389/fgene.2021.642227

Selecting Classification Methods for Small Samples of Next-Generation Sequencing Data

Jiadi Zhu[1], Ziyang Yuan[2], Lianjie Shu[3], Wenhui Liao[4], Mingtao Zhao[5]* and Yan Zhou[2]**

[1] Department of Mathematics and Statistics, Xidian University, Xi'an, China, [2] Shenzhen Key Laboratory of Advanced Machine Learning and Applications, College of Mathematics and Statistics, Institute of Statistical Sciences, Shenzhen University, Shenzhen, China, [3] Faculty of Business Administration, University of Macau, Macau, China, [4] GuangDong University of Finance, Guangzhou, China, [5] Institute of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong, Hong Kong

Reviewed by:
Qi Shen,
Icahn School of Medicine at Mount Sinai, United States
Xiaobo Guo,
Sun Yat-sen University, China

***Correspondence:**
Wenhui Liao
lwh_gduf@sina.com
Mingtao Zhao
mingtao.zhao@outlook.com
Yan Zhou
zhouy1016@szu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *15 December 2020*
Accepted: *01 February 2021*
Published: *04 March 2021*

Citation:
Zhu J, Yuan Z, Shu L, Liao W, Zhao M and Zhou Y (2021) Selecting Classification Methods for Small Samples of Next-Generation Sequencing Data. Front. Genet. 12:642227. doi: 10.3389/fgene.2021.642227

Next-generation sequencing has emerged as an essential technology for the quantitative analysis of gene expression. In medical research, RNA sequencing (RNA-seq) data are commonly used to identify which type of disease a patient has. Because of the discrete nature of RNA-seq data, the existing statistical methods that have been developed for microarray data cannot be directly applied to RNA-seq data. Existing statistical methods usually model RNA-seq data by a discrete distribution, such as the Poisson, the negative binomial, or the mixture distribution with a point mass at zero and a Poisson distribution to further allow for data with an excess of zeros. Consequently, analytic tools corresponding to the above three discrete distributions have been developed: Poisson linear discriminant analysis (PLDA), negative binomial linear discriminant analysis (NBLDA), and zero-inflated Poisson logistic discriminant analysis (ZIPLDA). However, it is unclear what the real distributions would be for these classifications when applied to a new and real dataset. Considering that count datasets are frequently characterized by excess zeros and overdispersion, this paper extends the existing distribution to a mixture distribution with a point mass at zero and a negative binomial distribution and proposes a zero-inflated negative binomial logistic discriminant analysis (ZINBLDA) for classification. More importantly, we compare the above four classification methods from the perspective of model parameters, as an understanding of parameters is necessary for selecting the optimal method for RNA-seq data. Furthermore, we determine that the above four methods could transform into each other in some cases. Using simulation studies, we compare and evaluate the performance of these classification methods in a wide range of settings, and we also present a decision tree model created to help us select the optimal classifier for a new RNA-seq dataset. The results of the two real datasets coincide with the theory and simulation analysis results. The methods used in this work are implemented in the open-scource R scripts, with a source code freely available at https://github.com/FocusPaka/ZINBLDA.

Keywords: RNA-seq data, classification, PLDA, NBLDA, ZIPLDA, ZINBLDA

1. INTRODUCTION

RNA sequencing (RNA-seq), which involves directly sequencing complementary DNAs and aligning the sequences to the reference genome or transcriptome, has emerged as a powerful technology for measuring gene expression (Mardis, 2008; Morozova et al., 2009; Wang et al., 2009). In recent years, the affordability and effectiveness of RNA-seq has resulted in its application in biological and medical studies, such as genomics research (Nagalakshmi et al., 2008; Trapnell et al., 2010) and clinical use (Berger et al., 2010; Biesecker et al., 2012). Unlike microarray technology, RNA-seq allows for the detection of novel transcripts with low background signals. One of the biological applications of RNA-seq is inferring differential expression (DE) genes between different conditions or tissues. Existing popular methods include edgeR (Robinson and Smyth, 2008; Robinson et al., 2010), DESeq2 (Love et al., 2014), and LFCseq (Lin et al., 2014). Another important application is the diagnosis of diseases. Numerous discriminant methods have been proposed for the diagnosis of diseases using microarray data, such as diagonal linear discriminant analysis and diagonal quadratic discriminant analysis in Dudoit et al. (2002). In previous RNA-seq experiments, the read counts (the number of short reads mapped to the reference genome) have been used to measure the expression level. However, because the expression matrix entries are non-negative integers, classification methods that follow a Gaussian distribution may not perform well for RNA-seq data.

Classification methods based on different discrete distributions have been proposed for RNA-seq data. Witten (2011) assumed RNA-seq data follow a Poisson distribution and proposed a Poisson linear discriminant analysis (PLDA) method. Comparison studies (Tan et al., 2014) have shown that PLDA performs much better than the method used for microarray data when classifying RNA-seq data. Considering the overdispersion of RNA-seq data, Dong et al. (2016) assumed that data follow a negative binomial distribution and developed a negative binomial linear discriminant analysis (NBLDA) method. Zhou et al. (2018) found excess zeros in real RNA-seq data and proposed a zero-inflated Poisson logistic discriminant analysis (ZIPLDA) method, which assumes RNA-seq data follow a mixture distribution with a point mass at zero and a Poisson distribution.

Due to the shallow sequence depth and dispersed biological replicates, there may be excess zeros and overdispersion in a real RNA-seq dataset, which should be considered when conducting data analysis. For instance, the real dataset TCGA-LIHC, which includes a cancerous and normal group, contains about 43.24% zeros of all numerical values, and the estimated dispersion parameter is 1.12. Therefore, a natural assumption would be to extend the existing discrete distribution to a mixture distribution with a point mass at zero and a negative binomial distribution. We call this method zero-inflated negative binomial logistic discriminant analysis (ZINBLDA). To obtain the model, which is similar to ZIPLDA, we built a mixture distribution with a point mass at zero and a negative binomial distribution for the remaining data. We then estimated the parameters in the model. Finally, we obtained a classifier by Bayes rule to predict for a future observation. We also analyzed the relationship between the above four classification methods, and the resulting discriminant scores for the four classification methods showed that they can transform into each other in some cases. We examined these four methods from the perspective of their parameters and determined how the parameters provide the link between the selected optimal method and the model classification performance. In addition, we built a decision tree to help us select the optimal classifier from these four methods for a new dataset.

The remainder of the article is organized as follows. In section 2, we review the existing three classification methods and propose the ZINBLDA method for overdispersion RNA-seq data with an excess of zeros. We also give the estimation of the parameters in the model in detail. We further discuss the transformation relations between the four methods. Section 3 discusses the results of the simulation studies that were conducted to evaluate the performance of the four methods in a wide range of settings. This section also presents a decision tree that was built to select the optimal classifier from these four methods for a new dataset. In section 4, we employ the four methods to analyze two real RNA-seq datasets and evaluate their performance. Finally, we conclude the work with a discussion of the findings and future directions.

2. CLASSIFICATION METHODS

There are three existing classification methods for RNA-seq data: PLDA (Witten, 2011), NBLDA (Dong et al., 2016), and ZIPLDA (Zhou et al., 2018). We propose a new discriminant analysis method to model overdispersion RNA-seq data with excess zeros. We examined these four methods from the perspective of their parameters and analyzed the transformation relations between the methods.

Before introducing the methods, we must first specify some notations used in this work. In this paper, K is the number of classes, and X_{ki_kg} denotes the number of read counts that are mapped to gene g in sample i_k of class k, where $k = 1, \cdots, K$; $i_k = 1, \cdots, n_k$; and $g = 1, \cdots, G$. Specifically, there are n_k samples in class k, and $n = \sum_{k=1}^{K} n_k$ denotes the total number of samples for all classes.

2.1. Principle of the Classifiers

The principle of the classifiers is applicable for the following four classifiers. Suppose that for the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ we wished to classify a new observation $\boldsymbol{x}^* = (X_1^*, \cdots, X_G^*)^T$. If y^* is the unknown label of $\boldsymbol{x}^*$, by Bayes' rule

$$P(y^* = k|\boldsymbol{x}^*) \propto f_k(\boldsymbol{x}^*)\pi_k, \tag{1}$$

where f_k is the probability density function of an observation in class k, and π_k is the prior probability that an observation belongs to class k. In general, we can use $\pi_k = n_k/n$ to satisfy $\sum_{k=1}^{K} \pi_k$=1. We define a discriminant score function as $d_k(\boldsymbol{x}^*) = \log[P(\boldsymbol{x}^*|y^* = k)\pi_k]$ on the basis of formula (1) and assign a new observation to the class for which the discriminant score is the highest.

2.2. Poisson Linear Discriminant Analysis

For PLDA, Witten (2011) assumed that RNA-seq data follow a Poisson distribution, that is,

$$X_{ki_kg}|y_{i_k} = k \sim Poisson(\mu_{ki_kg}), \quad \mu_{ki_kg} = d_{kg}s_{i_k}\lambda_g, \tag{2}$$

where μ_{ki_kg} is the expectation for gene g in sample i_k of class k, s_{i_k} is the size factor used to identify individuals in the kth class, λ_g is the total number of read counts for gene g, and d_{kg} allows for the differential expression of gene g between the different classes. Following the expression of (2), the probability density function is

$$P(X_{ki_kg} = x_{ki_kg}) = \frac{\mu_{ki_kg}^{x_{ki_kg}}}{(x_{ki_kg})!}e^{-\mu_{ki_kg}}.$$

Thus, according to formula (1), the discriminant score of PLDA is obtained by

$$d_k(\boldsymbol{x}^*) = \sum_{g=1}^{G} X_g^* \log(d_{kg}) - s^* \sum_{g=1}^{G} \lambda_g d_{kg} + \log \pi_k + C, \tag{3}$$

where s^* is the size factor of test observation, and C represents a constant that is unrelated to the class label.

2.3. Negative Binomial Linear Discriminant Analysis

Modeling RNA-seq data with a negative binomial distribution instead of a Poisson distribution is a natural extension. Dong et al. (2016) proposed NBLDA to allow for cases where variance is greater than or equal to the mean, and they also demonstrated that NBLDA is more suitable when biological replicates are available. The negative binomial distribution is expressed as

$$X_{ki_kg}|y_{i_k} = k \sim NB(\mu_{ki_kg}, \phi_g), \quad \mu_{ki_kg} = d_{kg}s_{i_k}\lambda_g, \tag{4}$$

where ϕ_g is a non-negative dispersion parameter, and the rest of parameters are the same as for PLDA. Therefore, the probability density function of $X_{ki_kg} = x_{ki_kg}$ in model (4) is

$$P(X_{ki_kg} = x_{ki_kg}) = \frac{\Gamma(x_{ki_kg} + \phi_g^{-1})}{(x_{ki_kg})!\Gamma(\phi_g^{-1})}\left(\frac{\mu_{ki_kg}\phi_g}{1+\mu_{ki_kg}\phi_g}\right)^{x_{ki_kg}} \left(\frac{1}{1+\mu_{ki_kg}\phi_g}\right)^{\phi_g^{-1}}.$$

Similarly, the discriminant score can be obtained by

$$\begin{aligned} d_k(\boldsymbol{x}^*) = &\sum_{g=1}^{G} X_g^*[\log(d_{kg}) - \log(1 + s^*\lambda_g d_{kg}\phi_g)] \\ &- \sum_{g=1}^{G} \phi_g^{-1} \log(1 + s^*\lambda_g d_{kg}\phi_g) + \log \pi_k + C. \end{aligned} \tag{5}$$

2.4. Zero-Inflated Poisson Logistic Discriminant Analysis

Considering data with excess zeros due to missing records or no observation signal, Zhou et al. (2018) proposed ZIPLDA method, which assumes that data follow a zero-inflated Poisson distribution. The distribution is expressed as

$$X_{ki_kg} \sim \begin{cases} \delta_{\{0\}}, & p_{ki_kg}, \\ \text{Poisson}(\mu_{ki_kg}), & (1 - p_{ki_kg}), \end{cases}$$

where $\delta_{\{0\}}$ denotes the point mass at zero, p_{ki_kg} is the probability of $\delta_{\{0\}}$ in gene g of sample i_k in class k, and μ_{ki_kg} is same as in the former two classifiers. Thus, the probability of X_{ki_kg} is written as

$$P(X_{ki_kg}) = \begin{cases} p_{ki_kg} + (1 - p_{ki_kg})e^{-\mu_{ki_kg}}, & X_{ki_kg} = 0, \\ (1 - p_{ki_kg})\frac{\mu_{ki_kg}^{X_{ki_kg}}}{(X_{ki_kg})!}e^{-\mu_{ki_kg}}, & X_{ki_kg} > 0. \end{cases}$$

Additionally, the probability density function of $X_{ki_kg} = x_{ki_kg}$ is

$$P(X_{ki_kg} = x_{ki_kg}) = \left[p_{ki_kg} + (1 - p_{ki_kg})e^{-\mu_{ki_kg}}\right]^{I_{(x_{ki_kg}=0)}} \left[(1 - p_{ki_kg})\frac{\mu_{ki_kg}^{x_{ki_kg}}}{(x_{ki_kg})!}e^{-\mu_{ki_kg}}\right]^{I_{(x_{ki_kg}>0)}}.$$

Finally, the discriminant score $d_k(x^*)$ is,

$$\begin{aligned} d_k(\boldsymbol{x}^*) = &\sum_{g=1}^{G} I_{(X_g^*=0)} \log\left(\hat{p}_{kg}^* + (1 - \hat{p}_{kg}^*)e^{-d_{kg}s^*\lambda_g}\right) \\ &- \sum_{g=1}^{G} I_{(X_g^*>0)} d_{kg}s^*\lambda_g + \sum_{g=1}^{G} I_{(X_g^*>0)} \log(1 - \hat{p}_{kg}^*) \\ &+ \sum_{g=1}^{G} I_{(X_g^*>0)} X_g^* \log(d_{kg}) + \log \pi_k + C. \end{aligned} \tag{6}$$

2.5. Zero-Inflated Negative Binomial Logistic Discriminant Analysis

2.5.1. Model

In this section, we extend the zero-inflated Poisson distribution to the zero-inflated negative binomial distribution and propose ZINBLDA to model overdispersion data with excess zeros. The distribution is expressed as

$$X_{ki_kg} \sim \begin{cases} \delta_{\{0\}}, & p_{ki_kg}, \\ \text{NB}(\mu_{ki_kg}, \phi_g'), & (1 - p_{ki_kg}). \end{cases}$$

Thus, the probability of X_{ki_kg} is written as

$$P(X_{ki_kg}) = \begin{cases} p_{ki_kg} + (1 - p_{ki_kg})\left(\frac{1}{1+\mu_{ki_kg}\phi'_g}\right)^{\phi_g'^{-1}}, & X_{ki_kg} = 0, \\ (1 - p_{ki_kg})\frac{\Gamma(X_{ki_kg}+\phi_g'^{-1})}{X_{ki_kg}!\Gamma(\phi_g'^{-1})}\left(\frac{\mu_{ki_kg}\phi'_g}{1+\mu_{ki_kg}\phi'_g}\right)^{X_{ki_kg}}\left(\frac{1}{1+\mu_{ki_kg}\phi'_g}\right)^{\phi_g'^{-1}}, & X_{ki_kg} > 0. \end{cases}$$

The probability density function of $X_{ki_kg} = x_{ki_kg}$ is

$$P(X_{ki_kg} = x_{ki_kg}) = \Big[(1 - p_{ki_kg})\frac{\Gamma(x_{ki_kg} + \phi_g^{'-1})}{x_{ki_kg}!\Gamma(\phi_g^{'-1})}(\frac{\mu_{ki_kg}\phi'_g}{1 + \mu_{ki_kg}\phi'_g})^{x_{ki_kg}} (\frac{1}{1 + \mu_{ki_kg}\phi'_g})^{\phi_g^{'-1}}\Big]^{I_{(x_{ki_kg}>0)}} \quad (7)$$

$$\Big[p_{ki_kg} + (1 - p_{ki_kg})(\frac{1}{1 + \mu_{ki_kg}\phi'_g})^{\phi_g^{'-1}}\Big]^{I_{(x_{ki_kg}=0)}} \quad (8)$$

By Bayes' rule, we obtain the discriminant score $d_k(\boldsymbol{x}^*)$ of ZINBLDA using

$$\begin{aligned} d_k(\boldsymbol{x}^*) = & \sum_{g=1}^{G} I_{(X_g^*=0)} \log\Big[(1 - \hat{p}_{kg}^*)(\frac{1}{1 + s^*\lambda_g d_{kg}\phi'_g})^{\phi_g^{'-1}} \\ & + \hat{p}_{kg}^*\Big] + \sum_{g=1}^{G} I_{(X_g^*>0)} \log(1 - \hat{p}_{kg}^*) \\ & + \sum_{g=1}^{G} I_{(X_g^*>0)} X_g^*[\log d_{kg} - \log(1 + s^*\lambda_g d_{kg}\phi_g^{'})] \\ & - \sum_{g=1}^{G} I_{(X_g^*>0)}\phi_g^{'-1} \log(1 + s^*\lambda_g d_{kg}\phi_g^{'}) \\ & + \log\pi_k + C. \end{aligned} \quad (9)$$

2.5.2. Parameters Estimation

Next, we estimate the parameters in the ZINBLDA model, which includes the class difference parameter d_{kg}, size factors s_{i_k} and s^*, dispersion parameter ϕ'_g, and the probability of excess zeros p_{ki_kg}.

2.5.2.1. Class Difference Parameter Estimation

Similar to the former three methods, to estimate d_{kg} we first obtain the maximum likelihood estimation $\hat{d}_{kg} = (\sum_{i_k=1}^{n_k} X_{i_kg})/(\sum_{i_k=1}^{n_k} s_{i_k}\lambda_g)$ and then take a Gamma(β, β) prior in case of $\sum_{i_k=1}^{n_k} X_{i_kg} = 0$. Therefore, the posterior mean

$$\hat{d}_{kg} = (\sum_{i_k=1}^{n_k} X_{i_kg} + \beta)/(\sum_{i_k=1}^{n_k} s_{i_k}\lambda_g + \beta)$$

is our estimation. For convenience and due to the small influence of β on the estimation result, we assume $\beta = 1$ in this work.

2.5.2.2. Size Factor Estimation

The total number of reads between samples differs due to various sequencing depths. Generally, data must be normalized by their size factor. The three existing classification methods (PLDA, NBLDA, and ZIPLDA) use three different normalization methods: total count (Dillies et al., 2013), median ratio (Love et al., 2014), and quantile (Bullard et al., 2010). Note that there is little difference in the performance of classification among the three normalization methods. In this work, we use total count to estimate the size factor for convenience. Therefore, the estimation of size factor $\hat{s}_{i_k}$ for the training data is

$$\hat{s}_{i_k} = \frac{\sum_{g=1}^{G} X_{i_kg}}{\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\sum_{g=1}^{G} X_{i_kg}},$$

and the estimation of size factor $\hat{s}^*$ for the testing data is

$$\hat{s}^* = \frac{\sum_{g=1}^{G} X_g^*}{\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\sum_{g=1}^{G} X_{i_kg}}.$$

2.5.2.3. Dispersion Parameter Estimation

Since ZINBLDA assumes that data follow a mixture distribution rather than a negative binomial distribution, the method used to estimate the dispersion parameter in NBLDA is not applicable in this case. Therefore, we used the maximum likelihood to estimate ϕ'_g. Based on equation (8), the log likelihood function of ZINBLDA is

$$\begin{aligned} L = & \sum_{g=1}^{G}\{I_{(x_{ki_kg}=0)} \log[\hat{p}_{ki_kg} + (1 - \hat{p}_{ki_kg}))(\frac{1}{1 + \hat{\mu}_{ki_kg}\phi'_g})^{\phi_g^{'-1}}] \\ & + I_{(x_{ki_kg}>0)}[\log(1 - \hat{p}_{ki_kg}) + \log\Gamma(x_{ki_kg} + \phi_g^{'-1}) \\ & - \log\Gamma(\phi_g^{'-1}) - \log\Gamma(x_{ki_kg}!) \\ & + x_{ki_kg}\log\hat{\mu}_{ki_kg}\phi'_g - x_{ki_kg}\log(1 + \hat{\mu}_{ki_kg}\phi'_g) \\ & - \phi_g^{'-1}\log(1 + \hat{\mu}_{ki_kg}\phi'_g)]\}. \end{aligned} \quad (10)$$

Because the parameter p_{ki_kg} must also be estimated, we cannot directly take the partial derivatives and let the result equal zero to get the estimation of dispersion parameter $\hat{\phi}'_g$ in formula (10). Therefore, we first set an initial value for parameters p_{ki_kg} and ϕ_g, and then we used the PORT routines optimization method (David, 1990) to get the estimation value $\hat{\phi}'_g$.

2.5.2.4. The Probability of Excess Zeros Estimation

Assuming the data for the classifier follow a zero-inflated mixture distribution, we needed to estimate the probability of excess zeros in the distribution. Based on the process proposed by Zhou et al. (2018), we assumed that the probability of zeros, the mean of the genes, and the sequencing depth have the following logistic relation:

$$log\{\frac{P(X_{ki_kg} = 0)}{1 - P(X_{ki_kg} = 0)}\} = \alpha + \beta_1(\frac{N_{ki_k}}{N_{1i_1}}) + \beta_2\mu_{ki_kg}. \quad (11)$$

Replacing $P(X_{ki_kg} = 0)$ in model (11) with $p_{ki_kg} + (1 - p_{ki_kg})(\frac{1}{1+\mu_{ki_kg}\phi_g^{'}})^{\phi_g^{'-1}}$, we get

$$\hat{p}_{ki_kg} = \frac{p_1 - (1 + p_1)(\frac{1}{1+\mu_{ki_kg}\hat{\phi}_g^{'}})^{\hat{\phi}_g^{'-1}}}{(1 + p_1)[1 - (\frac{1}{1+\mu_{ki_kg}\hat{\phi}_g^{'}})^{\hat{\phi}_g^{'-1}}]},$$

where $p_1 = exp\{\alpha + \beta_1(\frac{N_{ki_k}}{N_{1i_1}}) + \beta_2\mu_{ki_kg}\}$; $N_{ki_k} = \sum_{g=1}^{G} X_{ki_kg}$; and α, β_1, and β_2 are coefficients in the logistic model (11).

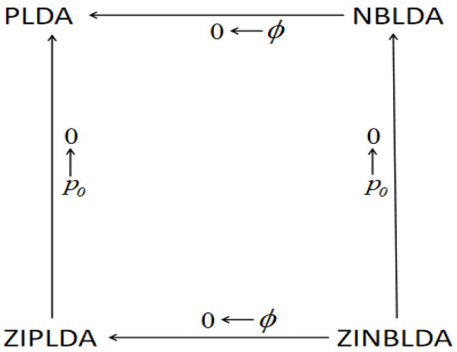


FIGURE 1 | The transformation relation between the four methods. Given a sufficient sample size, the four models can transform into one another according to the value of dispersion parameter ϕ and the average probability of excess zeros p_0.

2.6. Transformation Relation

Note that the above four models can transform into one another under some conditions.

(1) From the discriminant score function of NBLDA (formula 5), we found that if $s^*\lambda_g d_{kg}$ is bounded and $\phi_g \to 0$, then $\log(1 + s^*\lambda_g d_{kg}\phi_g) \to 0$ and $\phi_g^{-1}\log(1 + s^*\lambda_g d_{kg}) \to s^*\lambda_g d_{kg}$. Therefore, the discriminant score of NBLDA approaches that of PLDA (formula 3). That is, the NBLDA classifier reduces to the PLDA classifier when the dispersion value tends to zero.

(2) For the discriminant score function of ZIPLDA (formula 6), when $\hat{p}^*_{kg} \to 0$, then $\log(\hat{p}^*_{kg} + (1 - \hat{p}^*_{kg})e^{-d_{kg}s^*\lambda_g}) \to -d_{kg}s^*\lambda_g$, and the discriminant score of ZIPLDA approaches that of PLDA. Thus, with the probability of zeros decreased to zero, the ZIPLDA score reduces to the PLDA score.

(3) Similarly, for the discriminant score of ZINBLDA (formula 9), when $\phi'_g \to 0$, then $\phi_g^{'-1}\log(1 + s^*\lambda_g d_{kg}\phi'_g) \to d_{kg}s^*\lambda_g$ and $(1 + s^*\lambda_g d_{kg}\phi'_g)^{-\phi_g^{'-1}} \to exp\{-d_{kg}s^*\lambda_g\}$. That is, when dispersion tends to zero, the discriminant score of ZINBLDA reduces to that of the ZIPLDA. Furthermore, if $\hat{p}^*_{kg} \to 0$, then $\log[(1 - \hat{p}^*_{kg})(\frac{1}{1+s^*\lambda_g d_{kg}\phi'_g})^{\phi_g^{'-1}} + \hat{p}^*_{kg}] \to -\phi_g^{'-1}\log(1 + s^*\lambda_g d_{kg}\phi'_g)$. Therefore, when the probability of $\delta_{\{0\}}$ tends to zero, the ZINBLDA classifier reduces to the NBLDA classifier.

Figure 1 shows the above transformation relations, where ϕ denotes the dispersion parameter, and p_0 denotes the average probability of excess zeros. Starting at the bottom right of the figure and going clockwise, ZINBLDA reduces to ZIPLDA as $\phi \to 0$, and ZIPLDA reduces to PLDA as $p_0 \to 0$. Likewise, starting at the bottom right corner and going counterclockwise, ZINBLDA reduces to NBLDA as $p_0 \to 0$, and NBLDA reduces to PLDA as $\phi \to 0$. The transformation relationship between the four classification methods indicates that for data without dispersed biological replicates and excess zeros, PLDA may perform better than the other methods. However, NBLDA is good at dealing with overdispersion data, while ZIPLDA is designed to handle data with excess zeros. For data with excess zeros and dispersed biological replicates, ZINBLDA may be the optimal choice.

3. SIMULATION STUDIES

We evaluated the performance of the four methods by conducting simulations in various scenarios. We also built a decision tree to help us select the optimal classifier from the four methods for a new dataset.

3.1. Simulation Design

To ensure a fair comparison between the four classifiers, we followed the same process as Zhou et al. (2018) and generated simulation data from the following negative binomial distribution:

$$X_{ki_kg} \sim \mathrm{NB}(d_{kg}s_{i_k}\lambda_g, \phi).$$

We set $K = 2$ to illustrate the binary classification, and each class included about $n/2$ samples. We also considered multiple classifications with $K = 3$, with each class including approximately $n/3$ samples. The rest of the distribution parameters were as follows: the size factors s_{i_k} had a uniform distribution at [0.2, 2.2], the λ_g values had an exponential distribution with an expectation of 25, and the $\log d_{kg}$ values had a normal distribution with a location of 0 and scale of σ^2 (where $\sigma = 0.2$). In the simulation studies, the *DE rate* represented the proportion of differentially expressed genes, and p was the number of genes in the samples. For simplicity, we denoted p_0 as the probability of excess zeros. In each simulation study, we changed one parameter and fixed the others, then compared the misclassification rates of the four classifiers. We specified the values for p, *DE rate*, p_0, ϕ, and n in each simulation study. Each simulation was repeated 1,000 times, and the average misclassification rates were calculated for the four methods.

3.2. Simulation Results

Study 1 investigated the impact of the dispersion parameter on the performance of the four classification methods. Considering a binary classification, we set the probability of excess zeros of data to 0 and generated 50 training and 50 testing samples. Each sample included 100 genes, 20% of which were DE genes. **Figure 2** shows the average misclassification rates of the four methods with different dispersions. Overall, the misclassification rates of the four classifiers decreased when the dispersion parameters changed from 1 to 0. PLDA and ZIPLDA showed similar performance, and both were slightly worse than NBLDA and ZINBLDA in different dispersion settings. However, when the dispersion was reduced to zero, the misclassification rates of all four methods tended to zero. From the expressions of the negative binomial and Poisson distributions, the former reduced to the latter when the dispersion parameter was reduced to zero, which indicates that NBLDA and ZINBLDA (which are based on negative binomial distribution) are more suitable for classifying overdispersion data. In addition, we changed the probability of excess zeros of simulation data from 0 to 0.1, and the other parameters remained the same. **Supplementary Figure 1**

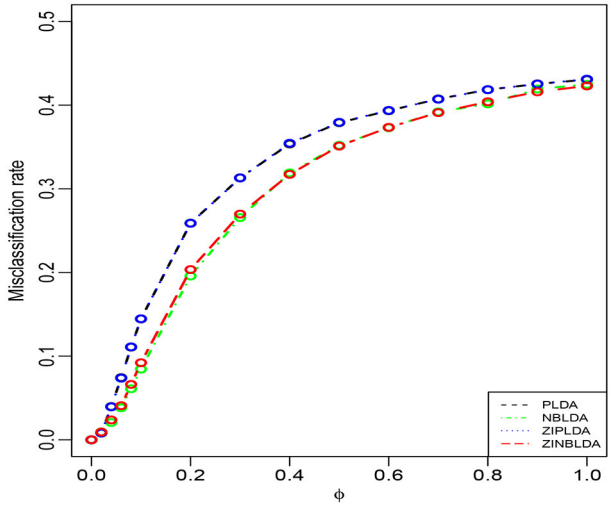


FIGURE 2 | The misclassification rates of the four methods with different dispersions (Study 1). Here, $\sigma = 0.2$, $p = 100$, $n = 50$, $p_0 = 0$, and *DE rate* = 0.2.

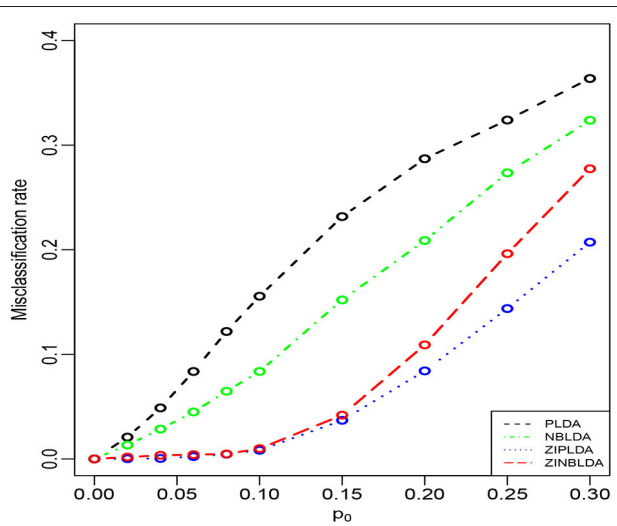


FIGURE 3 | The misclassification rates of the four methods with different probabilities of excess zeros (Study 2). Here, $\sigma = 0.2$, $p = 100$, $n = 50$, $\phi = 0$, and *DE rate* = 0.2.

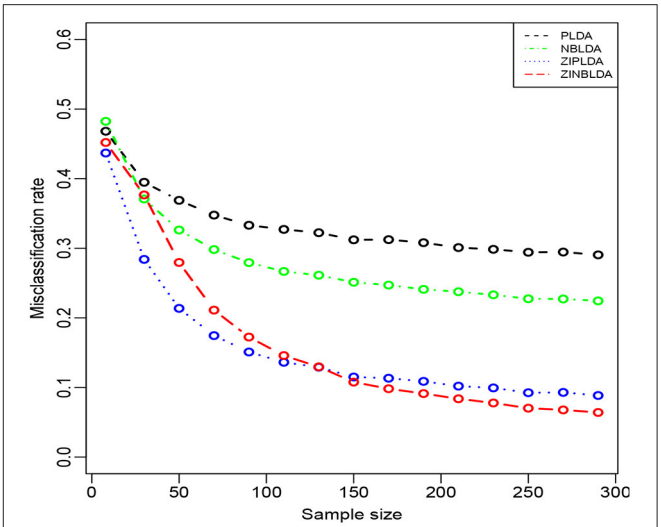


FIGURE 4 | The misclassification rates of the four methods with different sample sizes (Study 3). Here, $\sigma = 0.2$, $p = 100$, $\phi = 0$, $p_0 = 0.3$, and *DE rate* = 0.2.

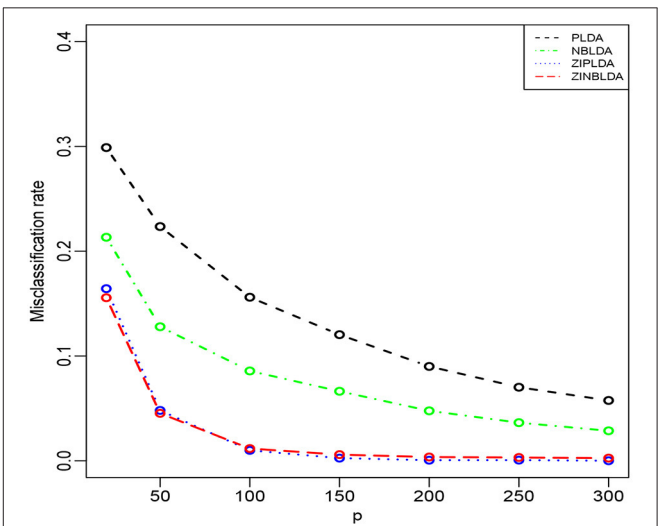


FIGURE 5 | The misclassification rates of the four methods with different numbers of genes (Study 4). Here, $\sigma = 0.2$, $n = 50$, $\phi = 0$, $p_0 = 0.1$, and *DE rate* = 0.2.

shows that when dispersion changed the value from 0 to 1, ZINBLDA outperformed the other methods. However, NBLDA and PLDA performed worse than ZINBLDA and ZIPLDA when the dispersion tended to zero. This result indicates that the probability of excess zeros has a major effect on the performance of the four methods.

Study 2 investigated the performance of the four methods with different probabilities of excess zeros. In this study, we set the dispersion parameter to 0, and $p_0 \in [0.1, 0.3]$; the rest of the parameters were the same as in Study 1. **Figure 3** shows that the average misclassification rates of the four classifiers increased as the probability of excess zeros increased. ZIPLDA performed slightly better than ZINBLDA when p_0 tended to 0. The performance of these two classifiers was far better than the other two classifiers, and PLDA performed the worst with different probabilities of excess zeros. This result demonstrates that ZIPLDA and ZINBLDA (which are designed for excess zeros) have a clear advantage over the other two methods when classifying data with excess zeros. Setting $\phi = 0$ could explain why ZIPLDA performed slightly better than ZINBLDA. In addition, when we reduced the sample size from 50 to 8, the result (**Supplementary Figure 2**) showed that ZIPLDA still

performed the best out of the four classifiers; however, ZINBLDA performed the worst in this case, which indicates that the sample size has a major effect on the performance of ZINBLDA.

Figure 4 shows the performance of the four classification methods when the sample size changes. In Study 3, we set the probability of excess zeros to 0.3, and the sample size was gradually changed from 8 to 300. The rest of the parameters were the same as in Study 2. The overall misclassification rates gradually declined to nearly a constant value for all four classifiers when the sample size increased. ZIPLDA showed superiority over the other methods when the sample size was less than 130, and ZINBLDA attained a lower misclassification rate when the sample size was over 150. The same pattern existed between NBLDA and PLDA. When the sample size was less than 20, PLDA

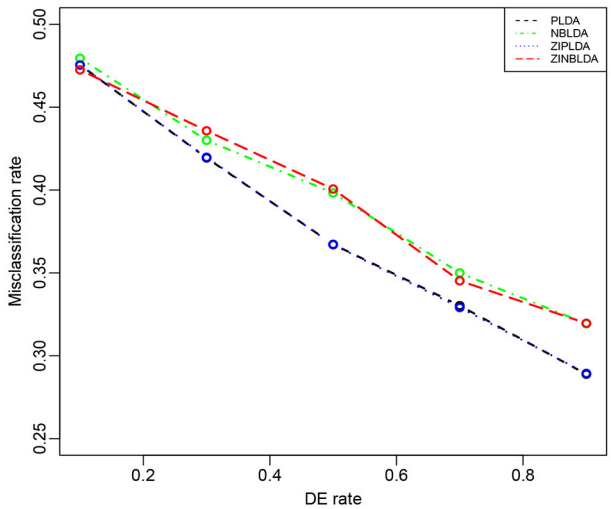


FIGURE 6 | The misclassification rates of the four methods with different probabilities of differential expression genes (Study 5). Here, $\sigma = 0.2$, $p = 100$, $n = 8$, $\phi = 0.5$, and $p_0 = 0$.

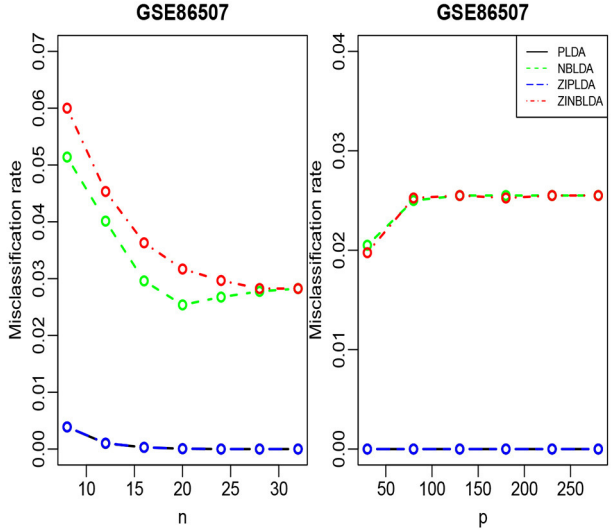


FIGURE 8 | The misclassification rates of the four classifiers for the GSE86507 dataset.

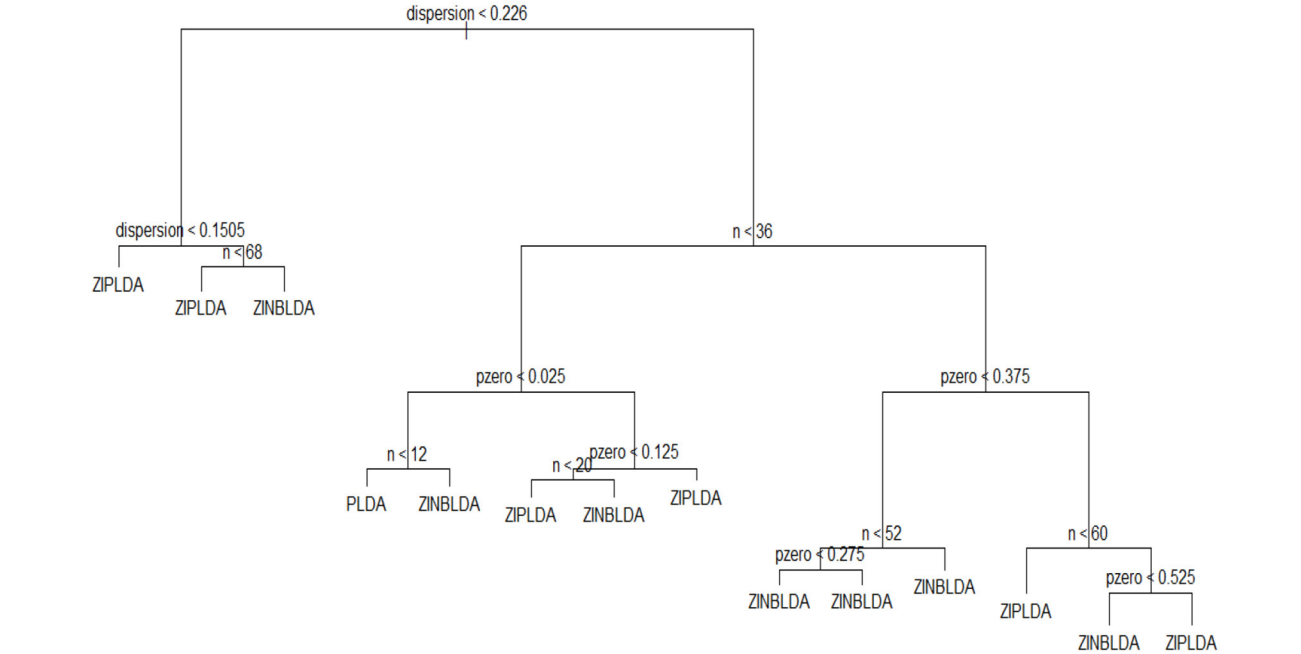


FIGURE 7 | The decision tree model used to choose an optimal classifier.

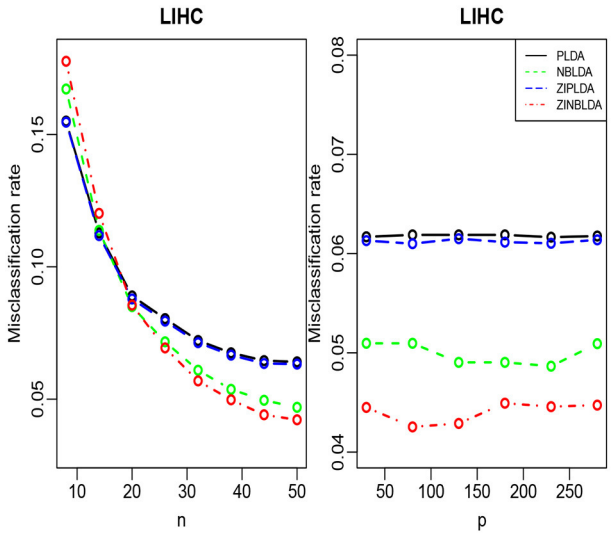


FIGURE 9 | The misclassification rates of the four classifiers for the LIHC dataset.

had a lower misclassification rate; however, NBLDA yielded a lower value when the sample size increased. The results illustrate that sample size has a huge impact on the performance of ZINBLDA and NBLDA, and ZINBLDA outperformed the other methods when a sufficient sample size was available. The reason for this may be that ZINBLDA requires a minimal number of samples to estimate the parameters in the model.

In the above three studies, we fixed the gene number at 100. In Study 4, we changed the number of selected genes and evaluated the performance of the four classifiers. The parameters were the same as in the former studies except $\phi = 0$ and $p_0 = 0.1$. **Figure 5** shows that the misclassification rates of the four methods declined as the number of genes selected increased. ZINBLDA and ZIPLDA showed similar performance and outperformed the other two methods, and PLDA performed the worst of the four methods. In **Supplementary Figure 3**, we changed the data dispersion from 0 to 0.2. A lower misclassification rate was obtained by ZINBLDA and NBLDA, and PLDA again performed the worst of the four methods. This result agrees with the conclusion that dispersion affects the performance of PLDA and ZIPLDA.

Study 5 investigated the influence of the probability of differential expression in the selected genes on the performance of the four classifiers. In this study, we set the dispersion parameter to 0.5, the probability of excess zeros was set at 0, and 100 genes were selected for all eight samples. **Figure 6** shows that the overall misclassification rates of the four methods decreased as the DE rate increased. PLDA and ZIPLDA showed similar performances, and both performed better than NBLDA and ZINBLDA with different DE rates. ZINBLDA and NBLDA performed nearly same with different probabilities of DE genes. This result demonstrates that the sample size has a marked impact on the performance of the four classifiers. In addition, **Supplementary Figure 4** shows that when the dispersion was set to 0.2 and the probability of zeros to 0.1, ZIPLDA performed remarkably better than the other methods with an increasing number of DE rates, followed by PLDA and then NBLDA, ZINBLDA performed the worst. This indicates that excess zeros in the data enable ZIPLDA to perform better than PLDA, and the sample size affects the performance of ZIPLDA notably.

We also conducted Simulation Studies 1–5 using multiple classifications ($K = 3$). **Supplementary Figures 5–9** show the performance of the four methods. The parameters were the same as in Studies 1–5 except for the sample sizes. We set $n = 75$ in **Supplementary Figures 5**, **6**, **8**; $n = 12$ in **Supplementary Figure 9**; and $n = 450$ in **Supplementary Figure 7** and compared the results with those of Studies 1–5. The performance of the four classifiers remained the same as in the binary classification.

In the simulation studies conducted above, the performance of the four classifiers was related to the attributes of the dataset, including sample size n, dispersion parameter ϕ, and the probability of excess zeros p_0. In the final simulation study, we considered a binary classification with three changeable parameters and compared the performance of the four methods for different combinations of those three parameters. We still selected 100 genes, 40% of which were DE genes. The probability of excess zeros was set at 0.001, 0.1, or 0.3, and the sample size was 8, 50, or 100. The dispersion parameters changed from 0.001 to 0.1 to 1 with every 0.2 steps. The average misclassification rates of the four methods are shown in **Supplementary Figure 10**. To clarify display the result, **Supplementary Table 1** shows the concrete values of each misclassification rate. Comparing the results of the three panels in each column, we found that for the first column (sample size of 8), the overall misclassification rates of the four methods increased when the probability of excess zeros increased from 0.001 to 0.3. When the probability of excess zeros was equal to 0.3, the misclassification rates approached 50%. When $p_0 = 0.001$, the performance of ZIPLDA and PLDA was better than NBLDA and ZINBLDA. However, when $p_0 = 0.1$, ZIPLDA outperformed the other methods, which indicates that ZIPLDA is more suitable for handling data with a small probability of excess zeros, and the sample size has less of an impact on it. When the sample size was increased to 50 (the second column), the overall performance of ZINBLDA was slightly better than that of ZIPLDA except when ϕ was small and $p_0 = 0.3$. The reason for this may be that there were not enough samples to estimate the parameters of ZINBLDA. Therefore, when the sample size increased to 100 (the third column), that ZINBLDA yielded a lower or equal misclassification rate compared to the other methods, which indicates that ZINBLDA can achieve the best classification result as long as enough samples are available. The performance of ZIPLDA also improved when p_0 increased from 0.1 to 0.3 due to the increase in the probability of excess zeros.

3.3. Optimal Classifier Selection

To select an optimal classification method for different datasets, we built a decision tree and a random forest. A decision tree is a machine learning algorithm that is widely used in many

scenarios because of its accuracy for the current algorithms. As its name implies, a decision tree is a decision support tool that uses a tree-like model. It is comprised of nodes and branches, and each sample is tested on an internal node. The outcome of the test determines which branch is followed, and this procedure continues until the leaf node that holds the class label of this sample is reached. Random forest is an ensemble of a decision tree, and it can achieve a more stable result than a decision tree.

To employ a tree-like model to select the optimal classifier, the chosen features were the sample number n, dispersion ϕ, and probability of excess zeros p_0, and y was regarded as the optimal classification method. The parameter region was divided to assign the value of the feature vector. The values of sample size n ranged from 8 to 100 with a step size of 8. The dispersion parameter ranged from 0.001 to 1.001 at intervals of 0.1. The probability of excess zeros p_0 ranged from 0 to 0.6 with a step size of 0.05. For each calculation, we took one value from each parameter set to generate the simulation data, allowing for multiple combinations of these three parameters. This procedure was repeated 1,000 times, and the classifier corresponding to the smallest value of the average misclassification rate was regarded as the optimal classification method. We used the obtained data to train a decision tree, and **Figure 7** displays the classification result. This model fits the data very well, with a misclassification rate of only 7.4%. To use this model, we only need to know or estimate the values of the three parameters, then use the conditional control statements in the decision tree to distinguish in each internal node, which will result in the optimal method when the leaf node is reached. In this way, this model can be used to help choose the optimal classification method. Similarly, we can obtain a random forest with a lower misclassification rate (2.2%). The classification results of decision tree and random forest are saved in R scripts, which could be used to choose the optimal classifier when inputting the parameters of dataset.

4. APPLICATION TO REAL DATA

We further compared the four methods by analyzing two real datasets: GSE86507 and TCGA-LIHC (Liver Hepatocellular Carcinoma). The details of these two RNA-seq datasets are as follows.

Woo et al. (2017) created the GSE86507 dataset to compare gene expression between two mouse models, Pkd1f/f: HoxB7-cre mice and Pkd2f/f: HoxB7-cre mice. Each group includes 18 samples, and there are a total of 29,996 transcripts in this dataset. It contains about 17.74% zeros of all numerical values.

The dataset TCGA-LIHC contains two groups of samples: the normal group (340 samples) and the cancerous group (50 samples). There are 60,487 genes in this dataset, which contains about 43.24% zeros of all numerical values.

We chose to classify parts of genes since the majority of genes in a dataset are not differentially expressed and thus do not contribute to the sample classification. Including entire genes in the model would reduce the classification accuracy and increase the computational complexity. Thus, selecting parts of genes not only improves the accuracy of classification but saves computation time. Following the steps outlined by Dudoit et al. (2002), we selected genes by first calculating the ratio of the sum of the squares between groups and within groups for each gene, then sorted all of the genes according to the ratio from greatest to least, and finally selected a certain number of genes for downstream analysis.

We randomly split the data into a training set and test set, with both datasets containing all classes. We selected the 300 most differentially expressed genes to train the model. This procedure was repeated 1,000 times, and the average misclassification rates for each method were recorded. The left panels of **Figures 8**, **9** show that for the test data, the average misclassification rates of the four methods decreased as the number of training data gradually increased. For the GSE86507 dataset, the misclassification rates of PLDA and ZIPLDA were lower than NBLDA and ZINBLDA, both of which were close to zero. However, for the TCGA-LIHC dataset, PLDA and ZIPLDA were superior to NBLDA and ZINBLDA when the sample size was small. As the training sample size increased, the misclassification rates of NBLDA and ZINBLDA decreased remarkably, and ZINBLDA outperformed the other three methods for a large sample size. We also evaluated the classification performance of the four methods by fixing 30 training sets and gradually increasing the number of selected DE genes. The right panel of **Figure 8** shows that PLDA and ZIPLDA outperformed the other two methods, whereas the right panel of **Figure 9** shows the superiority of ZINBLDA over the other methods in this case.

To assess the efficiency of the decision tree model, we estimated the dispersion and probability of excess zeros for the two datasets. The estimated dispersion of GSE86507 was $\phi = 0.12$, and the probability of excess zeros was 0.5%, which indicates that the dataset has slight overdispersion and almost no excess zeros. The estimated dispersion of TCGA-LIHC was $\phi = 1.12$, and the probability of excess zeros was 8%, which indicates that the dataset has high overdispersion and many excess zeros. According to the conclusions in section 2.6, PLDA should perform better with the GSE86507 dataset, and ZINBLDA should be the optimal method to classify the TCGA-LIHC dataset. We used the estimated parameters to select the optimal method according to the conditional control statements in the decision tree model (**Figure 7**). Based on the result, we recommend selecting ZIPLDA for the GSE86507 dataset and ZINBLDA for the TCGA-LIHC dataset, which coincides with the real analysis results.

5. DISCUSSION

RNA-seq data classification is vital to the diagnosis of diseases. In this work, we extended the existing classification methods and proposed a ZINBLDA method for overdispersion RNA-seq data with an excess of zeros. Concretely, we built a mixture distribution with a point mass at zero and a negative distribution to model the data, and a logistic regression was used to build a relation between the probability of zeros, the mean of the genes, and the sequencing depth. Most importantly, we examined four

classification methods from the perspective of their parameters, and we found that these four methods can transform into each other in some cases.

In the simulation studies, we evaluated the performance of the four methods in a wide range of settings. The simulation results showed that different methods perform better for different applications. In addition, we found that the application region of each method is associated with the attributes of the dataset, such as the dispersion, sample size, and probability of excess zeros. Therefore, we built a decision tree to help us select the optimal classification methods in different cases. In the real data analysis, we analyzed two real, next-generation sequencing datasets, and the results further confirmed the theory and simulation conclusions.

Although each of the four methods performed well in certain scenarios, there are numerous issues that remain to be solved, such as single cell RNA-seq data being particularly prone to dropout events due to the relatively shallow sequencing depth per cell. In this case, the existing classification methods may not provide a good result in practice. Therefore, we plan to develop a new classification method that employs deep learning technology to model scRNA-seq data to further improve our current work.

DATA AVAILABILITY STATEMENT

The dataset GSE86507 for this study can be found in the NCBI repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86507), and the dataset TCGA-LIHC for this study can be found in the GDC repository (https://portal.gdc.cancer.gov/projects/TCGA-LIHC).

AUTHOR CONTRIBUTIONS

YZ conceived the idea. JZ processed the data and conducted simulation and real dataset experiments. JZ, ZY, and YZ wrote the manuscript. LS, MZ, YZ, and WL revised the manuscript. All authors read and approved the final manuscript.

FUNDING

YZ's research was supported by the National Natural Science Foundation of China (Grant Nos. 12071305, 11871390, and 11871411), Natural Science Foundation of Guangdong Province of China under grant 2020B1515310008, Project of Educational Commission of Guangdong Province of China under grant 2019KZDZX1007. MZ's research was supported by the Social Science Foundation of China (Grant No. 15CTJ008). This paper is partially supported by the National Natural Science Foundation of China (No. 11901401).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.642227/full#supplementary-material

REFERENCES

Berger, M., Levin, J., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., et al. (2010). Integrative analysis of the melanoma transcriptome. *Genome Res.* 20, 413–427. doi: 10.1101/gr.103697.109

Biesecker, L., Burke, W., Kohane, I., Plon, S., and Zimmern, R. (2012). Next-generation sequencing in the clinic: are we ready? *Nat. Rev. Genet*. 13, 818–824. doi: 10.1038/nrg3357

Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

David, M. G. (1990). *Usage Summary for Selected Optimization Routines*. Computing Science Technical Report 153. AT T Bell Laboratories, Murray Hill, NY.

Dillies, M., Rau, A., Aubert, J., Christelle, H., Marine, J., Nicolas, S., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform*. 14, 671–683. doi: 10.1093/bib/bbs046

Dong, K., Zhao, H., Tong, T., and Wan, X. (2016). NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. *BMC Bioinformatics* 17:369. doi: 10.1186/s12859-016-1208-1

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*. 97:77–87. doi: 10.1198/016214502753479248

Lin, B. Q., Zhang, L. F., and Chen, X. (2014). LFCseq: a nonparametric approach for differential expression analysis of RNA-seq data. *BMC Genomics* 15:S7. doi: 10.1186/1471-2164-15-S10-S7

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

Morozova, O., Hirst, M., and Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet*. 10, 135–151. doi: 10.1146/annurev-genom-082908-145957

Nagalakshmi, U., Zhong, W., Karl, W., Chong, S., Debasish, R., Mark, G., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349. doi: 10.1126/science.1158441

Robinson, M., McCarthy, D., and Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Robinson, M., and Smyth, G. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 321–332. doi: 10.1093/biostatistics/kxm030

Tan, K., Petersen, A., and Witten, D. (2014). "Classification of RNA-seq data," in *Statistical Analysis of Next Generation Sequencing Data*, eds S. Datta and D. Nettleton (New York, NY: Springer), 219–246. doi: 10.1007/978-3-319-07212-8_11

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van, B. M., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Witten, D. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.* 5, 2493–2518. doi: 10.1214/11-AOAS493

Woo, Y., Kim, D., Koo, N., Kim, Y., Lee, S., Ko, J., et al. (2017). Profiling of mirnas and target genes related to cystogenesis in adpkd mouse models. *Sci. Rep*. 7:14151. doi: 10.1038/s41598-017-14083-8

Zhou, Y., Wan, X., Zhang, B., and Tong, T. (2018). Classifying next-generation sequencing data using a zero-inflated Poisson model. *Bioinformatics* 34, 1329–1335. doi: 10.1093/bioinformatics/btx768

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 March 2021
doi: 10.3389/fgene.2021.649376

Causal Linkage Between Inflammatory Bowel Disease and Primary Sclerosing Cholangitis: A Two-Sample Mendelian Randomization Analysis

Ying Xie[1†], Xuejie Chen[2†], Minzi Deng[2†], Yuhao Sun[1], Xiaoyan Wang[2], Jie Chen[1*], Changzheng Yuan[3] and Therese Hesketh[1,4]*

[1] Centre for Global Health, Zhejiang University School of Medicine, Hangzhou, China, [2] Department of Gastroenterology, The Third Xiangya Hospital, Central South University, Changsha, China, [3] Department of Big Data and Health Science, Zhejiang University School of Medicine, Hangzhou, China, [4] Institute for Global Health, University College London, London, United Kingdom

Background: Observational studies suggest an association between inflammatory bowel disease (IBD) [including ulcerative colitis (UC) and Crohn's disease (CD)] and Primary sclerosing cholangitis (PSC), but the causal association between the two diseases remains unclear.

Methods: We used two-sample Mendelian randomization (MR) to estimate the causal association between IBD and PSC. We chose single nucleotide polymorphisms (SNPs) data for analysis, obtained from previous genome-wide association studies (GWASs). Pleiotropy, heterogeneity, and sensitivity analyses were performed for quality control.

Results: We found that the causal associations between IBD (both UC and CD) and PSC were significant (e.g., IBD and PSC, Robust adjusted profile score (RAPS) OR = 1.29, 95% CI 1.16∼1.44, $p < 0.01$; UC and PSC, RAPS OR = 1.40, 95% CI 1.23∼1.58, $p < 0.01$; CD and PSC, RAPS OR = 1.13, 95% CI 1.02∼1.26, $p = 0.02$). MR Egger, IVW, and ML tests found statistical heterogeneity between determined IV estimates. The leave-one-out analysis also indicated the sensitivity of the SNPs (e.g., IBD and PSC, MR-Egger Q = 644.30, $p < 0.01$; UC and PSC, MR-Egger Q = 378.30, $p < 0.01$; UC and PSC, MR-Egger Q = 538.50, $p < 0.01$).

Conclusion: MR analyses support the positive causal effect of IBD (including UC and CD) on PSC in a European population. We provide suggestions for preventing and treating the two diseases.

Keywords: inflammatory bowel disease, ulcerative colitis, Crohn's disease, mendelian randomization, primary sclerosing cholangitis

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong, Hong Kong

Reviewed by:
Ming Ding,
Harvard University, United States
Tao Huang,
Peking University, China

****Correspondence:***
Xiaoyan Wang
wxy220011@163.com;
wangxiaoyan@csu.edu.cn
Jie Chen
med_chenjie@zju.edu.cn

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 January 2021*
Accepted: *23 February 2021*
Published: *18 March 2021*

Citation:
Xie Y, Chen X, Deng M, Sun Y, Wang X, Chen J, Yuan C and Hesketh T (2021) Causal Linkage Between Inflammatory Bowel Disease and Primary Sclerosing Cholangitis: A Two-Sample Mendelian Randomization Analysis. Front. Genet. 12:649376. doi: 10.3389/fgene.2021.649376

INTRODUCTION

Primary sclerosing cholangitis (PSC) is a rare, progressive cholestatic disease featuring impaired bile formation and chronic liver dysfunction, led by inflammation and fibrosis with a 0.77 per 100,000 person-years incidence rate (Molodecky et al., 2011; Karlsen et al., 2017; Dyson et al., 2018). Both genetic and environmental factors contribute to PSC, with the intestinal microbiome being considered as a pathogenetic factor. Inflammatory bowel diseases (IBDs) describe a series of chronic inflammatory disorders of the gastrointestinal tract including two main types: Crohn's disease (CD) and ulcerative colitis (UC) (Rosen et al., 2015; Hodson, 2016).

It has been reported that IBD and PSC are closely associated. According to a comprehensive review, the prevalence of IBD in PSC has reached two-thirds (Karlsen et al., 2017). It has been observed that total colectomy can reduce the recrudesce risk of PSC by 50%, prior to or within liver transplantation (Lindström et al., 2018; Ricciuto et al., 2018). It has also been reported that the inflammatory type of PSC differs from UC or CD (Karlsen et al., 2017). Genetically, PSC appears to be more like an autoimmune condition compared with IBD (Liu et al., 2013). Although the striking association has been found for more than 50 years (Warren et al., 1966), the mechanisms for the relationships between the two diseases remain elusive.

The causal relationship between IBD and PSC is important in exploring the function of the disease, and thus in informing evidence for effective treatment. Randomized controlled trials (RCTs) are the most reliable method for determining causal inference in treatment studies. However, due to the requirements of the design and implementation, difficultly to control, and the consideration of ethics, RCTs are difficult to conduct. We used the Mendelian randomization (MR) analysis to explore the likely causal relevance between exposure and outcome, based on observational epidemiological studies.

Because gametes follow Mendelian rules of inheritance (parental alleles are randomly assigned to offspring), genetic variation is not affected by confounders such as environmental exposure, socioeconomic status, and behavior. Furthermore, genetic variation comes from parents, thus the association with outcomes is chronological. Therefore, MR can overcome the problems of confounding and reverse causality.

The instrumental variables (IVs) in the MR study rely on three core assumptions: (a) the genetic variant (either combined or isolated with other variants) is associated with the exposure; (b) the genetic variant is not associated with confounders that are either known or unknown; (c) there is no pathway from the genetic variant to outcomes that do not include the exposure. In MR research, it is difficult to obtain an accurate estimate of causal association without any one of the above assumptions.

Genome-wide association studies (GWAS) featuring large sample sizes make genetic variants available. Based on the previous GWAS, we selected single nucleotide polymorphisms (SNPs) that are strongly relevant to IBD (including UC and CD) as IVs. The effect of IVs on the exposures (IBD) and outcomes (PSC) was from two independent samples. We used two-sample MR and statistical methods to analyze the quantitative effects of IBD (UC, CD) on PSC.

MATERIALS AND METHODS

Data Source

More SNP sites related to IBD were screened out by GWAS results combined with literature reports. This study chose SNPs from publicly available GWAS data bases associated with exposures (IBD, including UC and CD). IBD-associated SNPs were derived from a GWAS meta-analysis study of IBD in the European Genome-phenome Archive (EGA). The statistics came from an extended cohort of 86,640 European individuals and 9,846 non-Europeans (Liu et al., 2015). Studies showed that most of the risk loci were shared across divergent populations (Teslovich et al., 2010; Okada et al., 2014; Liu et al., 2015). The SNPs associated with PSC were selected from the largest GWAS of PSC up to date, the European population, including 4,796 cases and 19,955 population controls. Quality control, like the Inverse variance weighted (IVW) fixed-effects meta-analysis, was performed to test the evidence of association across the GWAS and cohorts. Bayesian tests were conducted in both studies to identify loci with strong evidence.

SNP Selection

From the collection of SNPs in previous studies, we then set some standards for including eligible SNPs. We chose SNPs that were significantly associated with exposures ($p \leq$ 5e-8) and that had a certain probability of mutation (Minor allele frequency, MAF $\geq$ 5%), without reported loci coincidence or linkage disequilibrium (LD) (R^2 <0.001). The palindromic SNPs which can introduce ambiguity into the identity of the effect allele in the exposure GWAS were also excluded. The SNPs that were both related to PSC and IBD were the excluded to meet the third core assumption, eliminating other pathways from the genetic variant to outcomes that do not include the exposure.

Effect Size Estimate

We estimated the causal association between exposures (IBD, UC and CD) and outcomes (PSC) with Inverse variance weighted (IVW), MR Egger, Weighted median (WM), Robust adjusted profile score (MR. RAPS), and Maximum likelihood (ML). IVW takes the inverse variance of each study as the weight to calculate the weighted average of effect sizes, to summarize the effect sizes of multiple independent studies (Lee et al., 2016). We also performed the Weighted median estimator (WME), with which causal effects can be accurately estimated with more than 50% weight using IVs when doing the analysis (Bowden et al., 2016a). A newly developed analysis called Robust adjusted profile score (MR. RAPS) considering the measurement error in SNP-exposure effects was conducted to reduce bias from weak IVs (Zhao et al., 2019). Maximum likelihood maximizing the likelihood function to estimate the probability distribution parameters was also used as a reference traditional method (Milligan, 2003). However, due to potential pleiotropic effects, the genetic variants may influence the outcome in an additional

way, thus causing a bias. Therefore, we used MR-Egger regression as well, the slope of which can estimate the magnitude of directional pleiotropy. The MR-Steiger directionality test is used to test the causal direction between the hypothesized exposure and outcomes as a verification of the reliability of the results (Hemani et al., 2017). The results were presented in odds ratios (OR) and 95% confidence intervals (CI). A two-sided p-value was considered statistically significant when it was less than 0.05 (**Figure 1**). All statistical analyses were performed in R 3.4.2 with the package "TwoSampleMR."

Sensitivity Analyses

In the MR analysis, it is necessary to consider whether SNPs, as instrumental variables, associate with other exposures. We used MR-Egger to test pleiotropy, verifying whether a single locus affects multiple phenotypes. The "Leave-one-out sensitivity analysis" algorithm is used as sensitivity analyses. With non-specific SNPs eliminated, if the correlation between other instrumental variables and outcomes was still statistically significant, it indicates more sufficient evidence of the causal association between exposure and outcomes. By removing SNPs one by one, the results are reanalyzed to draw the forest map with a stable result intuitively judged (**Supplementary Figures S1–S3**). As for the heterogeneity analysis, we conducted it for MR Egger, Inverse variance weighted, and Maximum likelihood. Heterogeneity was standardized with Cochran Q statistics; the weighted sum of the squared differences between the effect of each SNP and the summed effect of all SNPs. We also used a two-sided p-value and considered statistical significance at $p < 0.05$.

RESULTS

Based on the selection criteria above, we conducted linkage disequilibrium tests to choose SNPs that are both related to IBD and PSC. A total of 121, 76, and 104 SNPs were selected as IVs for IBD, UC, and CD, respectively. We then excluded 17 palindromic SNPs (nine for IBD, three for UC, and five for CD). Finally, we included 112, 73 m and 99 SNPs for IBD, UC, and CD, respectively (**Supplementary Tables S4–S6**).

The causal associations between IBD (UC, CD) and PSC were not accordant among the five methods. The RAPS indicated that IBD (both UC and CD) was significantly associated with PSC (IBD and PSC, RAPS OR = 1.29, 95% CI 1.16∼1.44, $p < 0.01$; UC and PSC, RAPS OR = 1.40, 95% CI 1.23∼1.58, $p < 0.01$; CD and PSC RAPS OR = 1.13, 95% CI 1.02∼1.26, $p = 0.02$) (**Table 1**). However, using MR-Egger, none were significantly associated with PSC (For IBD, OR = 1.16, 95% CI 0.82∼1.63, $p = 0.41$; For UC, OR = 1.13, 95% CI 0.80∼1.61, $p = 0.50$; For CD, OR = 1.06, 95% CI 0.75∼1.50, $p = 0.74$) (**Table 1**). When all genetic variants are valid, the causal effect may be underestimated due to the inflated type I error (Bowden et al., 2016b). Additionally, ML, WM showed the significant associations between UC, CD, and PSC while IVW did not reveal the associations between CD and PSC (**Table 1**). Based on the above five analyses, we concluded that the causal association between IBD and PSC were significant.

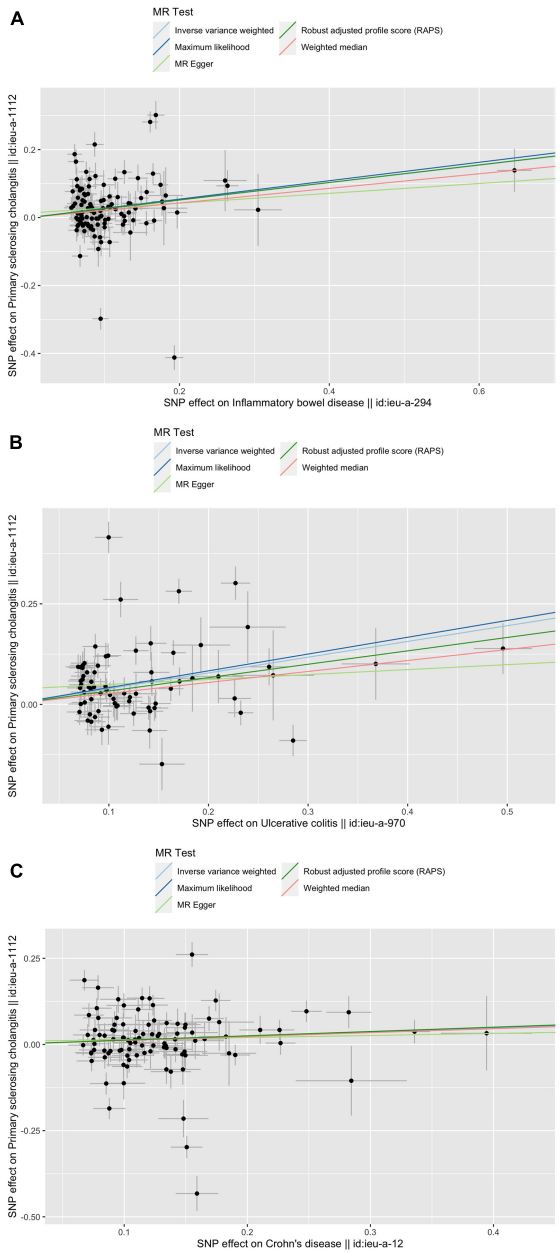


FIGURE 1 | Scatter plots of the genetic causal associations with IBD, UC, and CD against PSC using different MR methods. **(A)** IBD against PSC; **(B)** UC against PSC; and **(C)** CD against PSC. The slopes of the line represent the causal association for different methods. The light green line represents the MR Egger, the dark green line represents the Robust adjusted profile score (RAPS), the pink line represents the weighted median estimate, the light blue line represents the Inverse variance weighted estimate, and the dark blue line represents the Maximum likelihood estimate.

Pleiotropy, heterogeneity, and sensitivity analyses were performed for quality control. We used MR-Egger regression to test the pleiotropy, finding an unlikely bias caused by horizontal pleiotropy (IBD $p = 0.48$, UC $p = 1.03$, CD $p = 0.72$) (**Table 2**). To test the heterogeneity, we conducted MR Egger,IVW, and

ML, finding statistical heterogeneity between determined IV estimates(e.g., For IBD, MR-Egger Q = 644.30, $p < 0.01$; For UC, MR-Egger Q = 378.3, $p < 0.01$; For UC, MR-Egger Q = 538.50, $p < 0.01$.)(**Table 2**). For sensitivity, we conducted a Leave-one-out sensitivity analysis, finding that the MR estimates were reasonable considering the effect of single SNPs (**Supplementary Figures S1–S3**). Additionally, the MR-Steiger test supported a positive causal correlation between IBD (UC, CD) and PSC, also identifying IVs' affecting susceptibility to IBD traits and PSC. These results indicate the powerful relevance of MR assumption and the weak bias in the analysis.

DISCUSSION

To our knowledge, it is the first study to illustrate the causal relationship between IBD (UC, CD) and PSC using MR and GWAS. We found that IBD (including UC and CD) had a causal association with PSC, indicating that they may have a similar pathogenesis.

Several hypotheses have been proposed over the years, to explain the mechanisms of the model (Karlsen, 2016). An early RCT found that a small bowel bacterial overgrowth is associated with bile duct proliferation and destruction. Hypotheses based on the "leaky gut concept" indicated that bacteria and bacterial products could pass through damaged mucosa in IBD into the portal circulation (Lichtman et al., 1991). A review also reported that gut-derived mucosal T-cells expressing α4β7 would contribute to biliary inflammation. Barrier functions like the expression of pathogen pattern receptors are similar between the biliary and gut epithelium. The receptor CXCR6, is found to have a higher expression on liver-infiltrating and gut-infiltrating lymphocytes. Blocking the receptors is a developing treatment for inflammation (Adams et al., 2008). Another hypothesis suggests the possibility of FtsZ and TBB-5 antigens deriving from colonic content, which may drive the biliary inflammation. This is related to an abnormal immune response to intestinal microorganisms in susceptible individuals (Terjung et al., 2010). In summary, the association may result from hyperreactive bile duct proliferation, aberrant increased enterohepatic circulation pathogen-associated

TABLE 1 | MR estimates from each method assessing the causal effects of UC, CD, and IBD on PSC.

Exposure traits	MR methods	PSC				
		Number of SNPs	OR(95% CI)	SE	MR *p*-value	MR-Steiger test
IBD	MR Egger	112	1.16 (0.82∼1.63)	0.17	0.41	TRUE
	Inverse variance weighted	112	1.29 (1.12∼1.50)	0.08	< 0.01	
	Maximum likelihood	112	1.31 (1.23∼1.34)	0.03	< 0.01	
	Weighted median	112	1.24 (1.10∼1.40)	0.06	< 0.01	
	Robust adjusted profile score (RAPS)	112	1.29 (1.16∼1.44)	0.05	< 0.01	
UC	MR Egger	73	1.13 (0.80∼1.61)	0.18	0.50	TRUE
	Inverse variance weighted	73	1.48 (1.27∼1.72)	0.08	< 0.01	
	Maximum likelihood	73	1.52 (1.42∼1.63)	0.04	< 0.01	
	Weighted median	73	1.31 (1.15∼1.49)	0.06	< 0.01	
	Robust adjusted profile score (RAPS)	73	1.40 (1.23∼1.58)	0.06	< 0.01	
CD	MR Egger	99	1.06 (0.75∼1.50)	0.18	0.74	TRUE
	Inverse variance weighted	99	1.13 (0.99∼1.28)	0.07	0.07	
	Maximum likelihood	99	1.13 (1.07∼1.20)	0.03	< 0.01	
	Weighted median	99	1.12 (1.02∼1.24)	0.05	0.02	
	Robust adjusted profile score (RAPS)	99	1.13 (1.02∼1.26)	0.05	0.02	

TABLE 2 | Heterogeneity and pleiotropy analysis of UC, CD, and IBD with PSC, using different analytical methods.

Exposure traits	MR methods	PSC		
		Cochran Q statistic	Heterogeneity *p*-value	Pleiotropy *p*-value
IBD	MR Egger	644.30	<0.01	0.48
	Inverse variance weighted	647.28	<0.01	
	Maximum likelihood	643.80	<0.01	
UC	MR Egger	378.30	<0.01	0.10
	Inverse variance weighted	392.85	<0.01	
	Maximum likelihood	386.16	<0.01	
CD	MR Egger	538.50	<0.01	0.72
	Inverse variance weighted	539.21	<0.01	
	Maximum likelihood	538.40	<0.01	

molecular patterns (PAMPs), or an abnormal immune response (Tabibian et al., 2013).

All five MR methods indicated a significant relationship between UC and PSC. Another study also reported the strong association of PSC with UC (90%) (Adams et al., 2008). As for CD, MR Egger, IVW showed no significant relationship between CD and PSC, while the other three methods revealed the causal relationship. MR Egger and IVW are similar, both using the inverse of the outcome variance (Se^2) as the weight to carry out the fitting. The biggest difference between them is whether or not to consider the intercept term in the regression. Because of the low statistical power of MR-Egger, we usually focus on the consistency of the direction rather than the significance of estimates (Yeung and Schooling, 2020). From **Supplementary Figures S1–S3**, the consistent direction can be intuitively judged. Thus, we conclude that both UC and CD have a significant relationship with PSC.

In the Leave-one-out sensitivity analysis, we also found the specific SNPs that are strongly related to the disease (rs9836291 and rs2836883 for IBD; rs9836291 and rs2836883 for UC; rs3197999for CD). A previous study reported that the chromosome 3 SNP (rs3197999) is in the MST1 (Macrophage Stimulating 1) gene and is associated with MST1 protein levels. This SNP (rs3197999) can induce IBD by regulating the protein level of the Macrophage Stimulating Protein (MSP) (Di Narzo et al., 2017). Our results may provide inspiration for possible mechanism analyses in the future.

The causal association of IBD and PSC could contribute to improvement in PSC diagnostics and therapy, as well as prevention for IBD patients. For PSC, the diagnostics and therapy should better include IBD as a factor for improvement. According to PSC guidelines in the United States and Europe (Valuing Integrity, 2009; Lindor et al., 2015), major detection includes markers of cholestasis, bile duct lesions, and structuring on cholangiography with Magnetic resonance cholangiopancreatography (MRCP), along with a liver biopsy. Apart from these diagnostic investigations, we suggest regular colonoscopy surveillance for detecting IBD. For PSD patients with or without IBD, the clinical treatment and follow-up may be different. For example, clinical trials have tested the positive effect of antibiotics in PSC treatment (Tabibian et al., 2013). However, we should consider the potential consequent disturbance of gut microbiota (Karlsen, 2016), especially for IBD patients. For IBD, measures should be taken to prevent PSC at the very beginning. PSC-IBD has become an important public health issue due to the increased risk of malignancy (Rossi et al., 2016). Thus, regular physical examinations of PSC signs and symptoms are necessary for IBD patients. Currently, gut microbial signatures have been reported for their discriminatory function of determining early-stage PSC in IBD (Tabibian et al., 2013; Karlsen, 2016). Admittedly, the complex physiological machinery between IBD and PSC goes far beyond such simple models. Further studies are also needed to identify a potential mechanism for the association between IBD and PSC, to inform disease prevention.

Our study has several limitations. First, the SNP statistics we used were from a mixed population, 89.8% (86,640 in 96,486) of Europeans. However, the selected SNPs can explain 0.085 for IBD, 0.044 for UC, and 0.105 for CD of the phenotypic variation. Accordingly, the model fitness of PSC was also acceptable (0.06, 0.07, and 0.04). Second, although a series of sensitivity analyses have been conducted, we cannot guarantee that each SNP site meets the three basic conditions as instrumental variables. Considering the known confounding factors, we checked the confounders including smoking, drinking, and obesity and eliminated relative IVs. Admittedly, the influence of unknown possible confounders inevitably affects causal inference. Third, the MR model is based on the assumption of a linear effect association between exposure and outcome. Limited by the summary statistics, we did not perform a non-linearity of the association, which may be appropriate in some cases. Lastly, we found statistical heterogeneity between determined IV estimates, which may require further discussion (**Supplementary Tables S4–S6**).

CONCLUSION

MR analyses support the positive causal effect of IBD (including UC and CD) on PSC in the European population. Diagnostics and therapy improvement for PSC as well as the prevention of IBD should be promoted in clinical practice.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

XW and JC conceptualized and designed the study. YX, XC, and YS collected and analyzed the data in the study. YX and JC drafted the manuscript. All authors contributed to this article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (81970494) and Key Research and Development Program of Hunan Province (2019SK2041).

ACKNOWLEDGMENTS

We thank the blogger (Orange caramel) for helping us understand different MR methods.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.649376/full#supplementary-material

REFERENCES

Adams, D. H., Eksteen, B., and Curbishley, S. M. (2008). Immunology of the gut and liver: a love/hate relationship. *Gut* 57, 838–848. doi: 10.1136/gut.2007.122168

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016a). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40, 304–314. doi: 10.1002/gepi.21965

Bowden, J., Del Greco, M. F., Minelli, C., Davey, Smith G, Sheehan, N. A., and Thompson, J. R. (2016b). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int. J. Epidemiol.* 45, 1961–1974. doi: 10.1093/ije/dyw220

Di Narzo, A. F., Telesco, S. E., Brodmerkel, C., Argmann, C., Peters, L. A., Li, K., et al. (2017). High-throughput characterization of blood serum proteomics of IBD patients with respect to aging and genetic factors. *PLoS Genet.* 13:e1006565. doi: 10.1371/journal.pgen.1006565

Dyson, J. K., Beuers, U., Jones, D. E. J., Lohse, A. W., and Hudson, M. (2018). Primary sclerosing cholangitis. *Lancet* 391, 2547–2559. doi: 10.1016/S0140-6736(18)30300-3

Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* 13:e1007081. doi: 10.1371/journal.pgen.1007081

Hodson, R. (2016). Inflammatory bowel disease. *Nature* 540:S97. doi: 10.1038/540S97a

Karlsen, T. H. (2016). Primary sclerosing cholangitis: 50 years of a gut-liver relationship and still no love? *Gut* 65, 1579–1581. doi: 10.1136/gutjnl-2016-312137

Karlsen, T. H., Folseraas, T., Thorburn, D., and Vesterhus, M. (2017). Primary sclerosing cholangitis– a comprehensive review. *J Hepatol.* 67, 1298–1323. doi: 10.1016/j.jhep.2017.07.022

Lee, C. H., Cook, S., Lee, J. S., and Han, B. (2016). Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of z-scores. *Genomics Inform.* 14, 173–180. doi: 10.5808/GI.2016.14.4.173

Lichtman, S. N., Keku, J., Clark, R. L., Schwab, J. H., and Sartor, R. B. (1991). Biliary tract disease in rats with experimental small bowel bacterial overgrowth. *Hepatology* 13, 766–772. doi: 10.1002/hep.1840130425

Lindor, K. D., Kowdley, K. V., and Harrison, M. E. (2015). ACG clinical guideline: primary sclerosing cholangitis. *Am. J. Gastroenterol.* 110, 646–659; quiz660. doi: 10.1038/ajg.2015.112

Lindström, L., Jørgensen, K. K., Boberg, K. M., Castedal, M., Rasmussen, A., Rostved, A. A., et al. (2018). Risk factors and prognosis for recurrent primary sclerosing cholangitis after liver transplantation: a Nordic multicentre study. *Scand. J. Gastroenterol.* 53, 297–304. doi: 10.1080/00365521.2017.1421705

Liu, J. Z., Hov, J. R., Folseraas, T., Ellinghaus, E., Rushbrook, S. M., Doncheva, N. T., et al. (2013). Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet.* 45, 670–675. doi: 10.1038/ng.2616

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986. doi: 10.1038/ng.3359

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics* 163, 1153–1167.

Molodecky, N. A., Kareemi, H., Parab, R., Barkema, H. W., Quan, H., Myers, R. P., et al. (2011). Incidence of primary sclerosing cholangitis: a systematic review and meta-analysis. *Hepatology* 53, 1590–1599. doi: 10.1002/hep.24247

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381. doi: 10.1038/nature12873

Ricciuto, A., Kamath, B. M., and Griffiths, A. M. (2018). The IBD and PSC phenotypes of PSC-IBD. *Curr. Gastroenterol. Rep.* 20, 16. doi: 10.1007/s11894-018-0620-2

Rosen, M. J., Dhawan, A., and Saeed, S. A. (2015). Inflammatory bowel disease in children and adolescents. *JAMA Pediatr.* 169, 1053–1060. doi: 10.1001/jamapediatrics.2015.1982

Rossi, R. E., Conte, D., and Massironi, S. (2016). Primary sclerosing cholangitis associated with inflammatory bowel disease: an update. *Eur. J. Gastroenterol. Hepatol.* 28, 123–131. doi: 10.1097/MEG.0000000000000532

Tabibian, J. H., Talwalkar, J. A., and Lindor, K. D. (2013). Role of the microbiota and antibiotics in primary sclerosing cholangitis. *Biomed Res. Int.* 2013:389537. doi: 10.1155/2013/389537

Terjung, B., Söhne, J., Lechtenberg, B., Gottwein, J., Muennich, M., Herzog, V., et al. (2010). p-ANCAs in autoimmune liver disorders recognise human beta-tubulin isotype 5 and cross-react with microbial protein FtsZ. *Gut* 59, 808–816. doi: 10.1136/gut.2008.157818

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–713. doi: 10.1038/nature09270

Valuing Integrity (2009). EASL clinical practice guidelines: management of cholestatic liver diseases. *J. Hepatol.* 51, 237–267. doi: 10.1016/j.jhep.2009.04.009

Warren, K. W., Athanassiades, S., and Monge, J. I. (1966). Primary sclerosing cholangitis. A study of forty-two cases. *Am. J. Surg.* 111, 23–38. doi: 10.1016/0002-9610(66)90339-4

Yeung, C., and Schooling, C. M. (2020). Systemic inflammatory regulators and risk of Alzheimer's disease: a bidirectional Mendelian-randomization study. *Int. J. Epidemiol.* [Preprint]. Available online at: https://academic.oup.com/ije/advance-article/doi/10.1093/ije/dyaa241/6032243 (accessed March 5, 2021).

Zhao, Q., Chen, Y., Wang, J., and Small, D. S. (2019). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *Int. J. Epidemiol.* 48, 1478–1492. doi: 10.1093/ije/dyz142

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 26 March 2021
doi: 10.3389/fgene.2021.644101

Assessment of Bidirectional Relationships Between Polycystic Ovary Syndrome and Periodontitis: Insights From a Mendelian Randomization Analysis

Pengfei Wu[1,2†], Xinghao Zhang[3†], Ping Zhou[3], Wan Zhang[4], Danyang Li[4], Mingming Lv[5,6] and Xiaoyao Liao[7]*

[1] Hunan Key Laboratory of Animal Models for Human Diseases, Department of Laboratory Animals, Third Xiangya Hospital, Central South University, Changsha, China, [2] Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Center for Medical Genetics, Central South University, Changsha, China, [3] Department of Ultrasound, Third Xiangya Hospital, Central South University, Changsha, China, [4] Department of Biology, College of Arts and Sciences, Boston University, Boston, MA, United States, [5] Department of Oral Maxillofacial-Head Neck Oncology, College of Stomatology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [6] Shanghai Key Laboratory of Stomatology, National Clinical Research Center for Oral Diseases, Shanghai Research Institute of Stomatology, Shanghai, China, [7] School of Medicine, Dentistry and Nursing, University of Glasgow, Glasgow, United Kingdom

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong,
Hong Kong

Reviewed by:
Tao Huang,
Peking University, China
Changzheng Yuan,
Zhejiang University, China
Xihao Li,
Harvard University, United States

***Correspondence:**
Ping Zhou
zhouping1000@hotmail.com

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 19 December 2020
Accepted: 10 March 2021
Published: 26 March 2021

Citation:
Wu P, Zhang X, Zhou P, Zhang W, Li D, Lv M and Liao X (2021) Assessment of Bidirectional Relationships Between Polycystic Ovary Syndrome and Periodontitis: Insights From a Mendelian Randomization Analysis. Front. Genet. 12:644101. doi: 10.3389/fgene.2021.644101

Background: Observational studies have indicated an association between polycystic ovary syndrome (PCOS) and periodontitis, but it is unclear whether the association is cofounded or causal. We conducted a two-sample Mendelian randomization (MR) study to investigate the bidirectional relationship between genetically predicted PCOS and periodontitis.

Methods: From two genome-wide association studies we selected 13 and 7 single nucleotide polymorphisms associated with PCOS and periodontitis, respectively, as instrumental variables. We utilized publicly shared summary-level statistics from European-ancestry cohorts. To explore the causal effect of PCOS on periodontitis, 12,289 cases of periodontitis and 22,326 controls were incorporated, while 4,890 cases of PCOS and 20,405 controls in the reverse MR. Inverse-variance weighted method was employed in the primary MR analysis and multiple sensitivity analyses were implemented.

Results: Genetically determined PCOS was not causally associated with risk of periodontitis (odds ratio 0.97; 95% confidence interval 0.88–1.06; P = 0.50) per one-unit increase in the log-odds ratio of periodontitis. Similarly, no causal effect of periodontitis on PCOS was shown with the odds ratio for PCOS was 1.17 (95% confidence interval 0.91–1.49; P = 0.21) per one-unit increase in the log-odds ratio of periodontitis. Consistent results were yielded via additional MR methods. Sensitivity analyses demonstrated no presence of horizontal pleiotropy or heterogeneity.

Conclusion: The bidirectional MR study couldn't provide convincing evidence for the causal relationship between genetic liability to PCOS and periodontitis in the Europeans. Triangulating evidence across further observational and genetic-epidemiological studies is necessary.

Keywords: periodontitis, polycystic ovary syndrome, Mendelian randomization, causal inference, genetic epidemiology

INTRODUCTION

Polycystic ovary syndrome (PCOS) is a metabolic and hormonal disorder, which is prevalent in women of reproductive ages. About 15–20% premenopausal women are afflicted with PCOS in Europe (Sirmans and Pate, 2013; Escobar-Morreale, 2018) according to the Rotterdam criteria (Rotterdam ESHRE Group, 2004). PCOS is characterized by hyperandrogenism (HA), ovulatory dysfunction (OD), and polycystic ovarian morphology (PCOM). Unfavorable metabolic conditions, such as insulin resistance and endocrine-reproductive comorbidities, are commonly involved in the pathophysiology of PCOS (Azziz et al., 2019). As for the etiology, the complex interplay of genetic and environmental elements has been well recognized (Yu et al., 2018). However, the comprehensive links between PCOS and its downstream traits are waiting to be explored, as well as their underlying factors.

Periodontitis has posed an increasingly substantial burden on public health (Chaffee et al., 2020; Eke et al., 2020). Periodontitis features deterioration of local periodontal tissues, progressive destruction of alveolar bone and supporting ligament, and ultimately could result in tooth loss. Inadequate oral hygiene and microbe plaque accrual is known as initiation factors. Meanwhile, host susceptibilities to periodontopathic germs and inflammatory response are partly determined genetically; several loci have been identified in suggestive association with periodontitis (Divaris et al., 2013; Offenbacher et al., 2016; Munz et al., 2017, 2019; Kurushima et al., 2019). Moreover, periodontitis has much wider implications in multiple systems (Czesnikiewicz-Guzik et al., 2019; Bae and Lee, 2020; Sun et al., 2020) beyond oral health alone, albeit the underpinning is largely unidentified.

Emerging studies have proposed the association between PCOS and periodontitis (Kellesarian et al., 2017; Machado et al., 2020; Marquez-Arrico et al., 2020). As an essential parameter in the diagnosis of periodontitis, periodontal probing depth (PPD) was higher in PCOS in one recent cross-sectional research (Isik et al., 2020). The latest meta-analysis (Machado et al., 2020) suggested that patients with PCOS were at 28% higher odds of having periodontitis, and periodontitis increased the odds of PCOS by 46% (both $P < 0.001$). Notably, current evidence was mostly drawn from case-control or cohort studies, while high-quality studies like randomized controlled trails were scarce. Hence, effect estimates were prone to bias. Due to inherent weaknesses of traditional observational designs, we could not figure out whether the effects were causative. The direction of causality which was informative of risk factors and prevention strategies in the bidirectional association between PCOS and periodontitis, if existed, remains unknown.

Mendelian randomization (MR) is a powerful genetic-epidemiological tool to strengthen the causal inference and give the robust estimate (Burgess et al., 2019; Czesnikiewicz-Guzik et al., 2019; Bae and Lee, 2020; Sun et al., 2020), especially when well-powered randomized clinical trials are faced with financial challenges and ethical dilemmas. MR studies employs single nucleotide polymorphisms (SNPs) identified from the genome-wide association study (GWAS) as instrumental variables. MR design is supposed to give evidence whose strength is comparable to that of the meta-analysis (Davies et al., 2018). Therefore, we performed a bidirectional MR study to investigate the possible causal role for PCOS on periodontitis, along with the reverse causal effect of periodontitis on PCOS.

MATERIALS AND METHODS

Overall Study Design

This bidirectional MR study was undertaken in a framework as delineated in **Figure 1**. Causal effects of PCOS on periodontitis and the reverse causation were investigated separately. Three key assumptions underlie the MR analysis (Burgess et al., 2019). Firstly, the relevance assumption was met, considering genetic variants associated with exposures of interest were identified in sufficiently large-sample GWAS. Secondly, the independence assumption was validated. Mendel's laws make perfect randomization, that is, randomized segregation and independent assortment of alleles during gamete formation far precedes the onset of oucome diseases concerned. Hence, instrumental SNPs rarely links to confounders which are commonly involved in the traditional observation study examining the exposure-outcome relationship. Lastly, the exclusion-restriction assumption requires that instrumental variables exert influences on the outcome via no other pathways than the exposure, also known as pleiotropic effects. We have examined the potential pleiotropy through multiple sensitivity analyses. This MR study was conducted using publicly shared datasets, and the approval by concerned ethical committee and consent from all participants were obtained in the original GWAS studies and their contributing cohorts. Additional ethic statement or consent was not required.

Summary Statistics for PCOS

Day et al. (2018, 2019) has conducted the largest GWAS meta-analysis of PCOS in 10,074 cases and 103,164 controls of European ancestry and identified 14 SNPs at genome-wide significance ($P < 5 \times 10^{-8}$). Diagnosis of PCOS is

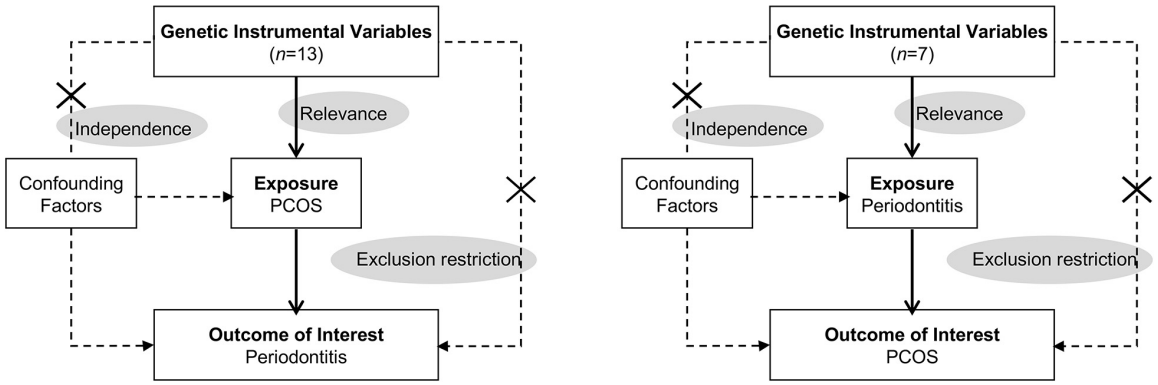


FIGURE 1 | Schematics for the bidirectional Mendelian randomization design. Mendelian randomization requires valid genetic instrumental variants satisfying three assumptions. PCOS, polycystic ovary syndrome.

based on National Institutes of Health criteria (Carmina, 2004), Rotterdam criteria (Rotterdam ESHRE Group, 2004), or self-report questionnaire (Day et al., 2015). Presence of both OD and HA satisfies the National Institutes of Health criteria, while Rotterdam criteria incorporates PCOM and requires two out of three principal traits to be met. Self-reported diagnosis was used in the 23 and Me (Mountain View, CA, United States) cohort, and due to the data shared policy, summary-level statistics from 4,890 cases of PCOS and 20,405 controls excluding this cohort were available. Nevertheless, 14 genome-wide significant loci manifested negligible heterogeneity in the effect direction and magnitude, after examining the odds ratio for PCOS as a function of diagnostic criteria. Shared genetic architecture across three diagnostic criteria was elaborated in the GWAS (Day et al., 2018). Hence, we utilized summary statistics for PCOS derived from as large sample fulfilling either criterion.

In the MR analysis exploring causal effects of PCOS on periodontitis, 13 SNPs were selected as instrumental variables (**Supplementary Table 1**). SNPs with minor allele frequency less than 1% or Hardy-Weinberg equilibrium test *P*-value less than 0.0001 will not be considered. Palindromic alleles with minor allele frequency above 0.45 was also excluded due to the ambiguous strand aligning issue. One such variant (rs853854, A/T, allele frequency, 0.499/0.501) reported in the original GWAS, hence was not selected as instrumental variables for PCOS. Linkage disequilibrium (threshold set at $R^2 > 0.01$, within 1 Mb window, EUR panel of 1,000 Genomes Project Phase 3) was examined (Myers et al., 2020) and the variant with the lowest *P*-value at each locus was retained. Look-up of potential pleiotropic associations (**Supplementary Table 2**) was performed in the GWAS Catalog (Buniello et al., 2019). Proportion of variance explained was calculated using the formula $2 \times MAF \times (1 - MAF) \times Beta^2$, where MAF was the minor allele frequency, Beta represented the estimated genetic effect on the risk of PCOS. Total variance explained by instrumental SNPs for PCOS approximated 6.2% (**Supplementary Table 3**). The strength of each SNP was assessed by *F*-statistic using the formula $R^2(N - 2)/(1 - R^2)$, where R^2 was the proportion of variance explained, N was the total sample size. *F*-statistic for individual variant ranged from 30.8 to 57.6; therefore, none was weak instrument ($F < 10$). GWAS results for PCOS are publicly available from Apollo[1]. Effect size has been adjusted for age and presented as beta (log-odds) per additional effect allele.

Summary Statistics for Periodontitis

Summary-level data for periodontitis were obtained from the newly released GWAS (Shungin et al., 2019). Totally, this GWAS incorporated 12,289 cases and 22,326 controls of European-ancestry from seven contributing cohorts in the Gene-Lifestyle Interactions in Dental Endpoints consortium. The clinical diagnostic criteria by the Centers for Disease Control and Prevention/American Academy of Periodontology (Page and Eke, 2007) and self-reported diagnosis from the Women's Health Study at Brigham and Women's Hospital (Yu et al., 2018) were primarily adopted, whereas additional inclusion criteria were defined as one of the following conditions, two or more tooth surfaces with PPD $\geq$ 5 mm, or four or more with PPD $\geq$ 4 mm, two or more tooth surfaces with PPD $\geq$ 5.5 mm or dental records of "gum surgery".

For the MR analysis of periodontitis on PCOS, seven instrumental SNPs associated with periodontitis ($P < 5 \times 10^{-6}$) were selected (**Supplementary Table 4**) since no genome-wide significant loci were identified in the European-ancestry GWAS (Shungin et al., 2019). No pleiotropic associations were identified through look-up in the GWAS Catalog (**Supplementary Table 5**). Considering allele frequency variable has been removed to prevent re-identification of individuals in the shared dataset, we utilized the reference minor allele frequency from 1,000 Genomes

[1]https://www.repository.cam.ac.uk/handle/1810/289950

European panel to calculate *F*-statistic. Although a more liberal threshold ($P < 5 \times 10^{-6}$) was adopted, there was no evidence of the existence of weak instrument (**Supplementary Table 6**). Initially, 20 SNPs reaching an arbitrary threshold for suggestive association ($P < 5 \times 10^{-6}$) were retrieved from the summary-level dataset. Criteria of instrumental SNPs for periodotitis, like minor allele frequency and linkage disequilibrium, were set similar to the criteria of instrumental SNPs for PCOS. Linkage disequilibrium was examined, and nine SNPs with the lowest *P*-value at each locus were kept. Two SNPs were further omitted, for whom or their proxies ($R^2 > 0.8$), corresponding statistics were not present in the PCOS dataset. Effect estimates denote log-odds of periodontitis given by the additive genetic model, adjusted for age and principal components as covariates. Summary-level statistics for periodontitis can be obtained from the dataset depository, University of Bristol[2].

Statistical Analysis

The statistical analysis was performed using the R software, version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria) and *TwoSample MR* and *MR-PRESSO* packages (Hemani et al., 2018; Verbanck et al., 2018). The inverse variance weighted (IVW) approach was implemented as the primary MR method to yield an overall estimate from multiple instrumental variables (Burgess et al., 2013). Specifically, for each variant SNP_k, its genetic effect on the exposure and outcome per additional effect allele, $\hat{\beta}_{X_k}$ and $\hat{\beta}_{Y_k}$, and their standard errors $\hat{\sigma}_{X_k}$ and $\hat{\sigma}_{Y_k}$, MR causal estimates can be given by the Wald ratio $\hat{\beta}_{Y_k} \Big/ \hat{\beta}_{X_k}$ with the standard error $\hat{\sigma}_{Y_k} \Big/ \hat{\beta}_{X_k}$. Then an overall causal estimator $\hat{\beta}_{\mathrm{IVW}}$ with standard error $\hat{\sigma}_{\mathrm{IVW}}$ can be derived as shown below.

$$\hat{\beta}_{\mathrm{IVW}} = \frac{\Sigma_k \hat{\beta}_{X_k} \hat{\beta}_{Y_k} \hat{\sigma}_{Y_k}^{-2}}{\Sigma_k \hat{\beta}_{X_k}^{2} \hat{\sigma}_{Y_k}^{-2}}$$

$$\hat{\sigma}_{\mathrm{IVW}} = \sqrt{1 \Big/ \Sigma_k \hat{\beta}_{X_k}^{2} \hat{\sigma}_{Y_k}^{-2}}$$

Inverse variance weighted estimates requires all instrumental variants to be valid and would be biased if average pleiotropic effects deviated from zero. Hence, robust analyses under weaker assumptions are required to provide valid causal inferences and to assess the sensitivity across these findings. Three complementary MR methods were adopted, MR-pleiotropy residual sum and outlier (MR-PRESSO), weighted median estimator and MR-Egger regression. MR-PRESSO (Verbanck et al., 2018) takes into account the horizontal pleiotropy and gives a causal estimate corrected for it, should instrumental variables with horizontal pleiotropy be identified via MR-PRESSO global test. Weighted median (Bowden et al., 2016) yields a pooled effect size more robustly if more than 50% of instrumental variables are valid (majority valid assumption). It is not as sensitively influenced by the presence of a handful of pleiotropic variants as the IVW method. MR-Egger regression (Bowden et al., 2015) has a lower statistical power with a wide range of causality estimates. It requires pleiotropic effects to be independent of the variant–exposure associations (instrument strength independent of direct effect assumption). MR-Egger method gives an causal estimate with the regression slope, meanwhile MR-Egger intercept also provides an assessment of unbalanced horizontal pleiotropy across all variants. Additional sensitivity analyses for heterogeneity detection were performed through Cochran's *Q* test, leave-one-out plots and funnel plots. Power calculations were conducted in *mRnd*, a web-based application (Brion et al., 2013) assuming a 80% power and 5% Type-I error rate. Statistical significance was set at 0.025 ($P = 0.05/2$ association tests) with the Bonferroni method to correct for multiple testing.

[2]https://data.bris.ac.uk/data/dataset/2j2rqgzedxlq02oqbb4vmycnc2

RESULTS

Estimated Causal Effect of PCOS on Periodontitis

Overall, there was no causal relationship between genetically predicted PCOS and periodontitis. Primary MR results (**Figure 2**) indicated that odds ratio (OR) of periodontitis is 0.97 [95% confidence interval (CI), 0.88–1.06; $P = 0.50$] per one-unit increase in log-OR of PCOS (equivalent to 2.718 fold change in the OR of PCOS) by the IVW method, and estimates by MR-PRESSO and weighted median methods (**Figure 3**) were consistent with respect to the effect size and direction. Causal estimates given by MR-Egger (OR = 1.04 per one-unit increase in log-OR of PCOS) with wide 95% CI (0.67–1.62) were less precise. No horizontal pleiotropy was identified (**Supplementary Table 7**), as shown by MR-Egger test (Intercept = -0.009; $P = 0.75$) and MR-PRESSO global test ($RSS_{obs} = 10.62$; $P = 0.69$). There was no significant heterogeneity detected through Cochran's *Q* test ($Q = 9.07$; $P = 0.70$). Elimination of single instrumental SNP would not lead to distortion of the overall MR estimate (**Supplementary Figure 1**), whereas overall symmetry of the funnel plot further demonstrated negligible heterogeneity and validated the robustness of the causal estimate given by the fixed-effect IVW method. Since the closer an OR approaching 1, the much larger a sample size is required to detect such a weak effect, this study was underpowered to detect an OR interval of 0.88–1.13 according to our power calculation.

Estimated Causal Effect of Periodontitis on PCOS

Genetic liability to periodontitis was not causally associated with risk of PCOS. By the IVW method (**Figure 2**), the OR of PCOS was 1.17 (95% CI 0.91–1.49; $P = 0.21$) per one-unit increment in the log-OR of periodontitis. Causal estimates by three additional approaches (**Figure 3**) did not reach nominal significance, either. There was no evidence for the existence of outlier SNPs, heterogeneity or horizontal pleiotropy by the MR sensitivity analyses (**Supplementary Table 7**). Examination of the leave-one-out plot and funnel plot (**Supplementary Figure 1**) suggested that the MR results were not driven by certain SNP

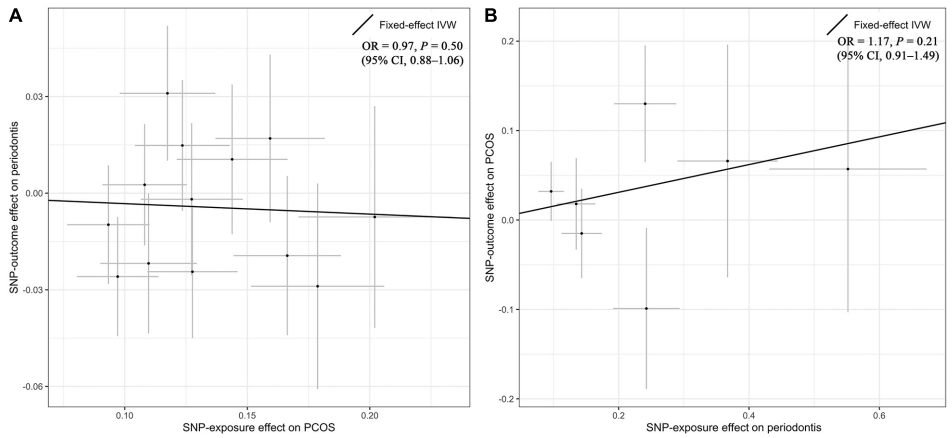


FIGURE 2 | Primary results by Mendelian randomization analysis using the inverse-variance-weighted method. The causal estimate (overall fitted line) for the effect of PCOS on periodontitis was shown in panel **(A)**, while the overall effect for the casual association of periodontitis with PCOS was presented in panel **(B)**. Individual SNP-effect on the outcome (point and vertical line) against its effect on the exposure (point and horizontal line) was delineated in the background. CI, confidence interval; IVW, inverse variance weighted; OR, odds ratio; PCOS, polycystic ovary syndrome; SNP, single nucleotide polymorphism.

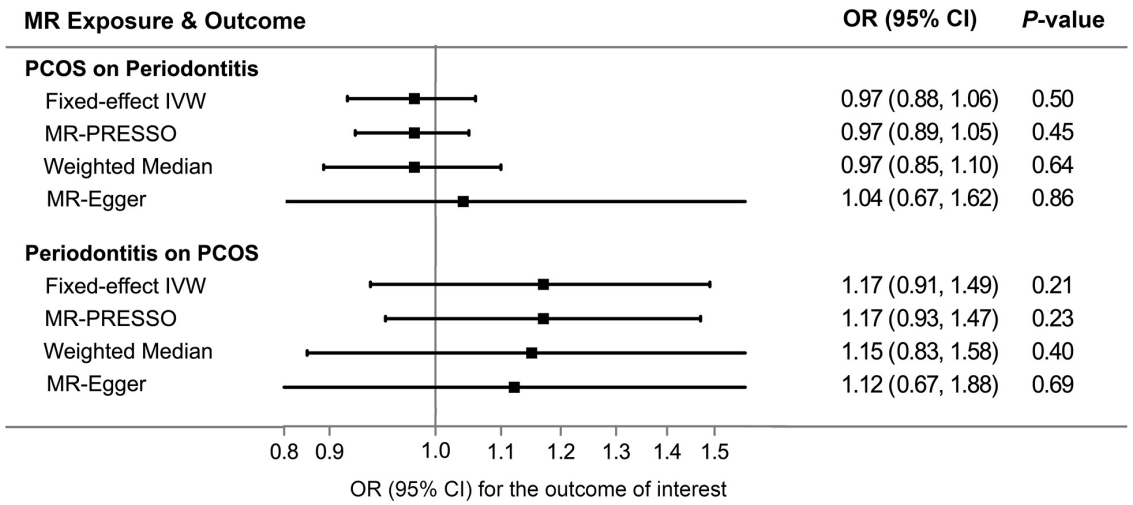


FIGURE 3 | Comparisons of Mendelian randomization results by different methods. CI, confidence interval; IVW, inverse variance weighted; MR, Mendelian randomization; OR, odds ratio; PCOS, polycystic ovary syndrome; PRESSO, pleiotropy residual sum and outlier.

alone and overall causal estimates were consistent and accurate on the whole. This study was underpowered to detect an OR interval of 0.70–1.26 according to our power calculation.

DISCUSSION

This study explored the bidirectional relationships between PCOS and periodontitis using a two-sample MR design for the first time. The MR analysis failed to identify a causal effect of PCOS on periodontitis in European women, and the statistical power was adequate if the observed effect (OR = 1.28) in the recent meta-analysis (Machado et al., 2020) represented a true causality. Meanwhile, our results provided no evidence that genetically predicted risk of periodontitis was causally associated with liability to PCOS. Previous established relationship between predisposition to PCOS and periodontitis might result from uncontrolled biases or cofounders in observational epidemiological studies.

Two systematic reviews (Kellesarian et al., 2017; Marquez-Arrico et al., 2020) have been conducted to address the hypothesis: whether a causal relationship exists between PCOS and periodontal diseases. The qualitative evidence suggested that a variety of periodontal parameters, such as PPD and bleeding on probing, together with altered immunoinflammatory and microbiological outcomes were observed in patients with PCOS. A positive association between PCOS and periodontal diseases was concluded. However, most included studies featured the case-control design, a small sample size ranging from 52 to 196, and non-follow-up. The strength of evidence should be taken

into account before promoting regular referral of PCOS patients to oral-health evaluation. Machado et al. (2020) conducted a quantitative synthesis and yielded an OR of 1.28 (95% CI 1.06–1.55; $P < 0.001$) for the effect of PCOS on periodontitis and an OR of 1.46 (95% CI 1.29–1.66; $P < 0.001$) for the reverse association. Notably, the effect estimates were derived from three Asian cohorts (Porwal et al., 2014; Tong et al., 2019; Saljoughi et al., 2020) in the meta-analysis. Discrepancies between this MR study and the recent meta-analysis in the assessment of bidirectional association might be partly explained by the population difference.

It has been postulated that PCOS and periodontitis is linked by systemic inflammation and oxidative status (Dursun et al., 2011; Saglam et al., 2018). Myeloperoxidase and nitric oxide, indicative of oxidative stress, were higher in women with PCOS than healthy controls (Marquez-Arrico et al., 2020). Likewise, increased levels of malondialdehyde and 8-hydroxy-2′-deoxyguanosine were identified in PCOS, which were prominent both in serum and gingival crevicular fluid (Saglam et al., 2018). Neutrophils are recognized to play a key role in the initiation of inflammatory responses to periodontal pathogens, and local oxidative stress is strengthened in periodontitis (Porwal et al., 2014). Hence, altered oxidative status in PCOS might contribute to the occurrence or progression of periodontitis. Considering the observed links between periodontitis and multiple diseases (Czesnikiewicz-Guzik et al., 2019; Bae and Lee, 2020; Sun et al., 2020), inflammatory response cascade in periodontitis might as well exert an influence on the risk of PCOS through molecular changes in the metabolic-endocrine networks.

Notably, PCOS is known as a heterogeneous disorder with three principal components, OD, HA, and PCOM. Several other diagnostic criteria have been adopted as well. To enhance the statistical power, there is a trade-off between phenotypic refinement and incorporating sufficiently large cohorts. However, Day et al. (2018) has demonstrated minimal heterogeneity of the SNP effect, except one SNP near GATA4/NEIL2 (rs804279, $P_{het} = 2.6 \times 10^{-5}$), across NIH, Rotterdam, and self-reported criteria. Therefore, we deemed that negligible bias should be incurred when utilizing summary statistics derived from multiple cohorts. Besides, with currently available summary-level statistics, we could not perform a comprehensive MR analysis exploring the effects of different subtypes of PCOS on periodontitis. Likewise, a broad spectrum of periodontal diseases, incorporating gingivitis, moderate chronic periodontitis, and severe aggressive periodontitis, could not be considered in the MR analysis.

There are several limitations in this study. Firstly, instrumental SNPs selected from GWAS were mainly based on a statistically driven hypotheses, and genome-wide significance alone cannot guarantee the plausibility of these variants. Especially, their biological implication and complexity has not been fully understood, let alone thoroughly examined. Notably, seven instrumental SNPs for periodontitis were only suggestively significant ($P < 5 \times 10^{-6}$). Therefore, we should be cautious with the null effect in the MR analysis of periodontitis on PCOS, which might be due to the lack of association strength of instrumental SNPs. Secondly, current GWAS design and analysis models cannot take all sources of potential bias into account, such as Collider bias, winner's curse and Beavis effects. Given that the genetic estimates of SNP-association underlie the further MR analysis, there might be bias incurred into the MR causal estimates as well. Thirdly, selected instrumental variants collectively explained a small proportion of variance of PCOS or periodontitis. Thus, we were not capable of detecting weak effects, although horizontal pleiotropy and weak instrument bias have been ruled out. Forthly, with summary-level data, we failed to conduct a stratified analysis exploring the effects of PCOS based on body mass index or obesity status, which has been proposed to account for diverged outcomes in PCOS. Moreover, GWAS estimates of PCOS were from studies in women while the corresponding estimates of periodontitis were not restricted to female participants, which could possibly introduce bias. Lastly, data sets were of European ancestry, and cautions should be exercised when interpreting and generalizing the MR results.

To conclude, this bidirectional MR study failed to provide convincing evidence to support the causal relationshipbetween genetic liability to PCOS and periodontitis. To elucidate previously observed links, high-qualified clinical trials and laboratory researches are warranted. Triangulation of evidence across multiple study designs is essential when assessing the association between PCOS and periodontitis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

PW, XZ, and PZ conceptualized the study. PW, XZ, WZ, DL, and ML took part in the data curation, methodology, software, and formal analysis. PW, XZ, PZ, ML, and XL were in charge of the validation and visualization. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We extend sincere thanks to Dr. Felix Day, Dr. Dmitry Shungin et al., GLIDE, UK Biobank and all concerned investigators and consortia for sharing GWAS summary statistics on PCOS and periodontitis. We thank all individuals for participating in the original GWAS studies. PW has received a visiting Ph.D. stipend from the China Scholarship Council.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.644101/full#supplementary-material

REFERENCES

Azziz, R., Kintziger, K., Li, R., Laven, J., Morin-Papunen, L., Merkin, S. S., et al. (2019). Recommendations for epidemiologic and phenotypic research in polycystic ovary syndrome: an androgen excess and PCOS society resource. *Hum. Reprod.* 34, 2254–2265. doi: 10.1093/humrep/dez185

Bae, S. C., and Lee, Y. H. (2020). Causal association between periodontitis and risk of rheumatoid arthritis and systemic lupus erythematosus: a Mendelian randomization. *Z. Rheumatol.* 79, 929–936. doi: 10.1007/s00393-019-00742-w

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525. doi: 10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40, 304–314. doi: 10.1002/gepi.21965

Brion, M. J., Shakhbazov, K., and Visscher, P. M. (2013). Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* 42, 1497–1501. doi: 10.1093/ije/dyt179

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37, 658–665. doi: 10.1002/gepi.21758

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., et al. (2019). Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res.* 4, 186. doi: 10.12688/wellcomeopenres.15555.1

Carmina, E. (2004). Diagnosis of polycystic ovary syndrome: from NIH criteria to ESHRE-ASRM guidelines. *Minerva Ginecol.* 56, 1–6.

Chaffee, B. W., Persai, D., and Vora, M. V. (2020). Interdental cleaning and oral health status in an adult cohort, 2015 to 2018. *J. Dent. Res.* 99, 1150–1156. doi: 10.1177/0022034520926139

Czesnikiewicz-Guzik, M., Osmenda, G., Siedlinski, M., Nosalski, R., Pelka, P., Nowakowski, D., et al. (2019). Causal association between periodontitis and hypertension: evidence from Mendelian randomization and a randomized controlled trial of non-surgical periodontal therapy. *Eur. Heart J.* 40, 3459–3470. doi: 10.1093/eurheartj/ehz646

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* 362:k601. doi: 10.1136/bmj.k601

Day, F., Karaderi, T., Jones, M. R., Meun, C., He, C., Drong, A., et al. (2018). Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* 14:e1007813. doi: 10.1371/journal.pgen.1007813

Day, F., Karaderi, T., Jones, M. R., Meun, C., He, C., Drong, A., et al. (2019). Correction: large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* 15:e1008517. doi: 10.1371/journal.pgen.1008517

Day, F. R., Hinds, D. A., Tung, J. Y., Stolk, L., Styrkarsdottir, U., Saxena, R., et al. (2015). Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat. Commun.* 6:8464. doi: 10.1038/ncomms9464

Divaris, K., Monda, K. L., North, K. E., Olshan, A. F., Reynolds, L. M., Hsueh, W. C., et al. (2013). Exploring the genetic basis of chronic periodontitis: a genome-wide association study. *Hum. Mol. Genet.* 22, 2312–2324. doi: 10.1093/hmg/ddt065

Dursun, E., Akalin, F. A., Guncu, G. N., Cinar, N., Aksoy, D. Y., Tozum, T. F., et al. (2011). Periodontal disease in polycystic ovary syndrome. *Fertil. Steril.* 95, 320–323. doi: 10.1016/j.fertnstert.2010.07.1052

Eke, P. I., Borgnakke, W. S., and Genco, R. J. (2020). Recent epidemiologic trends in periodontitis in the USA. *Periodontology* 2000, 257–267. doi: 10.1111/prd.12323

Escobar-Morreale, H. F. (2018). Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. *Nat. Rev. Endocrinol.* 14, 270–284. doi: 10.1038/nrendo.2018.24

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7:e34408. doi: 10.7554/eLife.34408

Isik, Y., Telatar, G. Y., Neselioglu, S., Bicer, C., and Gurlek, B. (2020). Evaluation of periodontal status in different phenotypes of polycystic ovary syndrome in untreated patients of early reproductive age: a case-control study. *J. Obstet. Gynaecol. Res.* 46, 459–465. doi: 10.1111/jog.14179

Kellesarian, S. V., Malignaggi, V. R., Kellesarian, T. V., Al-Kheraif, A. A., Alwageet, M. M., Malmstrom, H., et al. (2017). Association between periodontal disease and polycystic ovary syndrome: a systematic review. *Int. J. Impot. Res.* 29, 89–95. doi: 10.1038/ijir.2017.7

Kurushima, Y., Tsai, P. C., Castillo-Fernandez, J., Couto Alves, A., El-Sayed Moustafa, J. S., Le Roy, C., et al. (2019). Epigenetic findings in periodontitis in UK twins: a cross-sectional study. *Clin. Epigenetics* 11:27. doi: 10.1186/s13148-019-0614-4

Machado, V., Escalda, C., Proenca, L., Mendes, J. J., and Botelho, J. (2020). Is there a bidirectional association between polycystic ovarian syndrome and periodontitis? A systematic review and meta-analysis. *J. Clin. Med.* 9:1961. doi: 10.3390/jcm9061961

Marquez-Arrico, C. F., Silvestre-Rangil, J., Gutierrez-Castillo, L., Martinez-Herrera, M., Silvestre, F. J., and Rocha, M. (2020). Association between periodontal diseases and polycystic ovary syndrome: a systematic review. *J. Clin. Med.* 9:1586. doi: 10.3390/jcm9051586

Munz, M., Richter, G. M., Loos, B. G., Jepsen, S., Divaris, K., Offenbacher, S., et al. (2019). Meta-analysis of genome-wide association studies of aggressive and chronic periodontitis identifies two novel risk loci. *Eur. J. Hum. Genet.* 27, 102–113. doi: 10.1038/s41431-018-0265-5

Munz, M., Willenborg, C., Richter, G. M., Jockel-Schneider, Y., Graetz, C., Staufenbiel, I., et al. (2017). A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis. *Hum. Mol. Genet.* 26, 2577–2588. doi: 10.1093/hmg/ddx151

Myers, T. A., Chanock, S. J., and Machiela, M. J. (2020). LDlinkR: an R Package for rapidly calculating linkage disequilibrium statistics in diverse populations. *Front. Genet.* 11:157. doi: 10.3389/fgene.2020.00157

Offenbacher, S., Divaris, K., Barros, S. P., Moss, K. L., Marchesan, J. T., Morelli, T., et al. (2016). Genome-wide association study of biologically informed periodontal complex traits offers novel insights into the genetic basis of periodontal disease. *Hum. Mol. Genet.* 25, 2113–2129. doi: 10.1093/hmg/ddw069

Page, R. C., and Eke, P. I. (2007). Case definitions for use in population-based surveillance of periodontitis. *J. Periodontol.* 78(7 Suppl.), 1387–1399. doi: 10.1902/jop.2007.060264

Porwal, S., Tewari, S., Sharma, R. K., Singhal, S. R., and Narula, S. C. (2014). Periodontal status and high-sensitivity C-reactive protein levels in polycystic ovary syndrome with and without medical treatment. *J. Periodontol.* 85, 1380–1389. doi: 10.1902/jop.2014.130756

Rotterdam ESHRE Group (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). *Hum. Reprod.* 19, 41–47. doi: 10.1093/humrep/deh098

Saglam, E., Canakci, C. F., Sebin, S. O., Saruhan, N., Ingec, M., Canakci, H., et al. (2018). Evaluation of oxidative status in patients with chronic periodontitis and polycystic ovary syndrome: a cross-sectional study. *J. Periodontol.* 89, 76–84. doi: 10.1902/jop.2017.170129

Saljoughi, F., Nasri, K., and Bayani, M. (2020). Gingival crevicular fluid levels of visfatin in patients with chronic periodontitis and polycystic ovary syndrome. *Obstet. Gynecol. Sci.* 63, 87–93. doi: 10.5468/ogs.2020.63.1.87

Shungin, D., Haworth, S., Divaris, K., Agler, C. S., Kamatani, Y., Keun Lee, M., et al. (2019). Genome-wide analysis of dental caries and periodontitis combining clinical and self-reported data. *Nat. Commun.* 10:2773. doi: 10.1038/s41467-019-10630-1

Sirmans, S. M., and Pate, K. A. (2013). Epidemiology, diagnosis, and management of polycystic ovary syndrome. *Clin. Epidemiol.* 6, 1–13. doi: 10.2147/CLEP.S37559

Sun, Y. Q., Richmond, R. C., Chen, Y., and Mai, X. M. (2020). Mixed evidence for the relationship between periodontitis and Alzheimer's disease: a bidirectional mendelian randomization study. *PLoS One* 15:e0228206. doi: 10.1371/journal.pone.0228206

Tong, C., Wang, Y. H., Yu, H. C., and Chang, Y. C. (2019). Increased risk of polycystic ovary syndrome in taiwanese women with chronic periodontitis:

a nationwide population-based retrospective cohort study. *J. Womens Health (Larchmt)* 28, 1436–1441. doi: 10.1089/jwh.2018.7648

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50, 693–698. doi: 10.1038/s41588-018-0099-7

Yu, Y. H., Doucette-Stamm, L., Rogus, J., Moss, K., Zee, R. Y. L., Steffensen, B., et al. (2018). Family history of MI, smoking, and risk of periodontal disease. *J. Dent. Res.* 97, 1106–1113. doi: 10.1177/0022034518782189

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 26 April 2021
doi: 10.3389/fgene.2021.656637

Recovering Spatially-Varying Cell-Specific Gene Co-expression Networks for Single-Cell Spatial Expression Data

*Jinge Yu and Xiangyu Luo**

Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Recent advances in single-cell technologies enable spatial expression profiling at the cell level, making it possible to elucidate spatial changes of cell-specific genomic features. The gene co-expression network is an important feature that encodes the gene-gene marginal dependence structure and allows for the functional annotation of highly connected genes. In this paper, we design a simple and computationally efficient two-step algorithm to recover spatially-varying cell-specific gene co-expression networks for single-cell spatial expression data. The algorithm first estimates the gene expression covariance matrix for each cell type and then leverages the spatial locations of cells to construct cell-specific networks. The second step uses expression covariance matrices estimated in step one and label information from neighboring cells as an empirical prior to obtain thresholded Bayesian posterior estimates. After completing estimates for each cell, this algorithm can further predict or interpolate gene co-expression networks on tissue positions where cells are not captured. In the simulation study, the comparison against the traditional cell-type-specific network algorithms and the cell-specific network method but without incorporating spatial information highlights the advantages of the proposed algorithm in estimation accuracy. We also applied our algorithm to real-world datasets and found some meaningful biological results. The accompanied software is available on https://github.com/jingeyu/CSSN.

Keywords: Bayesian posterior estimates, cell-specific, gene co-expression network, prediction, single-cell spatial expression, neighborhood

OPEN ACCESS

Edited by:
Jiebiao Wang,
University of Pittsburgh, United States

Reviewed by:
Jinjin Tian,
Carnegie Mellon University, United States
Hao Dai,
Chinese Academy of Sciences (CAS), China

****Correspondence:***
Xiangyu Luo
xiangyuluo@ruc.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 January 2021*
Accepted: *18 March 2021*
Published: *26 April 2021*

Citation:
Yu J and Luo X (2021) Recovering Spatially-Varying Cell-Specific Gene Co-expression Networks for Single-Cell Spatial Expression Data. Front. Genet. 12:656637. doi: 10.3389/fgene.2021.656637

1. INTRODUCTION

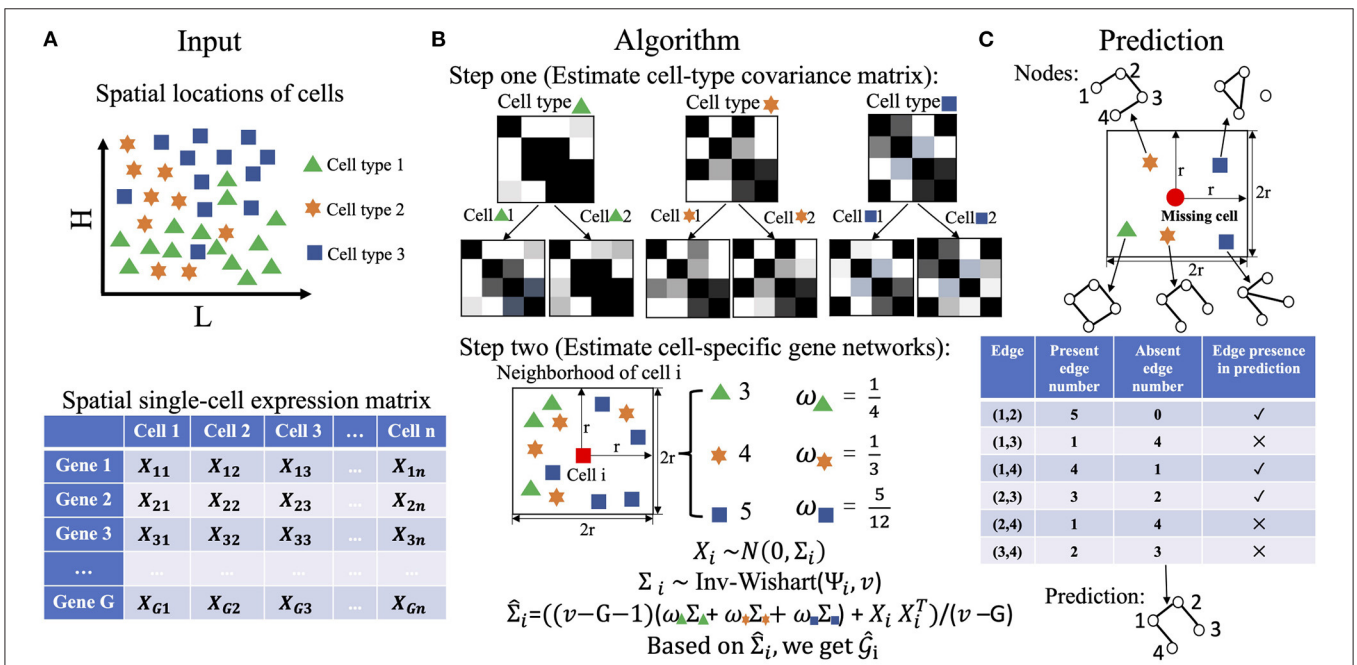
The last decade witnesses that the single-cell RNA-sequencing has revolutionized the focus of genomic analyses from bulk samples to single cells, but the technology loses important cell spatial information during tissue dissociation. Fortunately, recent technological advances have allowed for measurements of the gene expression levels at single-cell resolution while retaining the coordinates of cells in the tissue section (Chen et al., 2015; Moffitt et al., 2018; Wang et al., 2018). Specifically, various spatially resolved transcriptomic techniques have been developed to profile single-cell expression with cells' spatial information, including MERFISH (Chen et al., 2015), seqFISH (Lubeck et al., 2014), and FISSEQ (Lee et al., 2014), just to name a few. They are mainly based on either *in situ* hybridization or *in situ* sequencing. Fluorescence *in situ* hybridization (FISH) based approaches can measure hundreds of preselected marker genes, while *in situ* sequencing based approaches

can measure thousands of transcripts. Moreover, different techniques may have different strategies to capture transcriptomic spatial information. For example, MERFISH adopts an imaging-based way to map transcriptomic spatial organization for a three-dimensional tissue region. Usually, the region needs to be first sectioned into evenly spaced slices, and MERFISH is then performed on these slices, resulting in two-dimensional localization information. The information makes it possible to investigate spatial and functional organization of cells.

The amazing biological progress also offers rich opportunities to investigate the spatial patterns of cell-specific genomic features (Zhang et al., 2020). When features are genes, Sun et al. (2020) developed a statistical method to identify genes with spatially differential expressions. Li D. et al. (2020) utilized an expert system to predict signaling gene expression using information from nearby cells. However, as observed gene expressions may suffer from systematic biases (Köster et al., 2019) and are dynamically driven by an underlying regulation system, it is of more interest to study a more stable feature—gene co-expression network—(Dai et al., 2019) and learn its spatial pattern from one cell to another.

The gene co-expression network (Butte and Kohane, 2000; Stuart et al., 2003; Carter et al., 2004) can be encoded in an undirected graph, where nodes correspond to genes and an edge between nodes A and B indicates a significant association between expressions of the genes A and B. It has important biological applications including functional annotation for a set of unknown but highly connected genes (Serin et al., 2016) and single cell expression simulation (Tian et al., 2021). The pipeline to construct gene co-expression networks usually consists of two steps (Zhang and Horvath, 2005). In step one, we adopt a similarity measure (e.g., the absolute value of Pearson correlation) and calculate the similarity for all pairs of genes. In step two, we choose a threshold and genes with similarity larger than the threshold are thought of as co-expressed. Following the pipeline, Dai et al. (2019) proposed a hypothesis testing based approach to estimate cell-specific gene co-expression network, which is a breakthrough from "cell-type-specific" to "cell-specific" since most computational network methods for single-cell expression are restricted to a group of cells and ignore cell heterogeneity. Li L. et al. (2020) extends the approach to a conditional cell-specific network situation. Unfortunately, the method (Dai et al., 2019) does not incorporate the spatial information of cells and thus may lose power in estimating cell-specific gene co-expression structures, let alone carry out network prediction given a new cell location in the tissue.

To overcome the challenges, we present an easy-to-implement and computationally efficient two-step algorithm to recover cell-specific gene co-expression networks for single-cell spatial expression data. The input of the proposed algorithm is comprised of the spatial locations of cells, cell labels, as well as the gene-cell expression matrix (**Figure 1A**). If cell label information is not available, we can first carry out clustering using single-cell expression data clustering tools (Butler et al., 2018; Stuart et al., 2019). In step one, we estimate the sample expression covariance



	Cell 1	Cell 2	Cell 3	...	Cell n
Gene 1	X_{11}	X_{12}	X_{13}	...	X_{1n}
Gene 2	X_{21}	X_{22}	X_{23}	...	X_{2n}
Gene 3	X_{31}	X_{32}	X_{33}	...	X_{3n}
...	...	...	...	...	...
Gene G	X_{G1}	X_{G2}	X_{G3}	...	X_{Gn}

Edge	Present edge number	Absent edge number	Edge presence in prediction
(1,2)	5	0	✓
(1,3)	1	4	×
(1,4)	4	1	✓
(2,3)	3	2	✓
(2,4)	1	4	×
(3,4)	2	3	×

FIGURE 1 | An illustration of the two-step algorithm. **(A)** The input of the algorithm, including spatial coordinates of cells, the gene-cell expression matrix, and cell class information. Different shapes and colors represent different types of cells. **(B)** The two steps in the proposed algorithm. We first estimate cell-type covariance matrices, and then we use those estimates to refine cell-specific gene co-expression networks. **(C)** Gene co-expression network prediction based on the estimates from **(B)**.

matrix for each cell type, which serves as the "average" of the cell-specific covariance matrices in a given cell type (**Figure 1B**). In step two, for any given cell, we find its appropriate neighborhood and combine the cell label proportions in the neighborhood and the cell-type covariance matrices estimated in step one to assign an empirical prior to the covariance matrix of that cell. Subsequently, we apply the Bayes' rule to obtain the posterior mean estimates, transform it to the correlation matrix, and select a threshold to shrink absolute values of correlations less than it to zero, resulting in the cell's gene co-expression network (**Figure 1B**). After completing the estimates for each cell, we can further predict the network structures for a position where cells are not detected. We set a neighborhood of the location like in the estimation step two, and then an edge is present if and only if this edge appears more than or equal to half times among the gene networks of its neighboring cells (**Figure 1C**).

In the following, we introduce our proposed algorithm in detail in section 2. Section 3 provides the simulation study to compare the two-step algorithm against competing methods including traditional network construction methods (Zhang and Horvath, 2005) based on a group of cells and the cell-specific network construction approach (Dai et al., 2019). We use MERFISH data to demonstrate the good utility of the algorithm in section 4 and conclude the paper with a discussion in section 5.

2. METHOD

We first give some notations to clearly express the data preprocessing and our algorithm. Suppose that expression levels of G genes in n cells are measured and the expression of gene g in cell i is denoted by X_{gi}. We let $\mathbf{X} = (X_{gi})_{G\times n}$ represent the gene-cell expression matrix and use $\mathbf{X}_i$ to denote the *ith* column vector. The coordinates of cell i in the tissue section are denoted by (ℓ_i, h_i). We further assume that cells are from K distinct cell types and C_i indicates the membership of cell i. In other words, $C_i = k$ $(k = 1, \ldots, K)$ implies that cell i belongs to cell type k. Notice that the cell labels $\mathbf{C} = (C_1, \ldots, C_n)$ are assumed to be known in advance, and in case the cell label information is not available we can cluster cells using off-the-shelf single-cell expression tools. n_k is the cell number in cell type k, and $\mathbf{S}_k$ represents the index set $\{i : C_i = k\}$.

During data preprocessing, we need to normalize raw read count data to reduce the effects of different library sizes and other systematic biases. As we are interested in the pairwise gene correlations, the normalized expression values are further centered to zero and scaled to variance one within each cell type. If we still use X_{gi} to represent the normalized expression, then the transformed value is as follows. When $C_i = k$,

$$\widetilde{X}_{gi} = \frac{X_{gi} - \frac{1}{n_k}\sum_{j\in\mathbf{S}_k} X_{gj}}{\sqrt{\frac{1}{n_k-1}\sum_{j\in\mathbf{S}_k}(X_{gj} - \frac{1}{n_k}\sum_{j\in S_k} X_{gj})^2}}.$$

Next, we utilize the scaled expression matrix $\widetilde{\mathbf{X}} = (\widetilde{X}_{gi})_{G\times n}$ and its *ith* column vector $\widetilde{\mathbf{X}}_i$ in our algorithm.

In step one, we derive the sample expression covariance matrix for each cell type, which serves as the "average" of all cell-specific expression covariance matrices in that cell type and hence can be treated as an initial and coarse-grained estimate of the expression covariance matrix for each cell. Specifically, for cell type k, its sample expression covariance matrix is estimated by $\widehat{\mathbf{\Sigma}}^{(k)} := \frac{1}{n_k-1}\sum_{i\in\mathbf{S}_k}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T$ $(1 \le k \le K)$.

In step two, suppose the gene expression covariance matrix of cell i is denoted by $\mathbf{\Sigma}_i$. Biologically, $\mathbf{\Sigma}_i$ depends on both cell i's cell type as well as cell i's spatial circumstances. Taking this into account, we assume the following Bayesian statistical model for the observations,

$$\widetilde{\mathbf{X}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i) \quad (1)$$

$$\mathbf{\Sigma}_i \sim \mathcal{W}^{-1}(\mathbf{\Psi}_i, \nu), \quad (2)$$

where $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i)$ is a multivariate normal distribution with mean vector zero and covariance matrix Σ_i, and $\mathcal{W}^{-1}(\mathbf{\Psi}_i, \nu)$ is an inverse-Wishart distribution with scale matrix $\mathbf{\Psi}_i$ and ν degrees of freedom.

Equation (1) corresponds to the data-generating mechanism in which cell i's observation is sampled from its own distribution parameterized by $\mathbf{\Sigma}_i$. In the normal distribution, a zero element in $\mathbf{\Sigma}_i$ indicates that the corresponding two genes are independent, so $\mathbf{\Sigma}_i$ fully captures the gene co-expression network structure of cell i. Equation (2) reflects that we need to provide prior information for $\mathbf{\Sigma}_i$ to stabilize the estimate of $\mathbf{\Sigma}_i$; otherwise, only one sample is available, making the common maximal likelihood estimate very sensitive. We employ the inverse-Wishart distribution here as it is conjugate to the multivariate normal distribution (Gelman et al., 2013), which can enhance fast calculation of posterior estimates. Accordingly, we aim to borrow information from cell i's neighbors to define the hyper-parameter in the prior—the scale matrix $\mathbf{\Psi}_i$.

For each cell, we define its neighborhood as a square region with side length $2r$ and center at the location of the cell (**Figure 1B**). The choice of r depends on the cell density in the tissue section and our knowledge about the number of informative neighboring cells. We define the cell density as the ratio of the cell number (n) to the area where cells locate (A). As the area shape is often like a rectangle, we estimate A by $\hat{A} := (\max_i \ell_i - \min_i \ell_i)(\max_i h_i - \min_i h_i)$. If we believe that on average each cell has m_{info} informative neighboring cells, we then have the relationship $n/\hat{A} \times 4r^2 = m_{info}$, leading to $r = 0.5\sqrt{m_{info}\hat{A}/n}$. Based on our experience, we set $m_{info} = 70$ throughout our paper. Subsequently, we count the number of cells in this square region for each cell type and calculate proportions $(\omega_{i1}, \ldots, \omega_{iK})$ with $\omega_{ik} \ge 0$ and $\sum_{k=1}^{K}\omega_{ik} = 1$, where ω_{ik} is the proportion of type k cells in the neighborhood of cell i.

Next, we assign the weighted value $\sum_{k=1}^{K}\omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)}$ to the prior mean of $\mathbf{\Sigma}_i$, which is $\mathbf{\Psi}_i/(\nu - G - 1)$, resulting in the scale matrix $\mathbf{\Psi}_i = (\nu - G - 1)\sum_{k=1}^{K}\omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)}$. This prior reflects the information of nearby cells and helps stabilize the estimate of $\mathbf{\Sigma}_i$. We remark that the choice of the hyper-parameter $\mathbf{\Psi}_i$ depends on the data we are analyzing, so strictly speaking the approach is not fully Bayesian (Gelman et al., 2013).

Given the assigned prior, we estimate $\mathbf{\Sigma}_i$ by the posterior mean,

$$\widehat{\mathbf{\Sigma}}_i := \mathbb{E}(\mathbf{\Sigma}_i|\widetilde{\mathbf{X}}_i) = \frac{1}{\nu - G}(\mathbf{\Psi}_i + \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T)$$
$$= \frac{1}{\nu - G}((\nu - G - 1)\sum_{k=1}^{K}\omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)} + \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T),$$

where we set ν to $2G$ depending on the number of genes. $\widehat{\mathbf{\Sigma}}_i$ is then transformed to its corresponding correlation matrix $\widehat{\mathbf{R}}_i = \text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}\widehat{\mathbf{\Sigma}}_i\text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}$, where $\text{diag}(\widehat{\mathbf{\Sigma}}_i)$ is a diagonal matrix with diagonal elements the same as those of $\widehat{\mathbf{\Sigma}}_i$. Finally, we select a threshold d ($0 < d < 1$), and if the (g_1, g_2) element of the matrix $\widehat{\mathbf{R}}_i$, $\widehat{R}_{i,g_1g_2}$, has an absolute value larger than d, then we believe there is an edge between gene g_1 and g_2 in the gene co-expression network of cell i. Algorithm 1 displays the two-step estimation procedure.

Algorithm 1: Two-step gene co-expression network estimation.

Input: normalized gene expression matrix $\mathbf{X}$, cell labels $\mathbf{C}$, cell coordinates (ℓ_i, h_i), $1 \le i \le n$, and hyper-parameters (m_{info}, ν, d).
Output: cell-specific gene co-expression networks $\mathcal{G}_i$ $(1 \le i \le n)$.
Preprocessing:
for *i in* $1:n$ **do**
  for *g in* $1:G$ **do**
    $\widetilde{X}_{gi} = \frac{X_{gi} - \frac{1}{n_k}\sum_{j\in\mathbf{S}_k} X_{gj}}{\sqrt{\frac{1}{n_k-1}\sum_{j\in\mathbf{S}_k}(X_{gj} - \frac{1}{n_k}\sum_{j\in\mathbf{S}_k} X_{gj})^2}}$ when $C_i = k$
  end
end
Step 1: Obtain cell-type-specific covariance matrix:
for *k in* $1:K$ **do**
  $\widehat{\mathbf{\Sigma}}^{(k)} = \frac{1}{n_k-1}\sum_{i\in\mathbf{S}_k}\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T$
end
Step 2: Estimate cell-specific gene co-expression networks.
for *i in* $1:n$ **do**
  $\widehat{\mathbf{\Sigma}}_i = \frac{1}{\nu-G}((\nu - G - 1)\sum_{k=1}^{K}\omega_{ik}\widehat{\mathbf{\Sigma}}^{(k)} + \widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}_i^T)$
  $\widehat{\mathbf{R}}_i = \text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}\widehat{\mathbf{\Sigma}}_i\text{diag}(\widehat{\mathbf{\Sigma}}_i)^{-1/2}$
  for g_1 *in* $1:G$ **do**
    for g_2 *in* $(g_1+1):G$ **do**
      $\mathcal{G}_{i,g_1g_2} = \begin{cases} 0 \text{ if } |\widehat{\mathbf{R}}_{i,g_1g_2}| < d \\ 1 \text{ if } |\widehat{\mathbf{R}}_{i,g_1g_2}| \ge d \end{cases}$
    end
  end
end

After completing the network structure estimates for all cells, we can take advantage of the estimates to predict the gene co-expression network for any missing cell with a position in the studied tissue section area. If we are interested in an undetected cell at a new location (ℓ^*, h^*), its gene co-expression network is constructed as follows. We first find all detected cells in the neighborhood of (ℓ^*, h^*), and then we believe an edge between genes g_1 and g_2 in the prediction if there are more connections than disconnections for this pair of genes among the gene networks of (ℓ^*, h^*)'s neighboring detected cells. Algorithm 2 shows the steps of making gene co-expression network predictions.

Algorithm 2: Gene co-expression network prediction for a new cell position.

Input: Gene network estimates from Algorithm 1, cell coordinates (ℓ_i, h_i) for $1 \le i \le n$, hyper-parameter m_{info}, and a new cell position (ℓ^*, h^*).
Output: Gene co-expression network $\mathcal{G}_{(\ell^*,h^*)}$ for cell (ℓ^*, h^*).
Step 1: Find all cells in the neighborhood of (ℓ^*, h^*), denoted by $\text{Nei}_{(\ell^*,h^*)} := \{i \in \{1, 2, \ldots, n\} : (\ell_i, h_i)$ is in the neighborhood of $(\ell^*, h^*)\}$.
Step 2: Obtain gene co-expression network for cell (ℓ^*, h^*):
for g_1 *in* $1:G$ **do**
  for g_2 *in* $(g_1+1):G$ **do**
    $\mathcal{G}_{(\ell^*,h^*),g_1g_2} =$
    $\begin{cases} 0 \text{ if } \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 0\} > \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 1\} \\ 1 \text{ if } \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 0\} \le \#\{j \in \text{Nei}_{(\ell^*,h^*)} : \mathcal{G}_{j,g_1g_2} = 1\} \end{cases}$
  end
end

3. SIMULATION STUDY

In this section, we used simulated data to evaluate the performance of the proposed two-step algorithm. We set the gene number $G = 100$, the cell-type number $K = 5$, and the cell number for each cell type $(n_1, n_2, n_3, n_4, n_5) = (394, 373, 428, 274, 529)$. We chose a rectangle area as the tissue section with length $L = 1{,}000$ and width $H = 750$, where a total of $n = \sum_{k=1}^{K} n_k = 1998$ cells distribute on the section and display clear spatial patterns (**Figure 2A**). For example, cells from cell-type 5 concentrate on the left side, while cells from cell-type 1 enrich on the right side.

We then generated cell-type-specific covariance matrices $\Sigma^{(k)}$ for $k = 1, \ldots, K$. Genes that work together often form a gene module, which can exhibit a block structure in the covariance matrix. Hence, the covariance matrix of each cell type was set as a block diagonal matrix, where each block was a 20×20 positive definite matrix. Five different modules were used for this purpose and were as follows.

- In module 1 ($\mathcal{M}_1$), its (i, j) element $\sigma_{ij} = \rho^{|i-j|} + 0.5\mathbf{I}(i = j)$ for $1 \le i \le 20$ and $1 \le j \le 20$, where $\mathbf{I}(A)$ is an indicator function of event A. We took $\rho = 0.7$.

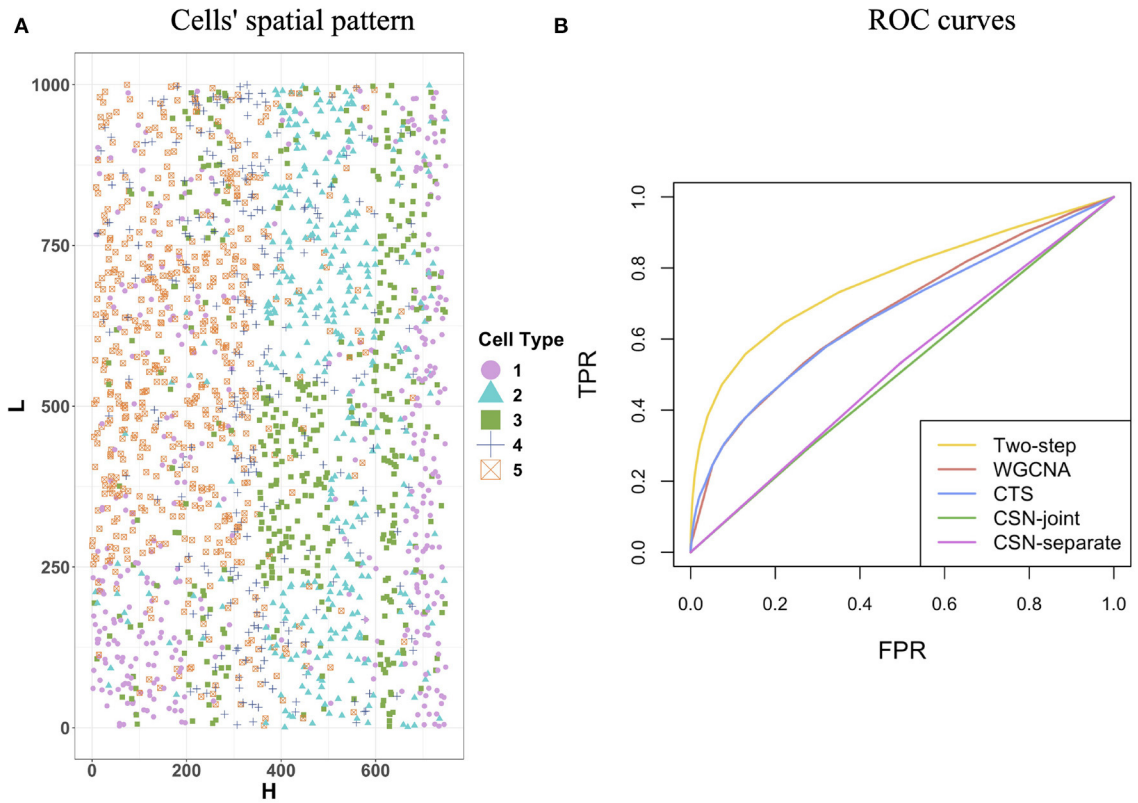


FIGURE 2 | Cells' spatial pattern and performance comparisons on the network structure recovery. **(A)** Cells' spatial pattern. Different shapes and colors correspond to different types of cells. **(B)** ROC curves of the two-step algorithm, WGCNA, CTS, CSN-joint, and CSN-separate.

- In module 2 ($\mathcal{M}_2$), $\sigma_{ij} = (1 - \frac{|i-j|}{10})_+$, which forms a banded matrix. The function $(x)_+$ equals x for $x \geq 0$ and zero for $x < 0$.
- In module 3 ($\mathcal{M}_3$), $\sigma_{ij} = \rho \mathbf{I}(|i-j| = 1) + 1.3\mathbf{I}(i = j)$ for $\rho = -0.3$.
- In module 4 ($\mathcal{M}_4$), $\sigma_j = (1 - \frac{|i-j|}{k})_+$, where $k = \lfloor G/2 \rfloor$.
- In module 5 ($\mathcal{M}_5$), the block was $F + \epsilon I_{20\times 20}$. $I_{20\times 20}$ is an identity matrix. $F = (f_{ij})_{20\times 20}$ is a symmetric matrix with independent upper triangle elements $f_{ij} = unif(-0.2, 0.8) \times Ber(1, 0.2)$, where $unif(-0.2, 0.8)$ is a random variable uniformly distributed on$(-0.2, 0.8)$, and $Ber(1, 0.2)$ is a Bernoulli random variable with the success probability 0.2. We set $\epsilon = \max\{-\lambda_{\min}(F), 0\} + 0.01$ to ensure that B is positive definite, where $\lambda_{\min}(F)$ is the smallest eigenvalue of F.

If we denote a block diagonal matrix with diagonal blocks being $\mathcal{M}_{i_1}, \mathcal{M}_{i_2}, \mathcal{M}_{i_3}, \mathcal{M}_{i_4}, \mathcal{M}_{i_5}$ in the order from the upper left to the lower right by $(\mathcal{M}_{i_1}, \mathcal{M}_{i_2}, \mathcal{M}_{i_3}, \mathcal{M}_{i_4}, \mathcal{M}_{i_5})$, then we specify $\mathbf{\Sigma}^{(1)}$=$(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5)$, $\mathbf{\Sigma}^{(2)}$=$(\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_2, \mathcal{M}_4, \mathcal{M}_5)$, $\mathbf{\Sigma}^{(3)}$=$(\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_2, \mathcal{M}_5, \mathcal{M}_4)$, $\mathbf{\Sigma}^{(4)}$=$(\mathcal{M}_3, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_5, \mathcal{M}_4)$, and $\mathbf{\Sigma}^{(5)}$=$(\mathcal{M}_3, \mathcal{M}_2, \mathcal{M}_5, \mathcal{M}_1, \mathcal{M}_4)$.

Next, we generated the cell-specific gene expression covariance matrix for each cell i. We first obtained the neighborhood of cell i using $r = 80$, then calculated cell-type proportions $q_{ik}, 1 \leq k \leq K$ in the neighborhood, and sampled Σ_i from the inverse-Wishart distribution $\mathcal{W}^{-1}(\sum_{k=1}^{K} q_{ik}\mathbf{49}\Sigma^{(k)}, G + 50)$. Moreover, to make the network sparse and covariance matrix positive definite, non-diagonal elements in the $\mathbf{\Sigma}_i$ with absolute values less than 0.5 were shrunk to zero, and the diagonal elements in Σ_i were added by five. Finally, we sampled the observed gene-cell expression matrix $\mathbf{X}_i = (X_{1i}, \ldots, X_{Gi})^T$ from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma_i)$ for $1 \leq i \leq n$.

To show the advantage of our algorithm in estimating cell-specific gene expression matrix, we compared it against the weighted gene co-expression network analysis (denoted by WGCNA) (Zhang and Horvath, 2005), the traditional hard-thresholding cell-type-specific network estimation approach (denoted by CTS), and the cell-specific gene network estimation method that does not make use of cell spatial information (denoted by CSN, Dai et al., 2019). Specifically, in WGCNA, we first calculated pairwise gene expression similarity using the absolute values of Pearson correlations, then utilized the "soft" power adjacency function to convert the similarity matrix, and finally obtain the topological overlap matrix based on the adjacency matrix. Regarding CTS, we used the cell-type-level gene network as the estimate for each cell in that cell type. For CSN, we adopted two versions: in the joint version (CSN-joint), we used the gene-cell expression matrix for all cells as the input of the CSN method; and in the separate version (CSN-separate), we

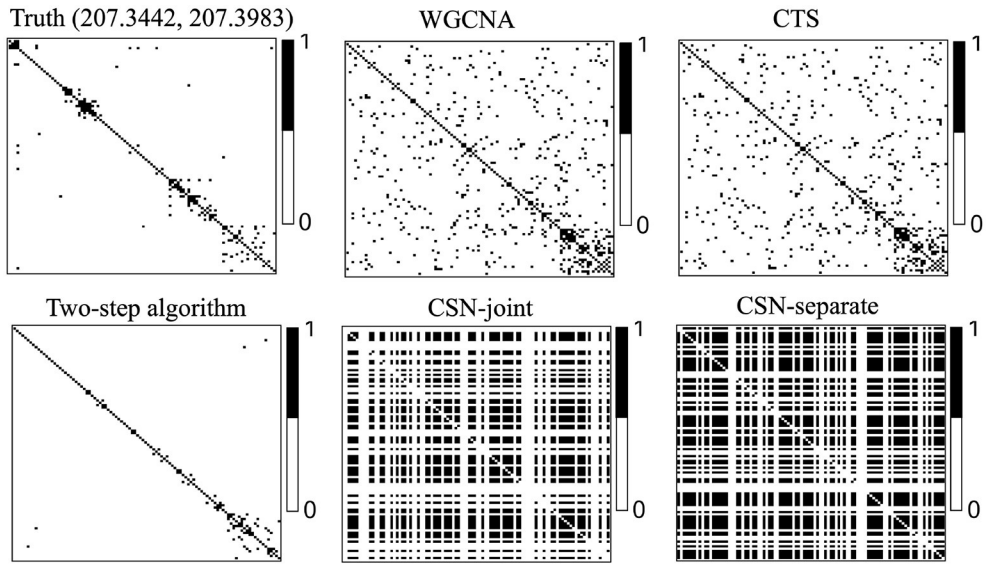


FIGURE 3 | Heatmaps of estimated gene co-expression networks for different methods and underlying truth. The network heatmaps of the cell on location (207.3442, 207.3983) was used for illustration.

only input the gene-cell expression matrix for cells coming from one cell type, repeat the procedure for each cell type, and also obtain cell-specific network estimates. In other words, for CSN-separate, the estimations for one cell only rely on the information of cells from the same cell type.

Figure 2B provides the receiver operating characteristic (ROC) curves for network structure recovery of the proposed algorithm (denoted by two-step algorithm) and other four competing approaches (WGCNA, CTS, CSN-joint, CSN-separate). The horizontal axis represents the false positive rate (FPR), which equals the ratio of the number of edges that were wrongly detected by the method for all cells to the number of absent edges in the underlying true networks for all cells, while the vertical axis corresponds to the true positive rate (TPR), describing the ratio of the number of edges that were correctly detected by the method for all cells to the number of edges in the underlying true networks for all cells. It is observed that the ROC curve of our algorithm is uniformly over the ROC curves of the other four approaches, indicating that given any FPR the TPR of the proposed algorithm is always higher than that of the other four competing methods. As WGCNA also estimates cell-type-specific networks, it does not outperform our algorithm but is slightly better than traditional CTS.

Figure 3 displays heatmaps of gene co-expression matrix of the cell with the coordinates (207.3442, 207.3983), both true and estimated gene co-expression matrix by two-step algorithm, WGCNA, CTS, CSN-joint, and CSN-separate are shown (the results of CSN-separate are similar to CSN-joint's). From **Figure 3**, we can observe that our two-step algorithm outperforms the other four methods in estimating cell-specific gene co-expression networks. To further quantify the network recovery error for these methods, we used the following error term $E: = 1/n \cdot \sum_{i=1}^{n} \sum_{g_1<g_2} |\mathcal{G}_{i,g_1g_2} - \mathcal{G}^{\text{true}}_{i,g_1g_2}|$. For WGCNA, we chose the truncation value 0.0001 for the topological overlap matrix; for the proposed algorithm and CTS, the threshold d for the gene-gene correlations was chosen as 0.1; for the two CSN methods, the significance level was set at 0.01. **Table 1** shows the errors based on ten replicates and indicates that the proposed method is more accurate than the others in terms of the network structure recovery.

TABLE 1 | Mean errors and corresponding standard deviations of five methods.

Methods	Two-step algorithm	WGCNA	CTS	CSN-joint	CSN-separate
Mean error	288.62	484.05	484.44	1869.60	2462.09
(standard deviation)	(14.15)	(18.78)	(18.80)	(347.49)	(95.57)

The degree of a gene is the number of edges connected to that gene. We investigated the degree distributions of the estimated cell-specific gene co-expression network and compared it to truth and other competing approaches on one gene for each cell type. **Figures 4A–E** show the violin plots of the degrees of gene 91 for each cell type. We can see that the distribution created by our proposed algorithm is much closer to the underlying truth than CSN-separate and CSN-joint, while WGCNA and CTS's distributions are just horizontal line segments as their network estimates are identical for all cells in one cell type.

Figure 4F shows the violin plot for the computation time in second for these methods based on ten replicates. It is reasonable that WGCNA and CTS have the minimum computing time as they only estimate K cell-type-specific gene-gene network, but their performances are obviously not good. The proposed algorithm has a similar computing time to CSN-separate

and is faster than CSN-joint. Hence, our algorithm not only performs well in estimating networks but also has relatively fast computing.

Given the network estimates by our method, we can easily use algorithm 2 to predict network structures for a new location. We randomly generated 50 new coordinates as the locations of 50 missing cells, simulated the true gene network of these 50 new cells following the data-generating procedure above, and then applied the prediction algorithm. The prediction error is 347.84 (in terms of E). WGCNA, CTS, CSN-joint, and CSN-separate do not have the ability to predict gene co-expression networks of missing cells, so the proposed algorithm provides an extra important function to make network predictions.

4. REAL APPLICATION

4.1. MERFISH Mouse Hypothalamus Data

Moffitt et al. (2018) combined single-cell RNA-sequencing and a single-cell transcriptome imaging method called MERFISH to obtain expression profiles at the cellular level as well as x-y coordinates of centroid positions for cells in the mouse hypothalamic preoptic region. In the MERFISH mouse hypothalamus data, class information of cells are also available. The single-cell spatial expression data can be downloaded from https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248.

We chose the expression data with animal id 35 and location 0.26 of the slice in bregma coordinates and removed cells labeled

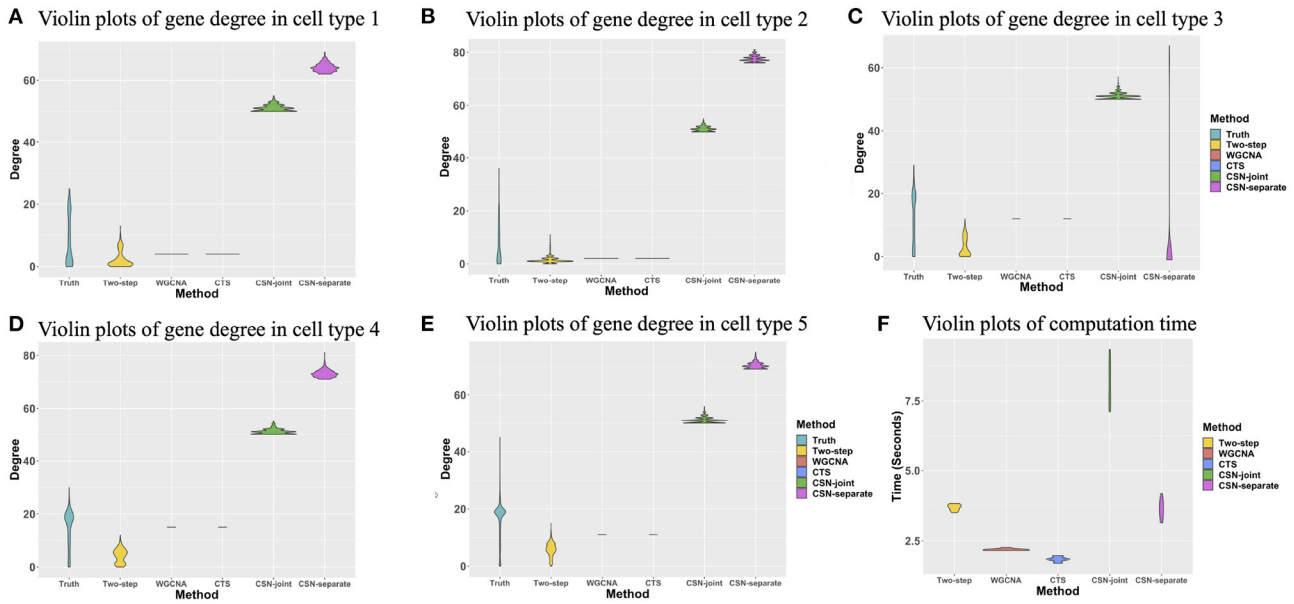


FIGURE 4 | (A–E) Violin plots of gene degrees of gene 91 in the five cell types in the simulation study. **(F)** Violin plots of the computational time (in seconds) for the five approaches based on ten replicates.

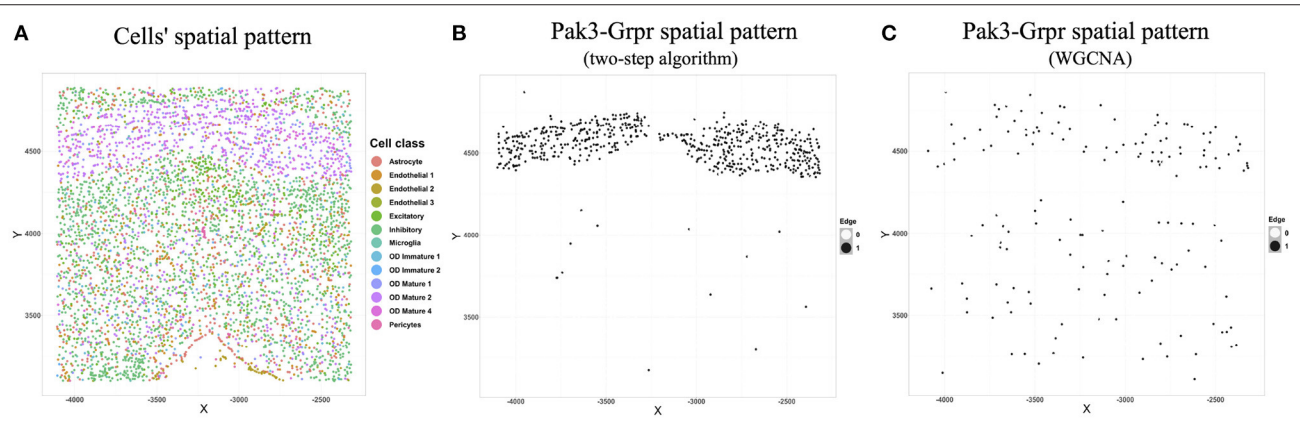


FIGURE 5 | (A) Cells' spatial distribution pattern, where different colors of points correspond to different cell classes. **(B)** The spatial pattern of the Pak3-Crpr connection obtained by two-step algorithm. The black point reflects that an edge exists between "Pak3" and "Crpr" in that cell. **(C)** The spatial pattern of the Pak3-Crpr connection obtained by WGCNA.

"Ambiguous" as well as cell types that contain less than 10 cells, resulting in 13 cell classes. The spatial pattern of the selected cells was displayed in **Figure 5A**. We further removed "blank" genes and genes whose expressions are zero across all the cells in one cell type, resulting in $G = 147$ genes and $n = 4,682$ cells. Subsequently, we applied the proposed two-step algorithm with informative neighboring cell number $m_{info} = 70$ and threshold parameter $d = 0.1$. We randomly selected two cells from cell classes "inhibitory neurons" and "excitatory neurons," respectively, and the gene co-expression networks of the two cells were shown in **Figure 6**. It is observed that the two gene co-expression networks have similar functional gene modules on the

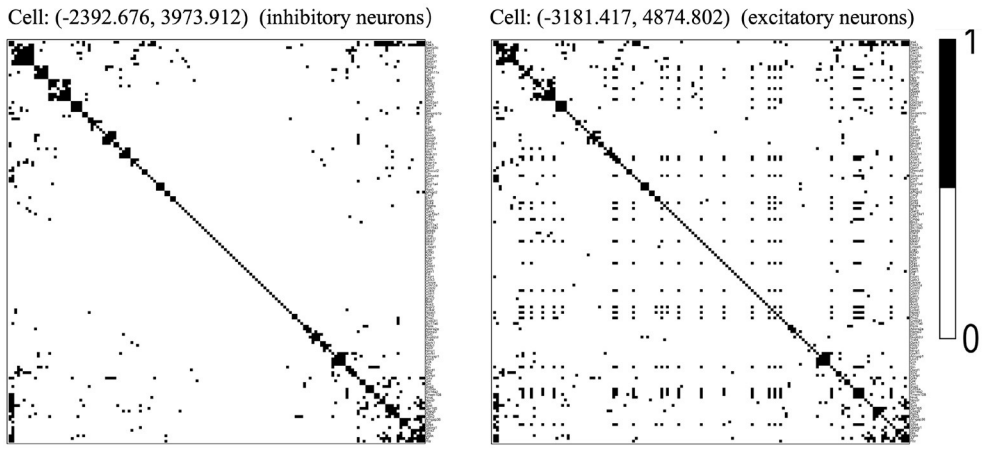


FIGURE 6 | Estimated gene co-expression networks of two selected cells in the MERFISH mouse hypothalamus data: the left panel corresponds to an inhibitory cell, while the right panel corresponds to an excitatory cell.

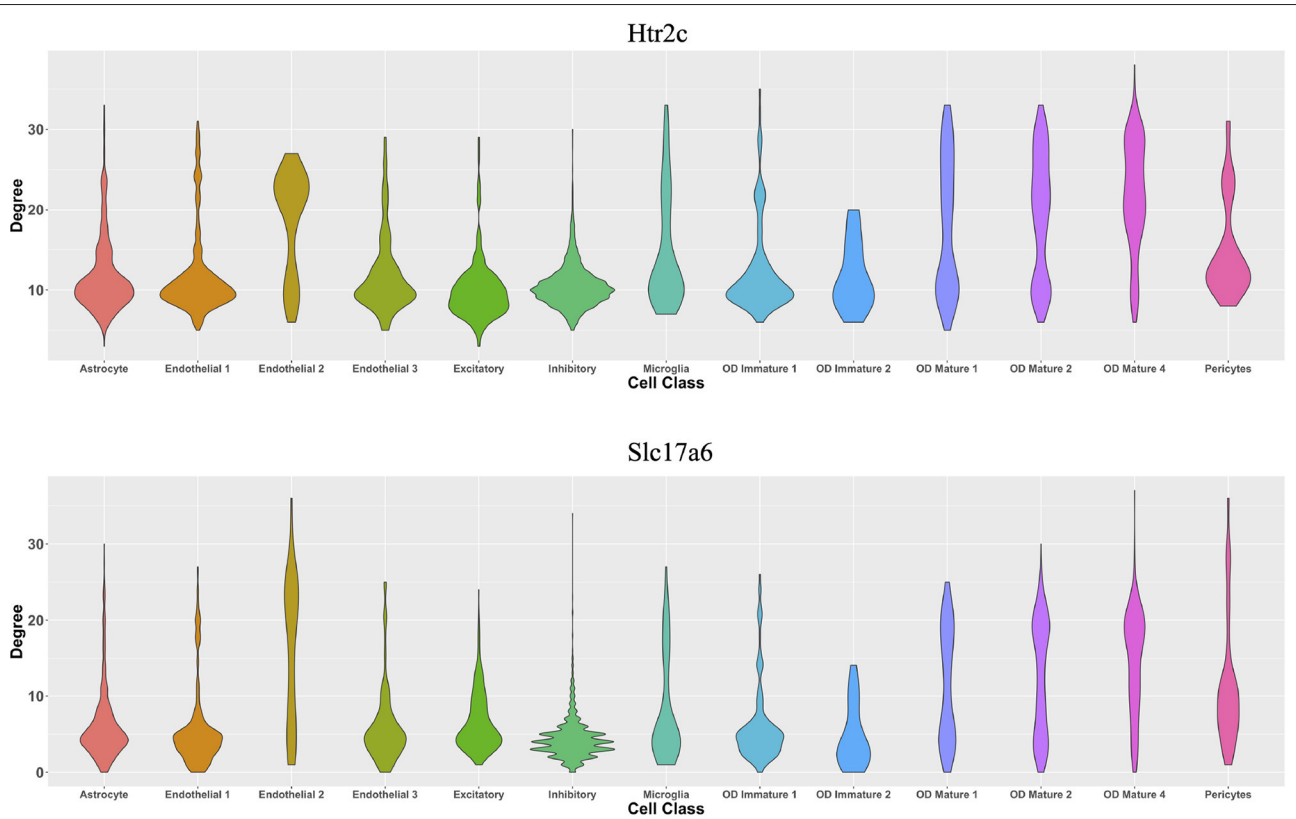


FIGURE 7 | Violin plots of two genes' degree distributions across thirteen cell classes in the MERFISH mouse hypothalamus data.

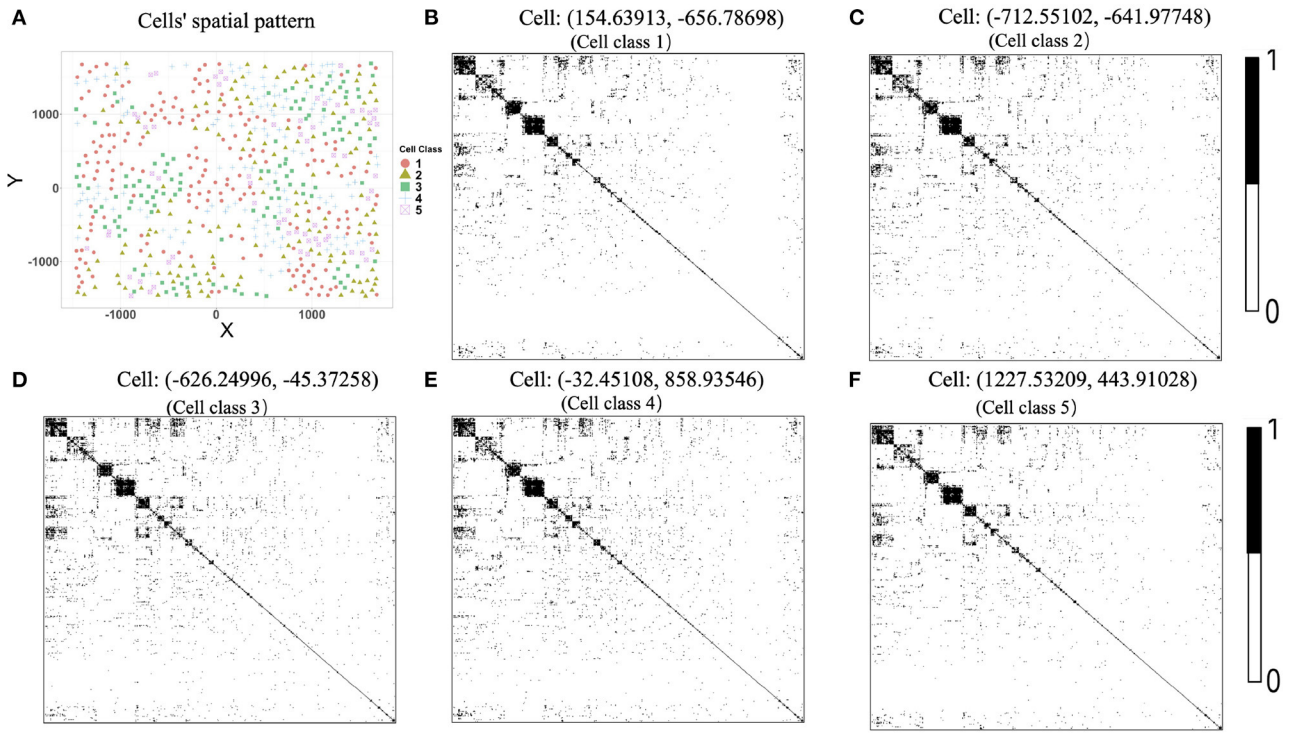


FIGURE 8 | (A) Cells' spatial distribution pattern, where different colors and shapes of points correspond to different cell classes. **(B–F)** Estimated gene co-expression networks of five randomly selected cells in the MERFISH U-2 OS cell line data.

diagonal possibly because both of them are neurons. Moreover, the network of the cell in excitatory neurons is denser than the network in inhibitory neurons, and the reason may be that the gene activity in cells controlling excitement is more active than that in cells controlling inhibition.

Cell-specific gene co-expression networks can provide insightful information about how genes' degrees vary in each cell type. To show that, in excitatory neuron cells, we selected 15 genes with the most variable degrees: Sln, Baiap2, Tmem108, Oprk1, Slc17a6, Nos1, Htr2c, Irs4, Gpr165, Slc18a2, Vgf, Pgr, Ar, Gabrg1, and Gabra1. To validate the functions of the gene set, we conducted gene set enrichment analysis (Subramanian et al., 2005) based on the gene ontology (GO) database (Gene Ontology Consortium, 2004). We found several significant annotations related to the excitatory neurons including GO_MODULATION_OF_EXCITATORY_POSTSYNAPTIC_POTENTIAL(biologicalprocess),GO_EXCITATORY_SYNAPSE (cellular component), and GO_NEURON_PROJECTION (cellular component). In terms of the inhibitory neurons, we identified 15 genes with the most variable degrees: Baiap2, Sox6, Irs4, Ar, Gda, Oprk1, Isl1, Cyr61, Prlr, Glra3, Gabra1, Dgkk, Tmem108, Sln, and Ano3. Using GO annotations, the gene set is associated with inhibitory neurons-related activities including GO_INHIBITORY_EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY (molecular function) and GO_NEURON_PROJECTION (cellular component). These observations show that estimated cell-specific networks have the potential to find genes with variable degrees for each cell type, which cannot be accomplished by cell-type-specific approaches.

We next illustrated the spatial feature of estimated gene co-expression networks in terms of gene-gene connections. We calculated the median degree for each gene. Gene Pak3 with the maximum median degree (31) and gene Grpr with the minimum median degree (0) were chosen for demonstration. **Figure 5B** shows that the Pak3-Grpr connection mainly appears in the region where "mature oligodendrocytes" are enriched. The observation indicates that the two genes may tend to work together in the mature oligodendrocytes. Actually, mutations on gene Pak3 are related to intellectual disability diseases, and its expression decreases in mature oligodendrocytes and may regulate oligodendrocyte precursor cell differentiation, as reported in a previous study (Renkilaraj et al., 2017). To demonstrate the advantage of estimating cell-specific networks, we further applied WGCNA (Zhang and Horvath, 2005) with truncation level 0.1 to obtain cell-type-specific networks. However, **Figure 5C** indicates that the cell-type-specific estimations by WGCNA cannot reveal the pattern provided by cell-specific estimations.

From the perspective of cell types, **Figure 7** demonstrates the cell-type-specific degree distributions of two genes, Htr2c and Slc17a6 (Campbell et al., 2017; Chen et al., 2017), which have the most degree variances across cells. It is observed that the degree distribution of one gene varies across cell types, and

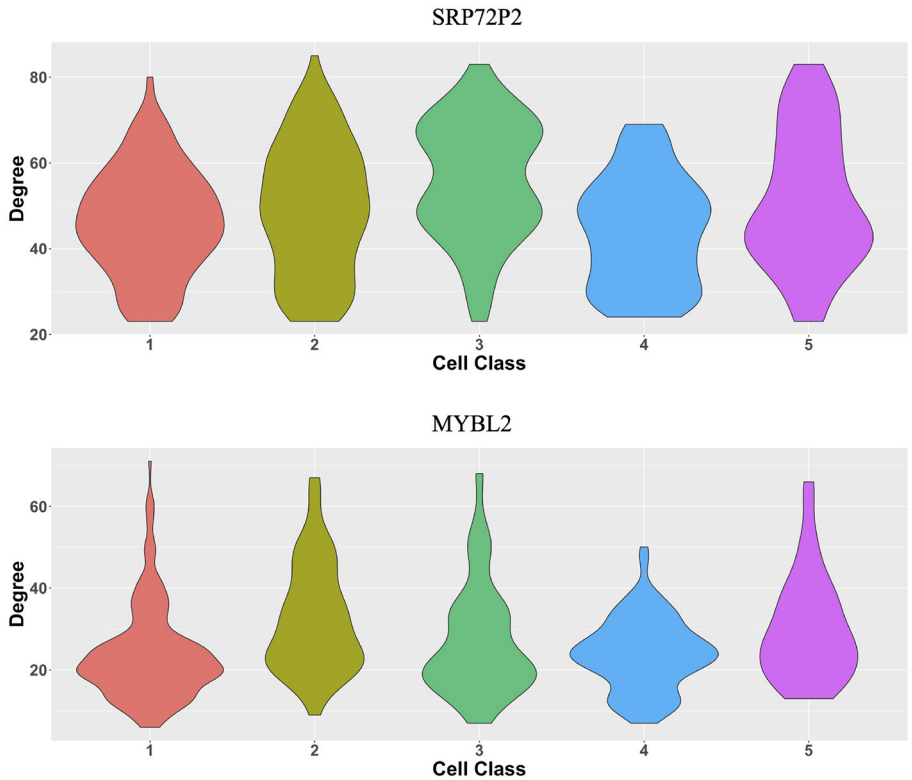


FIGURE 9 | Violin plots of two genes' degree distributions across five cell classes in the MERFISH U-2 OS cell line data.

this cannot be observed by traditional cell-type-specific gene co-expression networks.

4.2. MERFISH U-2 OS Data

We further provided some simple results of the proposed algorithms on another single-cell spatial expression dataset. Xia et al. (2019) carried out the MERFISH experiments on human osteosarcoma (U-2 OS) cells, and we downloaded the expression count data from https://www.pnas.org/content/116/39/19490/tab-figures-data. The data contain expression profiles for 10,050 genes and 1,368 cells in three batches. To avoid possible influences caused by batch effects, our analysis focuses on the batch one. We first removed "blank" genes, resulting in $n = 645$ cells and $G = 10,050$ genes. Since there is no cell-type annotation information, we first performed cell clustering procedure using Seurat (Butler et al., 2018; Stuart et al., 2019). By setting the resolution at 0.8 in Seurat clustering procedure, we obtained $K = 5$ cell classes, which is consistent with the cell type number in Xia et al. (2019). **Figure 8A** shows the cells' spatial distribution.

The original expression data were count data, so we normalized the data following the formula $x_{gi} \leftarrow \frac{10^6}{\sum_g x_{gi}} x_{gi}$, where x_{gi} is the expression level of gene g in cell i and then selected the most variable 500 genes to perform the proposed two-step algorithm. The informative neighboring cell number m_{info} was set to 70, and the threshold parameter d was set to 0.3. Accordingly, we randomly selected five cells from the five cell classes, respectively, and the gene co-expression networks of the five chosen cells were shown in **Figures 8B–F**. It is observed that the five gene networks from different cell types have similar gene modules. Moreover, we showed the degree distributions across five cell types for two genes, SRP72P2 and MYBL2, which have the most degree variances across cells. **Figure 9** tells us that the degree distributions of the two genes not only have variation within one cell type but also change from one cell type to another.

5. DISCUSSION

Recent technology advances enable us to gain deep insights into spatial cell-specific gene expressions. In this paper, we developed a simple and computationally efficient two-step algorithm to recover spatially-varying cell-specific gene co-expression networks. The simulation study shows that the proposed algorithm outperforms the traditional cell-type-specific gene network approach and cell-specific gene network estimation methods that do not employ spatial information. The application to the MERFISH data provides some interesting biological findings. In the meanwhile, there are some limitations in the proposed

algorithm we aim to improve in the future work. For example, we choose a hard threshold to identify a gene-gene connection, but an adaptive threshold selection needs to be derived.

We also acknowledge that using normal distributions to fit normalized gene expression data can lose power and be suboptimal compared to directly modeling the sequencing count data via Poisson distributions (Sun et al., 2017). Fortunately, in several previous bioinformatics works, using continuous multivariate normal distributions to model normalized single-cell sequencing data (Pierson and Yau, 2015; Chen and Zhou, 2017; Wang et al., 2020) or spatial single-cell expression data (Li D. et al., 2020) can still provide key biological findings. Moreover, in terms of computation, multivariate Poisson distributions (Karlis, 2003) largely increase the computational burden. Statistically, the covariance matrix in the multivariate Poisson distribution does not have a standard conjugate prior, thus failing to obtain an analytical form of the posterior mean. In real data, the cell number is often large (∼4,000 in our real application), which actually guarantees a satisfying normal approximation. Considering these issues, we chose the multivariate normal as the data distribution, but it is very interesting and challenging to extend the algorithm to directly model raw count data and we leave it for future work.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. MERFISH mouse hypothalamus data can be downloaded from https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248, and MERFISH U-2 OS cell line data is available via the link https://www.pnas.org/content/116/39/19490/tab-figures-data.

CODE AVAILABILITY STATEMENT

The codes that can reproduce results in simulation and real application are available on GitHub, https://github.com/jingeyu/CSSN_data_code. The associated CSSN package is available on GitHub, https://github.com/jingeyu/CSSN.

AUTHOR CONTRIBUTIONS

XL conceived the study. JY and XL developed the method, analyzed the real data, and wrote the paper. JY implemented the algorithm, prepared the software, and conducted simulation. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (11901572), the start-up research fund at Renmin University of China, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China (19XNLG08), and the fund for building world-class universities (disciplines) of Renmin University of China.

ACKNOWLEDGMENTS

We are very grateful to the Editor, Associate Editor, and reviewers for their constructive comments which greatly improve the paper. We also thank the High-performance Computing Platform of Renmin University of China for providing computing resources.

REFERENCES

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Butte, A., and Kohane, I. (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (Honolulu, HI), 418–429.

Campbell, J. N., Macosko, E. Z., Fenselau, H., Pers, T. H., Lyubetskaya, A., Tenen, D., et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* 20, 484–496. doi: 10.1038/nn.4495

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20, 2242–2250. doi: 10.1093/bioinformatics/bth234

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348:aaa6090. doi: 10.1126/science.aaa6090

Chen, M., and Zhou, X. (2017). Controlling for confounding effects in single cell RNA sequencing studies using both control and target genes. *Sci. Rep.* 7, 1–14. doi: 10.1038/s41598-017-13665-w

Chen, R., Wu, X., Jiang, L., and Zhang, Y. (2017). Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Rep.* 18, 3227–3241. doi: 10.1016/j.celrep.2017.03.004

Dai, H., Li, L., Zeng, T., and Chen, L. (2019). Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res.* 47:e62. doi: 10.1093/nar/gkz172

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Gene Ontology Consortium (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32(Suppl_1), D258–D261. doi: 10.1093/nar/gkh036

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *J. Appl. Stat.* 30, 63–77. doi: 10.1080/0266476022000018510

Köster, J., Brown, M., and Liu, X. S. (2019). A Bayesian model for single cell transcript expression analysis on MERFISH data. *Bioinformatics* 35, 995–1001. doi: 10.1093/bioinformatics/bty718

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., et al. (2014). Highly multiplexed subcellular RNA sequencing *in situ*. *Science* 343, 1360–1363. doi: 10.1126/science.1250212

Li, D., Ding, J., and Bar-Joseph, Z. (2020). Identifying signaling genes in spatial single cell expression data. *Bioinformatics*. doi: 10.1101/2020.07.27.221465. [Epub ahead of print].

Li, L., Dai, H., Fang, Z., and Chen, L. (2020). CCSN: single cell RNA sequencing data analysis by conditional cell-specific network. *bioRxiv [Preprint]*. doi: 10.1101/2020.01.25.919829

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell *in situ* RNA profiling by sequential hybridization. *Nat. Methods* 11:360. doi: 10.1038/nmeth.2892

Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Perez, J. D., et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 362:eaau5324. doi: 10.1126/science.aau5324

Pierson, E., and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16, 1–10. doi: 10.1186/s13059-015-0805-z

Renkilaraj, M. R. L. M., Baudouin, L., Wells, C. M., Doulazmi, M., Wehrlé, R., Cannaya, V., et al. (2017). The intellectual disability protein PAK3 regulates oligodendrocyte precursor cell differentiation. *Neurobiol. Dis.* 98, 137–148. doi: 10.1016/j.nbd.2016.12.004

Serin, E. A., Nijveen, H., Hilhorst, H. W., and Ligterink, W. (2016). Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.* 7:444. doi: 10.3389/fpls.2016.00444

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255. doi: 10.1126/science.1087447

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. III., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902. doi: 10.1016/j.cell.2019.05.031

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sun, S., Hood, M., Scott, L., Peng, Q., Mukherjee, S., Tung, J., et al. (2017). Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res.* 45:e106. doi: 10.1093/nar/gkx204

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* 17, 193–200. doi: 10.1038/s41592-019-0701-7

Tian, J., Wang, J., and Roeder, K. (2021). ESCO: single cell expression simulation incorporating gene co-expression. *Bioinformatics*. doi: 10.1093/bioinformatics/btab116. [Epub ahead of print].

Wang, J., Devlin, B., and Roeder, K. (2020). Using multiple measurements of tissue to estimate subject-and cell-type-specific gene expression. *Bioinformatics* 36, 782–788. doi: 10.1093/bioinformatics/btz619

Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361:eaat5691. doi: 10.1126/science.aat5691

Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 116, 19490–19499. doi: 10.1073/pnas.1912459116

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:Article17. doi: 10.2202/1544-6115.1128

Zhang, M., Sheffield, T., Zhan, X., Li, Q., Yang, D. M., Wang, Y., et al. (2020). Spatial molecular profiling: platforms, applications and analysis tools. *Brief. Bioinform.* doi: 10.1093/bib/bbaa145. [Epub ahead of print].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 06 May 2021
doi: 10.3389/fgene.2021.654821

BTOB: Extending the Biased GWAS to Bivariate GWAS

Junxian Zhu[1†], *Qiao Fan*[2†], *Wenying Deng*[3], *Yimeng Wang*[1] *and Xiaobo Guo*[1*]

[1] Department of Statistical Science, School of Mathematics, Sun Yat-sen University, Guangzhou, China, [2] Center for Quantitative Medicine, Duke-National University of Singapore Medical School, Singapore, Singapore, [3] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, United States

In recent years, a number of literatures published large-scale genome-wide association studies (GWASs) for human diseases or traits while adjusting for other heritable covariate. However, it is known that these GWASs are biased, which may lead to biased genetic estimates or even false positives. In this study, we provide a method called "BTOB" which extends the biased GWAS to bivariate GWAS by integrating the summary association statistics from the biased GWAS and the GWAS for the adjusted heritable covariate. We employ the proposed BTOB method to analyze the summary association statistics from the large scale meta-GWASs for waist-to-hip ratio (WHR) and body mass index (BMI), and show that the proposed approach can help identify more susceptible genes compared with the corresponding univariate GWASs. Theoretical results and simulations also confirm the validity and efficiency of the proposed BTOB method.

Keywords: GWAS, bivariate GWAS, summary association statistics, heritable covariate, biased

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong, Hong Kong

Reviewed by:
Lin Hou,
Tsinghua University, China
Dajiang Liu,
Pennsylvania State University, United States

***Correspondence:**
Xiaobo Guo
guoxb3@mail.sysu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *17 January 2021*
Accepted: *07 April 2021*
Published: *06 May 2021*

Citation:
Zhu J, Fan Q, Deng W, Wang Y and Guo X (2021) BTOB: Extending the Biased GWAS to Bivariate GWAS. Front. Genet. 12:654821. doi: 10.3389/fgene.2021.654821

1. INTRODUCTION

Genome-wide association studies (GWASs) have been greatly successful in identifying tens of thousands susceptible genes for complex diseases or traits, revealing the genetic architectures of complex diseases or traits in question (Visscher et al., 2012, 2017). These large scale studies produce extremely valuable resource for further studies. However, due to the privacy concerns and other logistical considerations, most GWASs publish the summary association statistics rather than the individual-level data. This limitation motivates the rapid development of developing methods for analyzing the summary association statistics, such as conditional association analysis (Yang et al., 2012), gene-based association tests (Hu et al., 2013; Lee et al., 2013), jointly analyzing multiple traits (Zhu et al., 2015; Liu and Lin, 2018; Ray and Michael, 2018). A recent publication systematically reviews the development of summary association statistics-based methods (Pasaniuc and Price, 2017).

In this study, we mainly focus on the summary association statistics obtained from the GWASs of human diseases or traits while adjusting for heritable covariate, such as the GWAS of waist-to-hip ratio (WHR) after adjusting for BMI (Heid et al., 2010; Randall et al., 2013), the GWAS of fasting glycemic traits and insulin resistance after adjusting for BMI (Manning et al., 2012). However, it has been known that the results from these GWASs are biased, which may result in biased genetic estimates or even false positive genetic discoveries (Aschard et al., 2015). If the aim is to increase the statistical power, it is suggested to use the bivariate analyse of the trait (or disease) of interest and the corresponding heritable covariate (Aschard et al., 2015). However, the practical issue is still under addressed for this suggestion, that is how to extend the existing the biased GWAS to the bivariate analyse. Recent efforts have indicated that the multivariate GWAS can be conducted

based on summary association statistics of the univariate GWASs (Zhu et al., 2015; Liu and Lin, 2018; Ray and Michael, 2018). However, these methods require the summary association statistics from the unbiased GWASs, that is the univariate GWASs without adjusting the heritable covariate. In reality, many studies only have the results from the GWAS after adjusting the heritable covariate. For example, in the GIANT (Genetic Investigation of ANthropometric Traits) consortium website, we can only download the summary association statistics for WHR adjusted BMI stratified by sex and age (Winkler et al., 2015). To obtain the results for WHR without adjusting for BMI, it needs to re-run a GWAS, which needs a great effort. To our best knowledge, there are no literatures addressing how to extend the biased GWAS to the bivariate GWAS.

In this paper, we develop a simple integration method called BTOB which extends the Biased GWAS TO Bivariate GWAS. We assess the valid and efficiency of BTOB using theoretical arguments and simulation studies. Finally, we apply the BTOB method to analyze the data downloaded from the GIANT consortium website.

2. METHOD

2.1. BTOB: Extending the Biased GWAS to Bivariate GWAS

Mathematically, the model used in the biased GWAS can be formulated as $Y_2 = G\beta_2 + Y_1\gamma_1 + Z_2\varsigma_2 + \varepsilon_2$, where Y_2 is the trait or disease of interest, Y_1 is the adjusted heritable covariate, G is the genotype score, and Z_2 is the adjusted non-heritable covariates. In reality, many studies also had conducted additional GWAS for Y_1, that is $Y_1 = G\beta_1 + Z_1\varsigma_1 + \varepsilon_1$. For example, the GIANT consortium had conducted the GWASs for WHR while adjusting for BMI, and the GWASs of BMI (Winkler et al., 2015). In addition, it is common that partial sample overlap between these two GWASs. For example, the sample size of the GWAS for BMI in men cohort with age greater than 50 is about 90,000, while the corresponding GWAS for WHR after adjusting BMI only use a sub-sample with about 60,000 sample. And the two studies may use different covariates adjustment strategies. In conclusion, the above real scenarios can be formulated as follows

$$\begin{pmatrix} Y_1^c \\ Y_1^{u_1} \end{pmatrix} = \begin{pmatrix} G^c \\ G^{u_1} \end{pmatrix} \beta_1 + Z_1\varsigma_1 + \varepsilon_1, \quad (1)$$

$$\begin{pmatrix} Y_2^c \\ Y_2^{u_2} \end{pmatrix} = \begin{pmatrix} G^c \\ G^{u_2} \end{pmatrix} \beta_2^* + \begin{pmatrix} Y_1^c \\ Y_1^{u_2} \end{pmatrix} \gamma_1 + Z_2\varsigma_2 + \varepsilon_2. \quad (2)$$

Where Y_1^c and Y_2^c are the overlap sample of two phenotypes with genotypes G^c, $Y_1^{u_1}$ is the unique sample only used in first model with genotypes G^{u_1}, and $Y_2^{u_2}$ and $Y_1^{u_2}$ are the unique sample only used in second model with genotypes G^{u_2}. Z_1 and Z_2 includes the intercept and covariates, which may consider different covariates for different GWAS. In **Supplementary Theorem 1**, we show that the estimates of the genetic effects $\hat{\beta}_1$ and $\hat{\beta}_2^*$ are independent. Under the null hypothesis H_0: none of Y_1 and Y_1 associates with G, we have $\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right)^2 \sim \chi_1^2, \left(\frac{\hat{\beta}_2^*}{se(\hat{\beta}_2^*)}\right)^2 \sim \chi_1^2$. Therefore, we can simply integrate the summary association statistics in model (1) and (2), that is

$$\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right)^2 + \left(\frac{\hat{\beta}_2^*}{se(\hat{\beta}_2^*)}\right)^2 \sim \chi_2^2 \quad (3)$$

which is a test statistics about testing the null hypothesis H_0: none of Y_1 and Y_2 associates with G. Hence the proposed procedure extends the biased GWAS to bivariate analyse, which is termed BTOB (extends the Biased GWAS to Bivariate GWAS).

2.2. Simulations

We simulate 1,000 replicates of correlated traits, the causal SNP G is generated with minor allele frequency of 0.3 assuming the Hardy Weinberg equilibrium. The traits are generated using a linear additive model

$$Y_k = \beta_k G + \varepsilon_k, k = 1, \ldots, K$$

where $(\varepsilon_1, \ldots, \varepsilon_K)^\top$ follows multivariate normal distribution with mean 0 and covariance matrix Σ. We set the sample size of Y_1 to be 5,000, and then vary the sample size of Y_2 to be 5,000, 4,000, and 3,000. We consider three scenarios:(1) The tested variant affects the bivariate traits in the same direction. The tested variant explains 0.5% of the variance of Y_1 and 0 to 0.5% of the variance of Y_2, or the tested variant explains 0.5% of the variance of Y_2 and 0 to 0.5% of the variance of Y_1. The correlation was set to be low ($\rho = 0.4$), moderate ($\rho = 0.6$), or high ($\rho = 0.8$), where ρ was the correlation coefficient between Y_1 and Y_2. (2) The tested variant affects one phenotype only. Specifically, we considered the following two scenarios: the tested variant explains 0.5% of the variance of Y_1 and 0% of the variance of Y_2, or the tested variant explains 0.5% of the variance of Y_2 and 0% of the variance of Y_1. The correlation coefficient between Y_1 and Y_2 is varied from -0.9 to 0.9. (3) The test variant affects the bivariate traits in the opposite directions. The tested variant explains 0.3% of the variance of Y_1 and 0.4% of the variance of Y_2 with the opposite directions, or the tested variant explains 0.4% of the variance of Y_1 and 0.3% of the variance of Y_2 with the opposite direction. The correlation between Y_1 and Y_2 is varied from 0 to 0.9.

2.3. Study Decription

We download the gender and age specific summary association statistics for WHR after adjustment for BMI, and the marginal summary association statistics of BMI by the GIANT consortium from website http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files (Winkler et al., 2015). We integrated the summary association statistics from the following univariate GWASs stratified by age and gender: BMI∼SNP, WHR∼SNP+BMI, resulting in the bivariate analysis of WHR and BMI. The aim of this study is to assess whether the proposed BTOB approach can contribute novel gene compared with the corresponding univariate GWASs. Hence, the gene is considered to be novel if the lead SNP in (or 400 KB flanking) a gene is genome-wide significant in the bivariate analysis, whereas none of the lead SNPs in (or 400 KB flanking) this

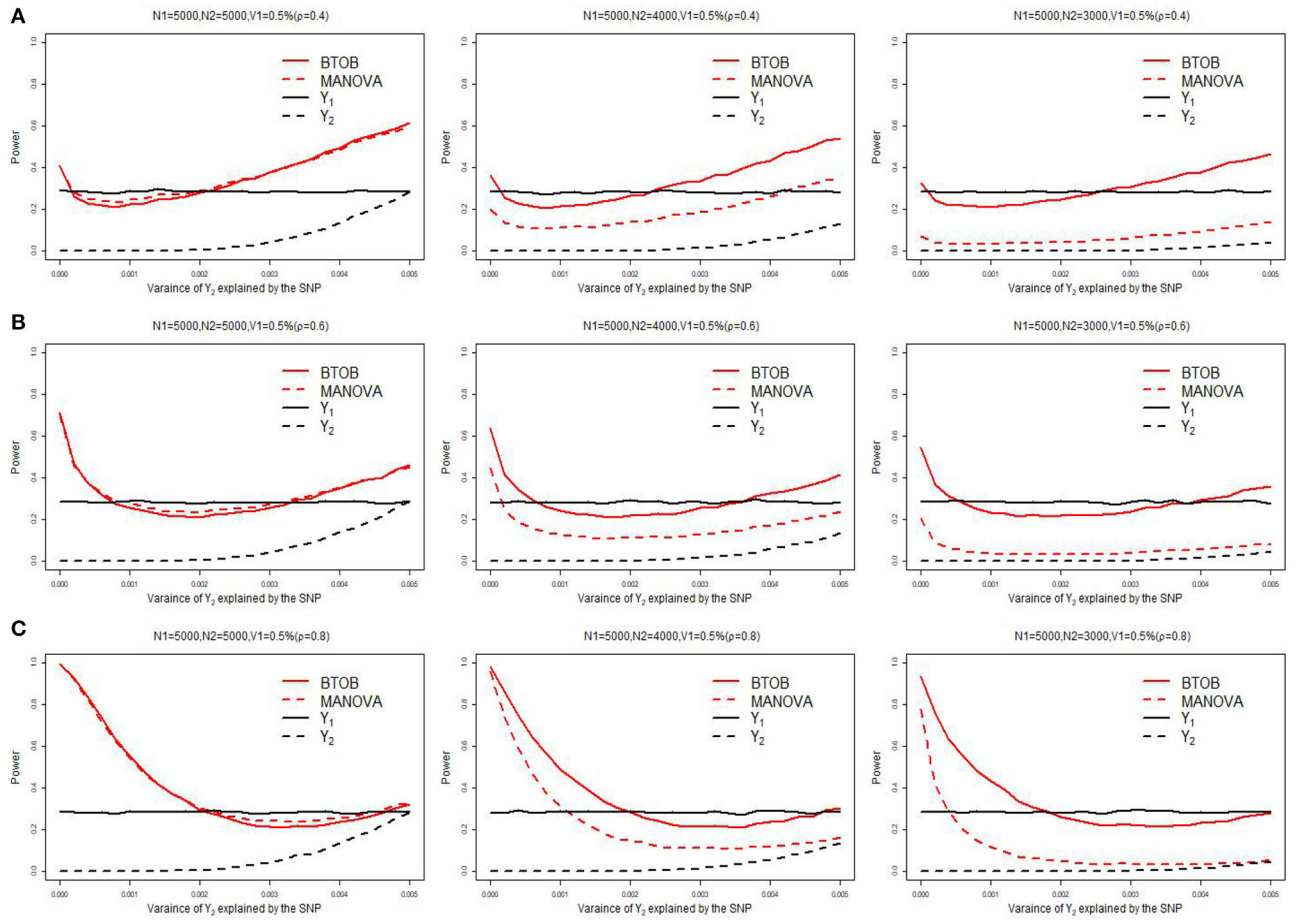


FIGURE 1 | Power comparison of BTOB, MANOVA, and the univariate analysis. The test variant explains 0.5% of the variance of Y_1, and the proportion of the test variant's variance for Y_2 varies from 0 to 0.5%. The genetic effects of Y_1 and Y_2 are with the same direction. The sample size of Y_1 is 5,000, and the sample size of Y_2 is set to be 5,000, 4,000 and 3,000, respectively. Three levels of correlation between Y_1 and Y_2 are investigated: low correlation with $\rho = 0.4$ **(A)**, moderate correlation with $\rho = 0.6$ **(B)**, and high correlation with $\rho = 0.8$ **(C)**.

gene reach genome-wide significance in the corresponding univariate GWASs. As we can only assess the HapMap II allele frequencies instead of pooled allele frequencies across all cohorts, we only included SNPs with sample size greater than 30,000, for which the HapMap allele frequencies may be representative.

3. RESULT

3.1. The Performance of BTOB in Integrating the Summary Association Statistics

For illustrate purpose, we conducted simulation studies to investigate the validity and efficiency of the proposed BTOB. As a comparison, we include the MANOVA method (Ray et al., 2016). Since MANOVA is not directly applicable to the summary association data, we use the overlap sample and re-run the multivariate association analysis using the MANOVA.

Supplementary Table 1 presents the type 1 error for BTOB, which shows that the proposed BTOB can control the type 1 error rate quite well. **Figure 1** presents the power comparisons when the tested variant affects the bivariate phenotypes in the same direction. The tested variant explains 0.5% of the variance of Y_1 and 0 to 0.5% of the variance of Y_2. We can observe from **Figure 1** that BTOB and MANOVA have nearly the same power when both phenotypes have the sample size 5,000, which indicates the validity and efficiency for BTOB. However, as the overlap sample size is set to be 4,000, BTOB performs much better than MANOVA. When the overlap sample size is set to be 3,000, the power discrepancy between BTOB and MANOVA is more obvious. This is expected as the sample size used in GWAS: $Y_1 = \mu_1 + \beta_1 G + Z_1\varsigma_1 + \varepsilon_1$ is often larger than the sample size used in GWAS: $Y_2 = \beta_2 G + Y_1\gamma_1 + Z_2\varsigma_2 + \varepsilon_2$. Traditional multivariate approaches, such as MANOVA, are only applicable to the overlap sample between Y_1 and Y_2. However, the proposed BTOB can make full use of the whole sample for Y_1, hence boosting the power compared with MANOVA. The same

TABLE 1 | The novel Genome-wide Significant loci which were identified by the proposed combining method but not found by the standard univariate approach for the analysis of WHR and BMI.

				BMI			WHR∼BMI				BTOB	
Cohort	**SNP**	**Chr**	**Gene**	**Beta**	**SE**	***P – value*[a]**	***N_1***	**Beta**	**SE**	***P – value*[a]**	***N_2***	***P – value*[b]**
Men(Age>50)	rs10923746	1	WARS2	–0.020	0.0051	5.3e-05	90,515	0.029	0.0063	4.4e-06	56,398	5.405e-09
Men(Age>50)	rs12073056	1	TBX15	–0.022	0.0049	6.7e-06	90,142	0.030	0.0062	9.9e-07	55,682	1.774e-10
Men(Age>50)	rs3817973	6	HCG23	–0.018	0.005	2.7e-04	91,470	0.031	0.0062	4.7e-07	56,924	3.019e-09
Men(Age>50)	rs9378213	6	HLA-DRA	–0.022	0.0051	1.6e-05	89,222	0.03	0.0063	3.2e-06	56,647	1.264e-09
Women(Age>50)	rs12998590	2	SLC38A11	–0.022	0.0054	6.3e-05	88,374	0.031	0.0067	3.2e-06	57,158	4.702e-09
Women(Age>50)	rs253393	5	POC5	–0.026	0.0058	8.40E-06	88,423	-0.026	0.0072	0.00024	57,159	4.24E-08
Women(Age>50)	rs6971365	7	KLF14	–0.017	0.0052	0.0013	104,946	0.033	0.0062	1.00E-07	71,909	3.09E-09
Women(Age>50)	rs11191295	10	TMEM180	0.017	0.0049	4.1e-04	97,313	–0.027	0.0058	3.3e-06	66,010	2.898e-08

[a]The *results for univariate phenotypes approach. The genome-wide Significant level is set to be 2.5E-08 with the Bonferroni correction.* [b]The *results for the BTOB approach. The genome-wide Significant level is set to be 5E-08. Chr, chromosome; N1, the sample size of GWAS for BMI; N2, the sample size of GWAS for WHR adjusting for BMI.*

phenomenons can be observed in **Figures 1B,C** with median and high correlation. In **Figure 1**, we also compare the power between the bivariate analysis and the univariate analysis after the Bonferroni correction. We can observe from **Figure 1** that BTOB approach performs better than the univariate approach in most scenarios. It should be noted that there is a decrease of power for BTOB when the proportion of the test variant's variance for Y_2 varies from 0 to a reasonably small value. This counterintuitive phenomenon can be explained by using the theoretical results given in a recent work (Guo et al., 2018). **Supplementary Figures 1–3** present the power comparison for two other scenarios: the tested variant affects one trait only, and the tested variant affects the bivariate traits in the opposite direction. All of the simulated results indicate the superior performance for BTOB compared with MANOVA when the overlap sample size is set to be 4,000 and 3,000, and the superior power for BTOB compared with univariate analysis in most scenarios.

3.2. Real Data Analysis

In total, 8 loci are novel compared with the univariate GWASs: 4 for bivariate analysis of WHR and BMI in the cohort of men aged over 50, and 4 for bivariate analysis of WHR and BMI in the cohort of women aged over 50 (**Table 1**). The genomic control (GC) inflation factors of these 4 bivariate analyses is presented in **Supplementary Table 2**.

Firstly, for the analyses of WHR and BMI in the cohort of women aged over 50, we identified 4 novel genes compared with the univariate GWASs (WARS2, leading SNP: rs10923746, *p*-value = 5.405E-09; TBX15, leading SNP: rs10923715, *p*-value = 4.88E-11; HCG23, leading SNP: rs3817973, *p*-value = 3.019e-09; HLA-DRA, leading SNP: rs9378213, *p*-value = 1.264e-09) (**Table 1**). Even though these 4 leading SNPs show evidence of association in the univariate analyses: GWAS for WHR after adjusting BMI and GWAS for BMI, these univariate analyses have no enough power to reach the genome-wide significance. What is more, for the analyse of WHR and BMI in the cohort of women aged over 50, BTOB method identified 4 novel loci compared with the univariate GWASs (SLC38A11, leading SNP: rs12998590, *p*-value = 4.702e-09; POC5, leading SNP: rs253393, *p*-value = 4.24E-08; KLF14, leading SNP: rs6971365, *p*-value = 3.09E-09; TMEM180, leading SNP: rs11191295, *p*-value = 2.898e-08) (**Table 1**). The real data analysis suggested that the BTOB method is capable to integrate moderate signals from the corresponding univariate analyses, hence leading to the identification of novel genetic signals compared with the univariate analyses. Further, six identified loci from the BTOB method, including TBX15, WARS2, POC5, KLF14, HLA-DRA, SLC38A11, were confirmed in the follow-up GWASs with at least ten times larger sample size (Pulit et al., 2019; Zhu et al., 2020), suggesting BTOB can help identify novel genes in the GWASs when the sample size is limited.

Finally, several studies have suggested a potential causal role of these identified genes in adipose development and function. For example, animal models have demonstrated that the important role of WARS2 in regulating brown adipose tissue function and consequently lipid and glucose metabolism, by regulating mitochondrial respiration, leading to the increased glucose oxidation in brown adipose tissues (Pravenec et al., 2017; Ejarque et al., 2019). TBX15 encodes a T-box transcription factor (TF) that has shown to be involved in various aspects of adipose development and maintenance, also to be associated with body fat distribution (Singh et al., 2005; Zhang et al., 2020). It has also been implicated the transcription factor KLF14, a member of the Krupple-like factor family (KLF), plays a key role in energy homeostasis by regulating lipid and glucose metabolism, and adipogenesis via promoting adipocyte differentiation (Chen et al., 2005; Birsoy et al., 2008).

4. DISCUSSION

There are several concerns that should be noted about multivariate approaches in GWAS. First, the proposed bivariate method or other multivaraite methods for summary association statistics from univariate GWASs have been shown to help identify novel genes compared with univariate GWASs. While the multivariate approaches can also fails some genes identified in the univariate GWASs. Hence, the multivaraite GWASs

should be considered as a valuable compensation rather substitution for univariate GWASs. Second, there is no single multivariate method that is uniformly most powerful in all scenarios. Hence, it is valuable to try several candidate methods in real case.

In summary, our proposed approach provides an efficient shortcut for extending the existing biased GWASs to the bivariate GWAS. Considering a great amount of large scale biased GWASs have been published (Hancock et al., 2010; Kaplan et al., 2011; Randall et al., 2013; Loth et al., 2014; Winkler et al., 2015; Pulit et al., 2019; Zhu et al., 2020), the proposed BTOB method is expected to be of great practical utility.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

XG conceived the idea and conducted the simulation. JZ, WD, YW, and QF processed the data and conducted the real dataset experiments. XG and QF wrote the manuscript. XG, QF, JZ, and WD revised the manuscript. All authors read and approved the final manuscript.

FUNDING

XG's research was supported by the NSFC (11771463), the Pearl River S&T Nova Program (201806010142).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.654821/full#supplementary-material

REFERENCES

Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* 96, 329–339. doi: 10.1016/j.ajhg.2014.12.021

Birsoy, K., Chen, Z., and Friedman, J. (2008). Transcriptional regulation of adipogenesis by klf4. *Cell Metab.* 7, 39–347. doi: 10.1016/j.cmet.2008.02.001

Chen, Z., Torrens, J. I., Anand, A., Spiegelman, B. M., and Friedman, J. M. (2005). Krox20 stimulates adipogenesis via c/ebpβ-dependent and-independent mechanisms. *Cell Metab.* 1, 93–106. doi: 10.1016/j.cmet.2004.12.009

Ejarque, M., Ceperuelo-Mallafré, V., Serena, C., Maymo-Masip, E., Duran, X., Díaz-Ramos, A., et al. (2019). Adipose tissue mitochondrial dysfunction in human obesity is linked to a specific DNA methylation signature in adipose-derived stem cells. *Int. J. Obes.* 43, 1256–1268. doi: 10.1038/s41366-018-0219-6

Guo, X., Zhu, J., Fan, Q., He, M., Wang, X., and Zhang, H. (2018). A univariate perspective of multivariate genome-wide association analysis. *Genet. Epidemiol.* 42, 470–479. doi: 10.1002/gepi.22128

Hancock, D. B., Eijgelsheim, M., Wilk, J. B., Gharib, S. A., Loehr, L. R., Marciante, K. D., et al. (2010). Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* 42:45. doi: 10.1038/ng.500

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., et al. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.* 42, 949–960. doi: 10.1038/ng.685

Hu, Y.-J., Berndt, S. I., Gustafsson, S., Ganna, A., Mägi, R., Wheeler, E., et al. (2013). Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *Am. J. Hum. Genet.* 93, 236–248. doi: 10.1016/j.ajhg.2013.06.011

Kaplan, R. C., Petersen, A.-K., Chen, M.-H., Teumer, A., Glazer, N. L., Döring, A., et al. (2011). A genome-wide association study identifies novel loci associated with circulating igf-i and igfbp-3. *Hum. Mol. Genet.* 20, 1241–1251. doi: 10.1093/hmg/ddq560

Lee, S., Teslovich, T. M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* 93, 42–53. doi: 10.1016/j.ajhg.2013.05.010

Liu, Z., and Lin, X. (2018). Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics* 74, 165–175. doi: 10.1111/biom.12735

Loth, D. W., Artigas, M. S., Gharib, S. A., Wain, L. V., Franceschini, N., Koch, B., et al. (2014). Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat. Genet.* 46, 669–677. doi: 10.1038/ng.3011

Manning, A. K., Hivert, M.-F., Scott, R. A., Grimsby, J. L., Bouatia-Naji, N., Chen, H., et al. (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* 44, 659–669. doi: 10.1038/ng.2274

Pasaniuc, B., and Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* 18, 117–127. doi: 10.1038/nrg.2016.142

Pravenec, M., Zidek, V., Landa, V., Mlejnek, P., Šilhavỳ, J., Šimáková, M., et al. (2017). Mutant wars2 gene in spontaneously hypertensive rats impairs brown adipose tissue function and predisposes to visceral obesity. *Physiol. Res.* 66, 917–924. doi: 10.33549/physiolres.933811

Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., et al. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of european ancestry. *Hum. Mol. Genet.* 28, 166–174. doi: 10.1093/hmg/ddy327

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., et al. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet.* 9:e1003500. doi: 10.1371/journal.pgen.1003500

Ray, D., and Michael, B. (2018). Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet. Epidemiol.* 42, 134–145. doi: 10.1002/gepi.22105

Ray, D., Pankow, J. S., and Basu, S. (2016). Usat: a unified score-based association test for multiple phenotype-genotype analysis. *Genet. Epidemiol.* 40, 20–34. doi: 10.1002/gepi.21937

Singh, M. K., Petry, M., Haenig, B., Lescher, B., Leitges, M., and Kispert, A. (2005). The t-box transcription factor tbx15 is required for skeletal development. *Mech. Dev.* 122, 131–144. doi: 10.1016/j.mod.2004.10.011

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *Am. J. Hum. Genet.* 90, 7–24. doi: 10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Winkler, T. W., Justice, A. E., Graff, M., Barata, L., Feitosa, M. F., Chu, S., et al. (2015). The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS Genet.* 11:e1005378. doi: 10.1371/journal.pgen.1005378

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* 44:369. doi: 10.1038/ng.2213

Zhang, X., Ehrlich, K. C., Yu, F., Hu, X., Meng, X.-H., Deng, H.-W., et al. (2020). Osteoporosis-and obesity-risk interrelationships: an

epigenetic analysis of GWAS-derived SNPs at the developmental gene tbx15. *Epigenetics* 15, 728–749. doi: 10.1080/15592294.2020.1716491

Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* 96, 21–36. doi: 10.1016/j.ajhg.2014.11.011

Zhu, Z., Guo, Y., Shi, H., Liu, C.-L., Panganiban, R. A., Chung, W., et al. (2020). Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK biobank. *J. Allergy Clin. Immunol.* 145, 537–549. doi: 10.1016/j.jaci.2019.09.035

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 May 2021
doi: 10.3389/fgene.2021.657499

Identifying Susceptibility Loci for Cutaneous Squamous Cell Carcinoma Using a Fast Sequence Kernel Association Test

*Manyan Huang[1], Chen Lyu[1], Xin Li[2,3], Abrar A. Qureshi[4], Jiali Han[2,3] and Ming Li[1]**

[1] Department of Epidemiology and Biostatistics, School of Public Health, Indiana University at Bloomington, Bloomington, IN, United States, [2] Department of Epidemiology, Richard M. Fairbanks School of Public Health, Indiana University – Purdue University Indianapolis, Indianapolis, IN, United States, [3] Melvin and Bren Simon Cancer Center, Indianapolis, IN, United States, [4] Department of Dermatology, Alpert Medical School, Brown University, Providence, RI, United States

Cutaneous squamous cell carcinoma (cSCC) accounts for about 20% of all skin cancers, the most common type of malignancy in the United States. Genome-wide association studies (GWAS) have successfully identified multiple genetic variants associated with the risk of cSCC. Most of these studies were single-locus-based, testing genetic variants one-at-a-time. In this article, we performed gene-based association tests to evaluate the joint effect of multiple variants, especially rare variants, on the risk of cSCC by using a fast sequence kernel association test (fastSKAT). The study included 1,710 cSCC cases and 24,304 cancer-free controls from the Nurses' Health Study, the Nurses' Health Study II and the Health Professionals Follow-up Study. We used UCSC Genome Browser to define gene units as candidate loci, and further evaluated the association between all variants within each gene unit and disease outcome. Four genes *HP1BP3, DAG1, SEPT7P2,* and *SLFN12* were identified using Bonferroni adjusted significance level. Our study is complementary to the existing GWASs, and our findings may provide additional insights into the etiology of cSCC. Further studies are needed to validate these findings.

Keywords: region-based association test, fast sequence kernel association test, cutaneous squamous cell carcinoma, rare variants, generalized genetic random field

OPEN ACCESS

Edited by:
Qi Yan,
Columbia University, United States

Reviewed by:
Rong Zhang,
Amgen, United States
Yalu Wen,
The University of Auckland, New Zealand

****Correspondence:***
Ming Li
li498@indiana.edu

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *23 January 2021*
Accepted: *09 April 2021*
Published: *10 May 2021*

Citation:
Huang M, Lyu C, Li X, Qureshi AA, Han J and Li M (2021) Identifying Susceptibility Loci for Cutaneous Squamous Cell Carcinoma Using a Fast Sequence Kernel Association Test. Front. Genet. 12:657499. doi: 10.3389/fgene.2021.657499

INTRODUCTION

Cutaneous squamous cell carcinoma (cSCC) is the second most common type of non-melanoma skin cancers, accounting for about 20% of all skin cancers and the majority of deaths attributable to non-melanoma skin cancers (Chitsazzadeh et al., 2016; Motaparthi et al., 2017; Parekh and Seykora, 2017; Que et al., 2018a). The incidence of cSCC in the United States has been increasing over the last few decades, with over 1 million annual cases in recent years (Nguyen et al., 2014; Muzic et al., 2017; Que et al., 2018a,b). The increase is also expected to continue because of the longer life expectancy, aging population and chronic ultraviolet exposure (Nguyen et al., 2014; Motaparthi et al., 2017; Waldman and Schmults, 2019). The growing mortality and morbidity of cSCC has posed immense economic burden on the national healthcare systems. Though the remission rate of cSCC cases has substantially improved, many cases were still associated with higher probability of recurrence,

metastasis and poor prognosis after surgery (Motaparthi et al., 2017; Que et al., 2018a; Waldman and Schmults, 2019). It is of crucial importance to understand the pathogenesis of cSCC and to reduce the public health impact of the disease.

The etiology of cSCC has not been fully understood. However, the risk of the disease can be influenced by multiple environmental exposures. For example, higher risk of cSCC is found to be associated with increased age, fair skin color, male gender, exposure to ultraviolet radiation, immunosuppression and human papillomavirus (Chahal et al., 2016; Parekh and Seykora, 2017; Que et al., 2018a; Waldman and Schmults, 2019). Similar to all cancers, genetic susceptibility also plays an important role in the development of cSCC. Familial aggregation provides direct evidence for the heritability of cSCC (Hussain et al., 2009; Asgari et al., 2015). A few known cancer-related genes, such as *TP53, CDKN2A*, *Ras,* and *NOTCH1* were also causal to skin cancers (Que et al., 2018a). Mutations with these genes may disrupt normal cell growth, cell circle and cellular signal transduction, leading to the development of the disease. In the past decade, genome-wide association studies (GWAS) have become a commonly used strategy to identify genetic variants for complex human diseases in the general population. A few GWASs have identified multiple genetic variants that are associated with the risk of cSCC, such as *CADM1, AHR, SEC16A*, and *DEF8* (Nan et al., 2011; Asgari et al., 2016; Chahal et al., 2016; Siiskonen et al., 2016). Many findings were also successfully replicated in independent populations. These findings have provided valuable insights into the genetic etiology of cSCC.

Despite of these successes, it was estimated that the genetic variants identified by existing GWASs only account for ∼8.5% of the cSCC heritability (Sarin et al., 2020). The genetic causes of the disease remain largely unknown (Chahal et al., 2016). This challenge may be due to a number of limitations of the existing GWASs, such as insufficient statistical power to detect small to moderate genetic effects, burden of multiple testing adjustment, and overlooking potential interactions among variants (Mo et al., 2015; Nettiksimmons et al., 2016). As an alternative to the single-locus analysis, gene- or region-based analysis can be a complementary approach addressing some of those limitations. It may integrate effects of multiple genetic variants, especially rare variants, within a genetic region for improved power, reduce the computational intensities and alleviate the burden of multiple testing (Wu et al., 2010). In recent years, a number of statistical methods have been developed for conducting region-based association test. For example, a sequence kernel association test (SKAT) has been a commonly used method that evaluates the joint effects of genetic variants in a region on a disease outcome while adjusting for covariates (Wu et al., 2011). It uses flexible kernel functions to integrate the effects from multiple variants and allows the effect of causal variants to be bi-directional. Further, a fast sequencing kernel association test (fastSKAT) has been developed to implement SKAT in a computational efficient fashion, especially for large-scale studies with thousands of subjects (Lumley et al., 2018). In this article, we assessed the validity of region-based fastSKAT by replicating 18 GWAS-identified SNPs using single-locus testing. We further tested the association between approximately 23,000 gene regions and cSCC outcome in five independent study populations. The results from each population were further integrated by a Fisher's combined probability test.

MATERIALS AND METHODS

Ethics Statement

The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required.

Study Population

Our study included 26,014 individuals from three large prospective cohort studies in the U.S., including the Nurses' Health Study (NHS), the Nurses' Health Study 2 (NHS2), and the Health Professionals Follow-up Study (HPFS). The subjects were selected under a nested case-control design based on cSCC status. Cases were defined as individuals diagnosed with invasive cSCC, while controls were individuals free of cSCC or any primary type of cancers. The individuals' characteristics, genotypes and other covariates information were collected in the NHS, the NHS2 and the HPFS studies. In this study, we partitioned the subjects into five independent sub-populations based on their genotyping platforms, including "Affymetrix," "Illumina," "OmniExpress," "OncoArray" and "HumanCore." In the following, we used these platforms to represent five populations. After the quality control process, the five populations included a total of 5,533, 3,314, 5,354, 5,267, and 6,646 subjects, respectively. More details about the study design and data collection were described elsewhere (Chahal et al., 2016; Duffy et al., 2018).

Genomic Imputation and Quality Control

The genomic datasets, imputation and quality control procedures were conducted separately in each population and were described with details in previous publications (Lindström et al., 2017; Duffy et al., 2018). Briefly, the participants from five sub-populations were genotyped at different times and by different genotyping platforms. The subjects in "Affymetrix" were genotyped by the Genome-wide Human SNP Array 6.0. The subjects in "Illumina" were genotyped by either Illumina HumanHap300 BeadChip, HumanHap550-Quad BeadChip, Human610-Quad BeadChip, or Human660W-Quad BeadChip. The subjects in "OmniExpress" were genotyped by Illumina HumanOmniExpress-12 BeadChip. The subjects in "OncoArray" were genotyped by Infinium OncoArray-550K BeadChip. The subjects in "HumanCore" were genotyped by Illumina HumanCoreExome-12v1-0 BeadChip.

Variants with low call rate (<95%) were removed. A pairwise identity-by-descent (IBD) analysis was conducted to identify duplicates. For individuals who may be genotyped for more than once using different genotyping platforms, one of the duplicated pair was excluded by the order of "Affymetrix," "Illumina," "OmniExpress," "OncoArray," and "HumanCore." For individuals with different cohort IDs but a high genetic concordance rate, both of the pairs were removed. Genome

imputation was further conducted in each population using the 1000 Genomes Project Phase 3 Integrated Release Version 5 as reference panels. Software *ShapeIT* (v2.r837) was used for genotype phasing, and the phased genotypes were further imputed to ∼ 47 million variants using *Minimac3* (O'Connell et al., 2014; Das et al., 2016).

Replication of GWAS Identified SNPs Using Single-Locus Testing

To evaluate the validity of fastSKAT, we used 18 SNPs identified in two previous GWAS as positive controls (Chahal et al., 2016; Sarin et al., 2020). In these previous GWASs, ten SNPs were identified involving 3 independent populations (i.e., "Affymetrix," "Illumina," and "OmniExpress"), and 8 SNPs were identified using all 5 populations. For comparison purpose, we first used fastSKAT to test the association between each of these SNPs and cSCC, and further conducted a Fisher's combined probability test to evaluate the overall association across three or five populations consist with their analysis in the previous GWASs. For fair comparison, we calculated *p*-values by applying fastSKAT to the same NHS and HPFS populations used in previous publications. In particular, "Affymetrix," "Illumina," and "OmniExpress" were used in Chahal et al. (2016), while "Affymetrix," "Illumina," "OmniExpress," "OncoArray," and "HumanCore" were all used in Sarin et al. (2020). The *p*-values were compared to those of previous GWAS publications for consistency.

Genomic Region Selection

To identify biologically meaningful loci, we used UCSC Genome Browser (assembly GRCh37/hg19) to define gene units as candidate loci for region-based analysis. Software *bedtools* were used to merge the redundant and overlapping genomic regions based on the gene annotation (Kindlon ARQaN, 2009–2019; Quinlan and Hall, 2010). A candidate locus was then defined as 7.5KB upstream and downstream the corresponding gene region. Ultimately, a total of 25,437 regions were extracted. During the data processing, SNPs with an imputation quality score less than 0.3 were removed. We also extracted common and rare variants separately for each region using *PLINK2.0* (Purcell et al., 2007; Purcell). Common and rare variants were defined based on whether the minor allele frequency (MAF) was larger than 5%. Because previous GWAS has comprehensively evaluated each single variant for association with cSCC, we only considered regions with two or more variants for region-based association analysis.

Region-Based Association Test

We evaluated the association between genomic regions and cSCC using the fastSKAT, a region-based association test that is computationally efficient for large-scale genomic datasets (Lumley et al., 2018). Similar to the SKAT method, it is a variance component score test that integrates the effect of multiple genetic variants within the same region (Wu et al., 2011). The improvement of computational speed over SKAT was achieved by accurately approximating the tail probability for the asymptotic distribution of the test statistics (Lumley et al., 2018). Instead of computing all the eigenvalues of the genotypic similarity matrix, only the top ones were computed through random projections (Halko et al., 2011; Tropp, 2011). The tail probability can then be approximated by the top eigenvalues and a reminder term computed using Satterthwaite approximation, which approximates the sum of weighted chi-square distributions with a single chi-square distribution. The fastSKAT has been implemented in R package "bigQF" (Lumley et al., 2018). For each gene region, the method was applied for rare variants (MAF $<$ 5%) and common (MAF $\geq$ 5%) variants separately, and also for all variants together, adjusting for age, gender and the first five genetic principal components. A weighted linear kernel was used with each variant weighted by $Beta(MAF, 1, 25)$, the beta distribution density function. After testing each region within each of the five sub-populations, we further adopted the Fisher's combined probability test to integrate the *p*-values from sub-populations for an overall *p*-value.

Cross-Check With Expression Quantitative Trait Loci (eQTL) Database

The majority of variants identified by existing GWASs were located in the non-coding regions of the genome, and were therefore likely to be involved in gene regulation. One hypothesis is that that causal genetic variants for complex diseases may function through regulating the expression level of genes within specific tissues. To prioritize our findings, we further examined if the identified genes harbor any known expression quantitative trait locus (eQTL) in the database. We used the Genotype-Tissue Expression (GTEx) database (GTEx Consortium, 2013) for cross checking. There are two main types of skin tissues available in the GTEx, including sun-exposed skin at lower leg and sun-unexposed skin in suprapubic region. We summarized the number of eQTLs located within each identified region for either of skin tissue types.

RESULTS

Study Population

Our study included a total of 1,710 cSCC cases and 24,304 controls, partitioned into five sub-populations based on genotyping platforms. The number of subjects and their characteristics by each population is summarized in **Table 1**. The case-control ratios ranged from 1:6 to 1:31 across five populations. Gender was statistically different between cases and controls in four populations ($p < 0.05$), which was consist with the fact that the incidence rate was higher in men than in women (Karagas et al., 1999; Nguyen et al., 2014). Age, a well-established risk factor, was associated with cSCC in all populations ($p < 0.001$).

Replication of GWAS Identified SNPs Using Single-Locus Testing

For a total of 18 SNPs identified by previous GWASs, we used fastSKAT to test each variant for association with the disease outcome and compared the testing *p*-values with those reported in previous publications. The comparison is presented in **Figure 1** and summarized in **Table 2**. We found that the Fisher's *p*-values

combining fastSKAT results of multiple populations were highly correlated with the reported *p*-values in previous publications. The Fisher's combined *p*-values tend to be smaller, especially for variants with relatively small testing *p*-values (e.g., <0.01), leading to a higher level of statistical significance for the association. The results suggested that testing with fastSKAT in each population and combining with Fisher's combined probability test was able to reliably identify the gene-disease association with improved statistical power.

TABLE 1 | Study population characteristics and number of regions tested in each population.

Population	n (%)	Male		Age	
		n (%)	*p*-value[a]	Mean (SD)	*p*-value[a]
Affy (*n* = 5,533)					
Cases	340 (6.1)	166 (48.8)	0.004	50.34 (9.53)	<0.001
Controls	5193 (93.9)	2118 (40.8)		48.10 (9.48)	
Illumina (*n* = 3,314)					
Cases	200 (6.0)	63 (31.5)	0.002	48.25 (8.70)	<0.001
Controls	3114 (94.0)	683 (21.9)		43.72 (8.71)	
Omni (*n* = 5,354)					
Cases	737 (14.0)	281 (38.1)	0.310	48.51 (9.52)	<0.001
Controls	4517 (86.0)	1631 (36.1)		46.90 (8.90)	
Onco (*n* = 5,267)					
Cases	226 (4.3)	94 (41.6)	<0.001	47.80 (9.77)	<0.001
Controls	5041 (95.7)	866 (17.2)		41.01 (8.87)	
HumanCore (*n* = 6,646)					
Cases	207 (3.1)	102 (49.3)	<0.001	48.40 (10.24)	<0.001
Controls	6439 (96.9)	1262 (19.6)		40.96 (9.54)	

[a] *p-value by two-sample t-test for age and by Chi-square test for gender.*

Region-Based Association Test

Approximately 23,000 candidate regions were extracted and tested in each population. The numbers differed slightly across populations and was listed in **Table 3**. For each candidate region, the rare variants, common variants and all variants were tested separately for association with cSCC outcome using fastSKAT. The distribution of testing *p*-values were examined against a uniform distribution via quantile-quantile plots (**Supplementary Figures 1–3** for rare, common and all variants, respectively). The genomic inflation factors ranged between 0.974 and 1.07, suggesting well-controlled type I error rates. The Manhattan plots based on fastSKAT and Fisher's method are provided in **Figures 2–4**.

A total of four genomic regions were identified by Fisher's combined probability test at the Bonferroni adjusted significance level. The genomic regions and their testing *p*-values are listed in **Table 4**. Four regions were identified via rare variants association, and one of them was also identified via all variants analysis. No regions reached statistical significance after Bonferroni adjustment via common variants analysis. While the overall significant association was largely driven by one particular population for most of these regions, the association for one region was replicated by one additional population in the study. In particular, a region (gene

TABLE 2 | Comparison of *p*-values for 18 SNPs identified by published GWASs and computed by fastSKAT.

Publication	SNP	Chro	Gene[c]	*p*-value in paper[d]	*p*-value by fastSKAT[e]
Sarin et al., 2020[a]	rs10399947	1	*ARNT–[]–SETDB1*	2.31×10^{-2}	9.41×10^{-1}
	rs10200279	2	*ALS2CR12*	3.34×10^{-1}	2.59×10^{-1}
	rs10944479	6	*BACH2*	5.99×10^{-2}	3.73×10^{-1}
	rs7834300	8	*TRPS1*	1.58×10^{-1}	6.89×10^{-1}
	rs1325118	9	*[]–TYRP1*	8.60×10^{-2}	2.08×10^{-1}
	rs7939541	11	*ZNF143–[]–WEE1*	8.55×10^{-2}	1.80×10^{-1}
	rs657187	12	*KRT6A–[]–KRT5*	3.25×10^{-1}	4.20×10^{-1}
	rs721199	12	*HAL*	1.08×10^{-3}	3.07×10^{-1}
Chahal et al., 2016[b]	rs12203592	6	*IRF4*	3.10×10^{-6}	1.33×10^{-10}
	rs1805007	16	*MC1R*	4.90×10^{-5}	1.88×10^{-7}
	rs35407	5	*SLC45A2*	5.50×10^{-2}	8.56×10^{-2}
	rs1126809	11	*TYR*	3.30×10^{-1}	1.15×10^{-2}
	rs6059655	20	*RALY-ASIP*	5.40×10^{-1}	5.51×10^{-2}
	rs1800407	15	*OCA2*	8.30×10^{-1}	4.76×10^{-1}
	rs57994353	9	*SEC16A*	4.70×10^{-1}	5.65×10^{-1}
	rs10810657	9	*BNC2, CNTLN*	1.20×10^{-2}	1.70×10^{-3}
	rs74899442	11	*CADM1, BUD13*	1.80×10^{-1}	1.85×10^{-1}
	rs117132860	7	*AHR*	4.00×10^{-2}	1.94×10^{-1}

[a] *Sarin et al. (2020). Genome-wide meta-analysis identifies eight new susceptibility loci for cutaneous squamous cell carcinoma. Nat Commun 11, 820.*
[b] *Chahal et al. (2016). Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. Nat Commun 7, 12048.*
[c] *The format gene–[]– indicates SNPs are located within intergenic regions.*
[d] *p-values reported in previous publications using either three or five NHS/HPFS populations.*
[e] *p-values of Fisher's method combining fastSKAT p-values from NHS/HPFS populations used in previous publications.*

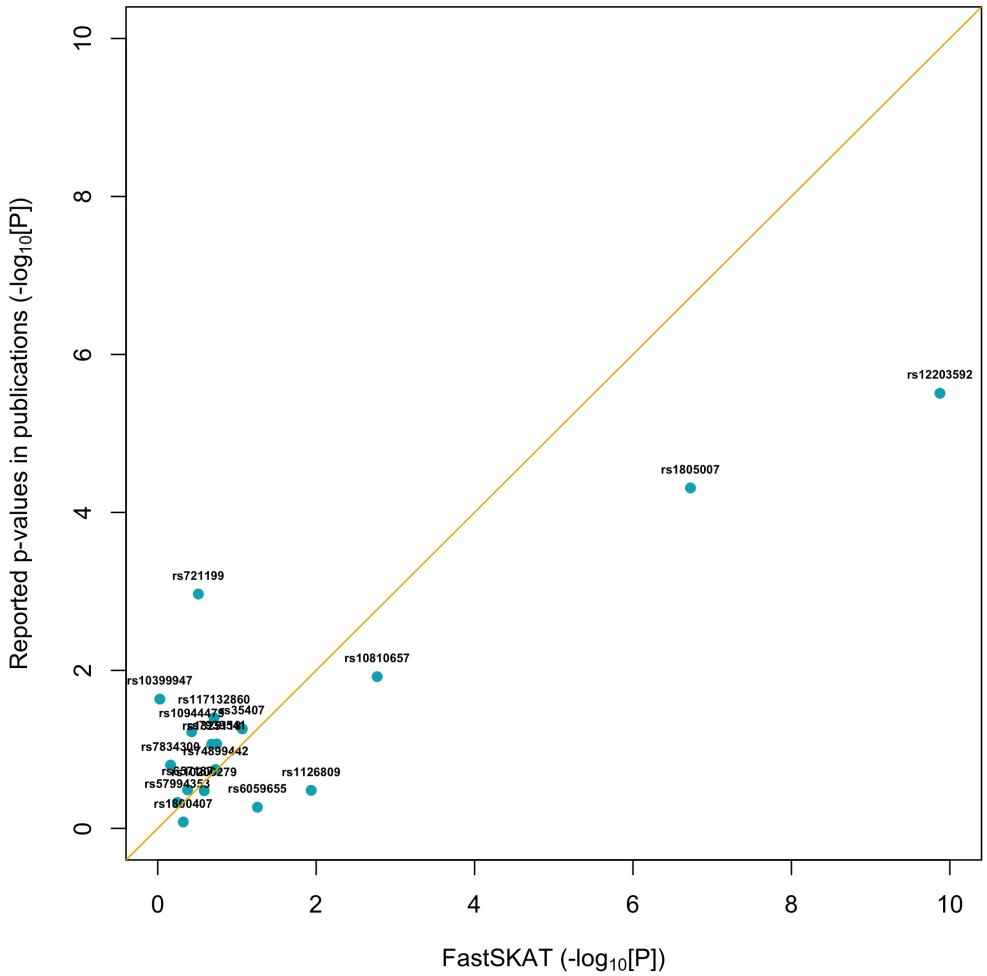


FIGURE 1 | Replication of 18 GWAS identified SNPs using fastSKAT. The *p*-values of fastSKAT were based on Fisher's method combining its testing *p*-values from the same NHS and HPFS populations used in previous publications.

TABLE 3 | Total number of regions and genetic variants tested in each population.

Population	Rare variants				Common variants				All variants			
	# of regions	# of SNPs in regions		Significance level[a]	# of regions	# of SNPs in regions		Significance level[a]	# of regions	# of SNPs in regions		Significance level[a]
		Range	Median			Range	Median			Range	Median	
Affy	23,566	2–26,354	131	2.12×10^{-6}	23,552	2–13,667	79	2.12×10^{-6}	23,675	2–40,021	210	2.11×10^{-6}
Illumina	23,565	2–26,485	131	2.12×10^{-6}	23,518	2–13,673	80	2.13×10^{-6}	23,661	2–40,158	211	2.11×10^{-6}
Omni	23,645	2–27,077	157	2.11×10^{-6}	23,619	2–13,700	80	2.12×10^{-6}	23,729	2–40,777	230	2.11×10^{-6}
Onco	23,546	2–24,220	120	2.12×10^{-6}	23,540	2–13,655	79	2.12×10^{-6}	23,673	2–37,875	198	2.11×10^{-6}
HumanCore	23,734	2–18,549	109	2.11×10^{-6}	23,699	2–13,648	79	2.11×10^{-6}	23,823	2–32,197	214	2.10×10^{-6}
Fisher	23,844	–	–	2.10×10^{-6}	23,803			2.10×10^{-6}	23,897	–	–	2.09×10^{-6}

[a] *Bonferroni adjusted significance level.*

SLFN12) was located on chromosome 17, BP: 33,737,940–33,760,195. The rare variant association test achieved statistical significance after Bonferroni correction ($p = 2.40 \times 10^{-8}$). The association was highly significant in "OncoArray" population ($p = 5.05 \times 10^{-9}$) and was replicated in "HumanCore" population ($p = 3.73 \times 10^{-3}$).

We further looked into the significant findings within each population. In **Table 5**, we summarized the regions that were identified in a particular population by both rare variants and all variants association test. In **Table 6**, we summarized the regions that were identified by rare variants association test only. The *p*-values computed in five populations for these regions were

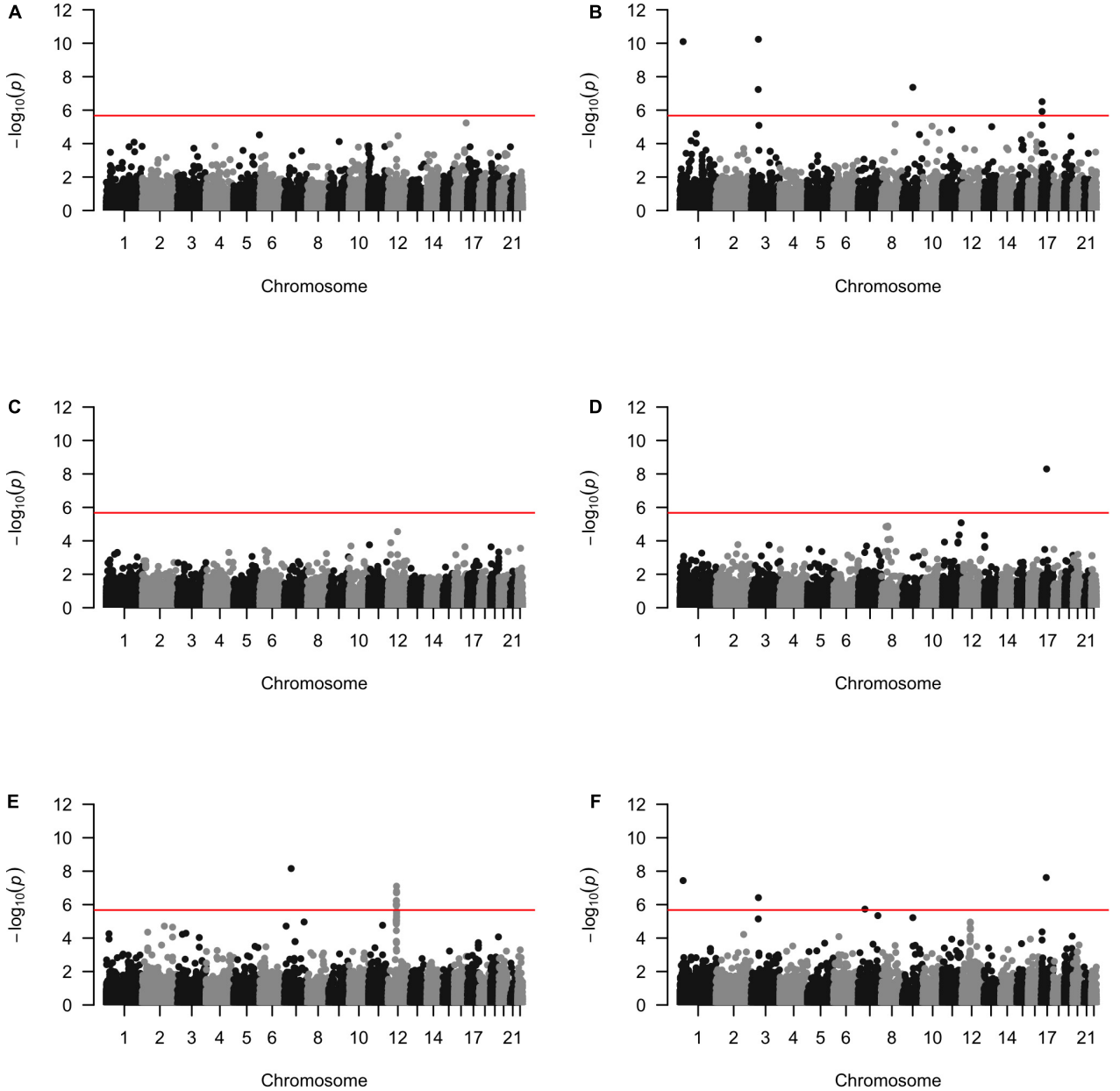


FIGURE 2 | The Manhattan plots by rare variants analysis in each population **(A)** Affymetrix. **(B)** Illumina. **(C)** OmniExpress. **(D)** OncoArray. **(E)** HumanCore. **(F)** Fisher.

summarized in **Supplementary Tables 1, 2**. In particular, the results suggested that multiple gene regions on chromosome 12 and chromosomes 17 were identified for association with the disease outcome. For example, two regions close to each other on chromosome 17 (gene *LOC101928000*, BP: 5,015,229–5,017,677 and gene *USP6*, BP: 5,019,732–5,078,326) were identified for both rare and all variants association. A different region on chromosome 17 was identified for rare variants association. While the underlying genetic mechanism and causal SNPs were not clear, we think the rare variants association test may provide findings that are complementary to existing GWAS that usually are limited to relatively common variants. For common variants analysis, we were not able to identify any regions after Bonferroni adjustment. In **Table 7**, we summarized regions with suggestive significance (i.e., 10^{-5}) in a particular population. In particular, the association for region *SPATA2L* was marginally significant in "OmniExpress" and was also nominally significant in both "Illumina" and "OncoArray."

Cross-Check With Expression Quantitative Trait Loci (eQTL) Database

To provide additional insights on the possible involvement of these identified regions in regulating gene expression, we

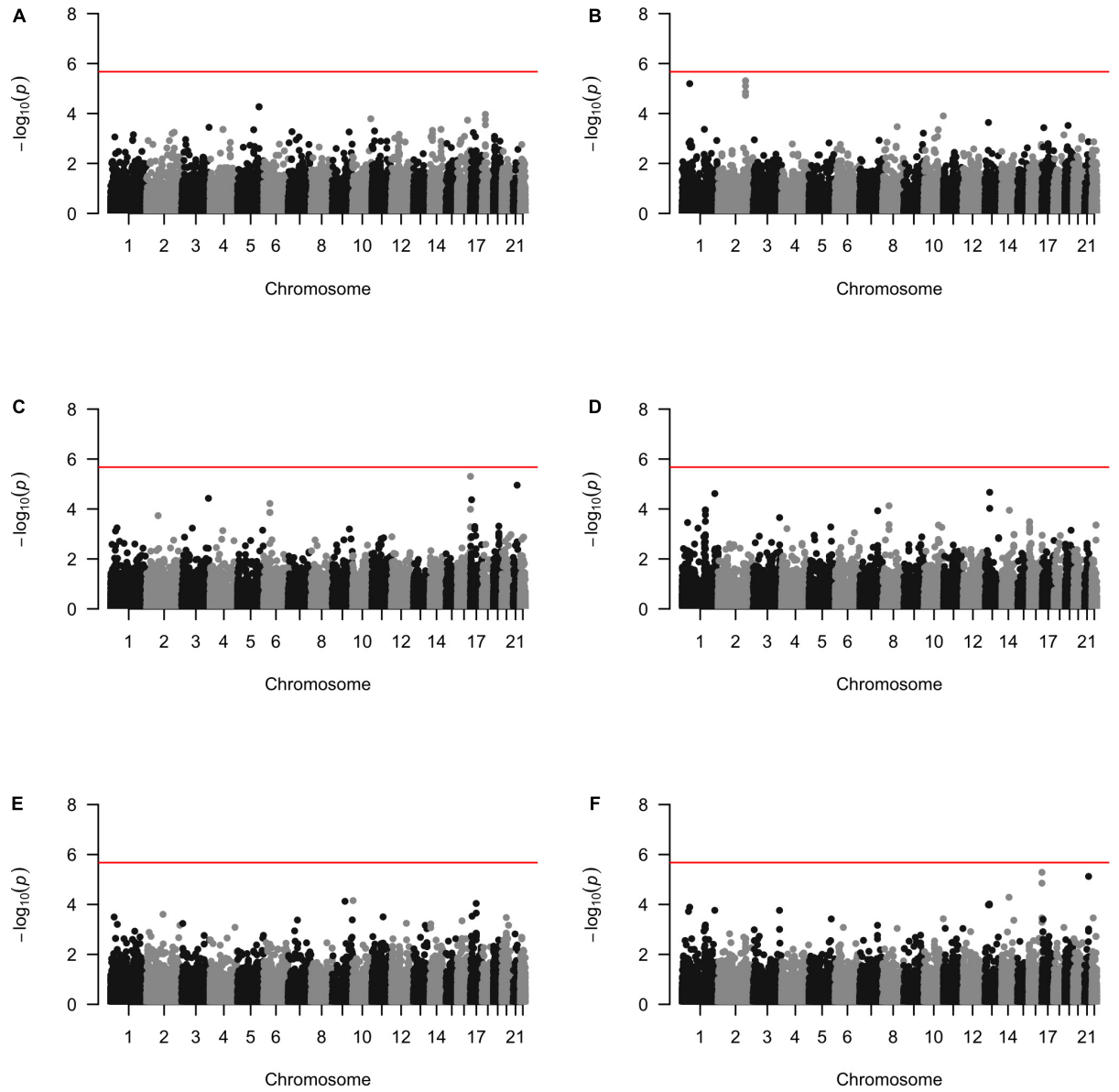


FIGURE 3 | The Manhattan plots by common variants analysis in each population **(A)** Affymetrix. **(B)** Illumina. **(C)** OmniExpress. **(D)** OncoArray. **(E)** HumanCore. **(F)** Fisher.

summarized the number of known eQTLs within each region (**Table 8**). Most of those loci (15 out of 18) included at least one eQTL either in not-sun-exposed or sun-exposed skin tissues. Among 24,279 regions being tested, a total of 16,534 contained at least one eQTL in the GTEx database. To evaluate the overrepresentation of eQTL in the identified region, we calculated an exact *p*-value using a hyper-genomic distribution as:

$$p_{val} = \sum_{k=15}^{k=18} \frac{\binom{16,534}{k}\binom{24,279-16,534}{18-k}}{\binom{24,279}{16,534}} = 0.126$$

It is also worthwhile to note that most of existing studies of eQTL were also based on single-locus association test between each genetic variants and gene expression data. Though the *p*-value was not statistically significant at 0.05 level, the large proportion of identified regions harboring known eQTL suggested their possible involvement of gene expression within skin tissues.

DISCUSSION

In this study, we identified 18 cSCC-associated genomic regions using gene-based fastSKAT method. One region (i.e., *SLFN12*)

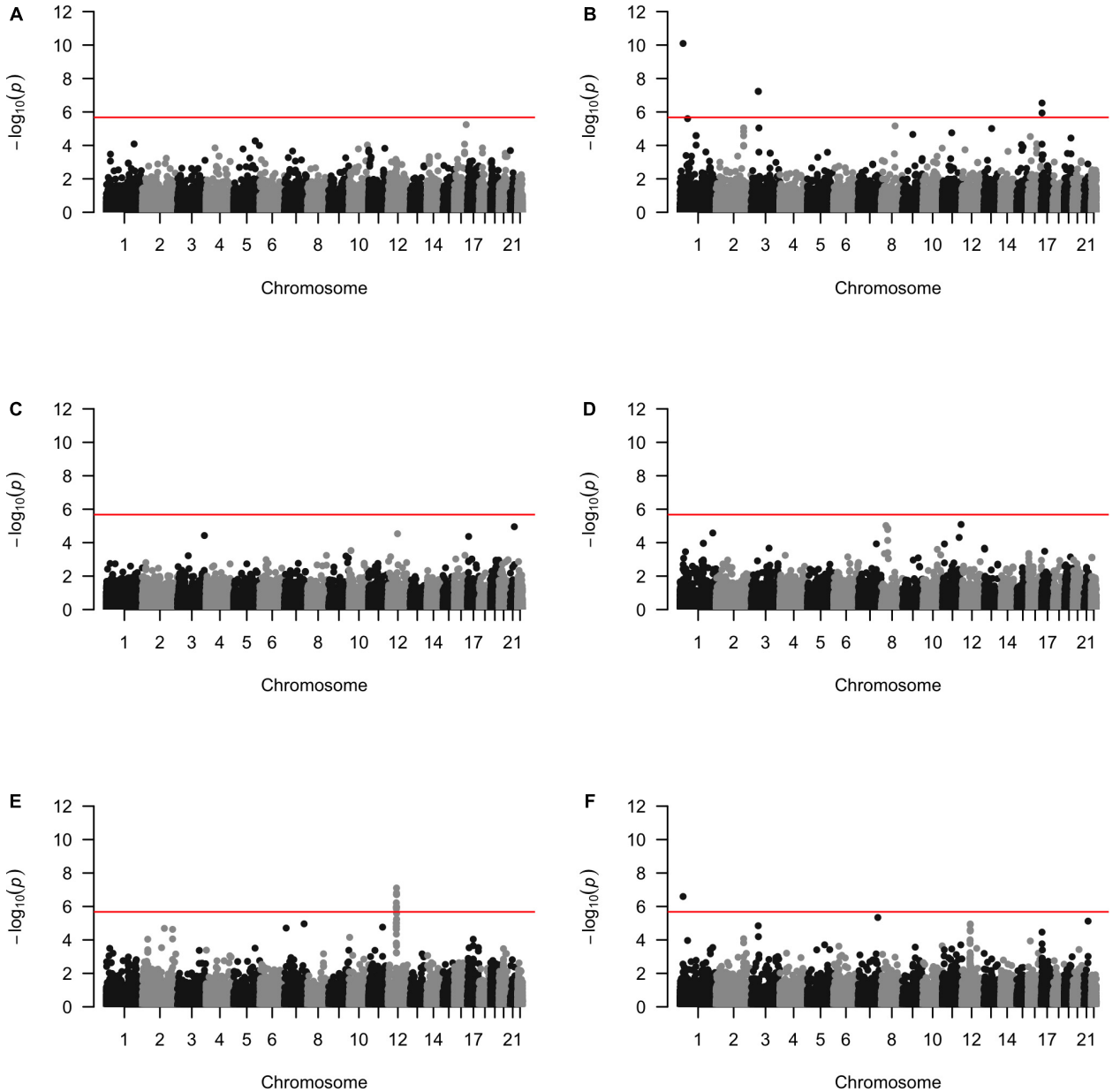


FIGURE 4 | The Manhattan plots by all variants analysis in each population **(A)** Affymetrix. **(B)** Illumina. **(C)** OmniExpress. **(D)** OncoArray. **(E)** HumanCore. **(F)** Fisher.

was statistically significant in one population and replicated in another population. The eQTL analysis further supported the possible biological contribution of those regions to the genetic susceptibility of cSCC. The replication of previous GWAS-identified SNPs also demonstrated the reliability of fastSKAT in identifying susceptibility loci with improved statistical power. To our knowledge, our study is among the first ones to investigate the region-based association for cSCC on a genome-wide level.

As an effective and powerful tool, GWAS has been commonly used to investigate the genetic architecture of complex diseases, including squamous cell carcinoma. The goal of our study is to provide a complementary strategy to address a few limitations of the GWAS, especially to evaluate the rare variants with low frequencies in the populations. In our study, although the total sample size was relatively large (∼26K), the number of cases were relatively small in each sub-population (<800). In such a situation, the single-locus-based GWAS is expected to be under-powered to identify rare variants (Tong et al., 2011; Mo et al., 2015). In addition, the highly unbalanced numbers of cases and controls may also present additional challenge to both conventional GWAS and rare-variants association tests. Recent studies have suggested that the number of cases and case to control ratio may both have an impact on the statistical power and type I errors, especially under large control group scenarios

TABLE 4 | Regions identified by Fisher's combined probability test after Bonferroni adjustment.

	Chro	Regions	Gene	*p*-value					
				Affy	Illumina	Omni	Onco	HumanCore	Fisher
Rare variants analysis	1	21,069,170–21,113,181	*HP1BP1*	7.90×10^{-1}	**7.97×10^{-11}**	8.47×10^{-1}	3.62×10^{-1}	6.99×10^{-2}	**3.65×10^{-8}**
	3	49,506,135–49,573,051	*DAG1*	8.62×10^{-1}	**5.80×10^{-11}**	8.30×10^{-1}	7.00×10^{-1}	7.32×10^{-1}	**3.83×10^{-7}**
	7	45,763,385–45,808,617	*SEPT7P2*	5.35×10^{-1}	7.72×10^{-1}	1.07×10^{-1}	4.56×10^{-1}	**6.94×10^{-9}**	**1.86×10^{-6}**
	17	33,737,940–33,760,195	*SLFN12*	1.64×10^{-1}	6.11×10^{-1}	4.38×10^{-1}	**5.05×10^{-9}**	**3.73×10^{-3}**	**2.40×10^{-8}**
All variants analysis	1	21,069,170–21,113,181	*HP1BP1*	8.29×10^{-1}	**8.03×10^{-11}**	5.86×10^{-1}	9.51×10^{-1}	3.52×10^{-1}	**2.54×10^{-7}**

Bold values indicate significant association after Bonferroni adjustment in the discovery phase or nominal significant association in the replication phase.

TABLE 5 | Regions identified by both rare and all variants analysis in a particular population after Bonferroni adjustment.

Population	Chro	Regions	Gene	Rare variants analysis			All variants analysis		
				p-value in this population	Fisher's *p*-value	# of SNPs in region	*p*-value in this population	Fisher's *p*-value	#of SNPs in region
Illumina	1	21,069,170–21,113,181	*HP1BP3*	7.97×10^{-11}	3.65×10^{-8}	224	8.03×10^{-11}	2.54×10^{-7}	296
	3	48,445,260–48,471,460	*PLXNB1*	5.82×10^{-8}	7.17×10^{-6}	155	5.82×10^{-8}	1.43×10^{-5}	187
	3	49,506,135–49,573,051	*DAG1*	5.80×10^{-11}	3.83×10^{-7}	169	5.99×10^{-8}	6.37×10^{-5}	304
	17	5,015,229–5,017,677	*LOC101928000*	1.20×10^{-6}	4.25×10^{-5}	78	1.14×10^{-6}	1.72×10^{-4}	119
	17	5,019,732–5,078,326	*USP6*	3.11×10^{-7}	1.31×10^{-4}	253	2.92×10^{-7}	3.43×10^{-5}	406
HumanCore	12	56,512,003–56,516,280	*ZC3H10*	9.95×10^{-7}	1.37×10^{-4}	54	1.05×10^{-6}	1.16×10^{-4}	71
	12	56,521,985–56,538,460	*ESYT1*	1.14×10^{-6}	1.68×10^{-4}	102	1.16×10^{-6}	1.66×10^{-4}	122
	12	56,546,203–56,551,771	*MYL6B*	6.04×10^{-7}	7.77×10^{-5}	61	6.04×10^{-7}	9.85×10^{-5}	76
	12	56,660,641–56,664,750	*COQ10A*	5.68×10^{-7}	9.10×10^{-5}	27	1.38×10^{-6}	5.74×10^{-4}	53
	12	57,623,355–57,628,718	*SHMT2*	1.57×10^{-7}	2.49×10^{-5}	70	1.57×10^{-7}	2.49×10^{-5}	86
	12	57,628,685–57,634,475	*NDUFA4L2*	1.90×10^{-7}	2.81×10^{-5}	52	1.90×10^{-7}	2.81×10^{-5}	66
	12	57,637,237–57,644,976	*STAC3*	7.88×10^{-8}	1.23×10^{-5}	55	7.88×10^{-8}	1.23×10^{-5}	70
	12	57,647,546–57,824,788	*R3HDM2*	1.96×10^{-7}	1.10×10^{-5}	501	1.96×10^{-7}	1.11×10^{-5}	729
	12	57,828,467–57,845,845	*INHBC*	1.06×10^{-6}	2.94×10^{-5}	85	1.06×10^{-6}	2.94×10^{-5}	133

TABLE 6 | Regions identified by rare variants analysis in a particular population after Bonferroni adjustment.

Population	Chro	Regions	Gene	Rare variants analysis		
				p-value in this population	Fisher's *p*-value	# of SNPs in region
Illumina	9	71,650,478–71,715,094	*FXN*	4.32×10^{-8}	6.01×10^{-6}	394
Onco	17	33,737,940–33,760,195	*SLFN12*	5.05×10^{-9}	2.40×10^{-8}	154
HumanCore	7	45,763,385–45,808,617	*SEPT7P2*	6.94×10^{-9}	1.86×10^{-6}	97
HumanCore	12	56,631,590–56,652,143	*ANKRD52*	9.60×10^{-7}	1.50×10^{-4}	49

(Zhang et al., 2019). It was also found that SKAT can reach reasonably high power with well-controlled type I error if the number of cases is larger than 200. In our study, the number of cases ranged between ∼200 and 700 across five subpopulations, and the results appeared to be consistent with previous studies. The QQ-plot and estimated genomic inflation factors suggested well-controlled type I errors. While we expect the statistical power will improve with additional cases, the current results also suggested that region-based association test was able to identify genomic regions though rare variants association.

A number of gene units were identified to harbor genetic variants that may contribute to the susceptibility of cSCC. One gene was identified with replicated association in two sub-populations. Gene *SLFN12*, or Schlafen family member 12, belongs to a group of genes mediating growth-inhibition as cell cycle regulators (Katsoulidis et al., 2010). Many studies have found that *SLFN12* played a key role in generating anti-tumor effects triggered by certain drugs or interventions (Katsoulidis et al., 2010; An et al., 2019; Lewis et al., 2019). For example, the drug Anagrelide (ANA) can only inhibits cancer cell growth when both *PED3A* and *SLFN12* are expressed.

A number of other gene units were identified to be associated with cSCC in one population without replication. However, they have been reported in the literature for involvement with cancer development. For example, the identified gene units *HP1BP1* and *SEPT7P2* have been found to be involved in cancer growth and progression (Dutta et al., 2014; Wang et al., 2019). In addition, gene *SPATA2L* have been identified to be associated

TABLE 7 | Regions reaching suggestive significance level of 10^{-5} by common variants analysis.

Identification platform	Chro	Regions	Gene	*p*-values in each population					
				Affy	Illumina	Omni	Onco	Human core	Fisher
Illumina	1	52,254,865–52,344,609	*NRDC, MIR761*	2.95×10^{-1}	**6.39×10^{-6}**	2.50×10^{-1}	1.13×10^{-1}	4.91×10^{-1}	**1.29×10^{-4}**
	2	190,627,505–190,630,282	*OSGEPL1-AS1*	3.97×10^{-1}	**7.95×10^{-6}**	8.37×10^{-1}	8.93×10^{-1}	8.73×10^{-1}	**3.50×10^{-3}**
	2	190,634,992–190,649,097	*ORMDL1*	4.16×10^{-1}	**4.94×10^{-6}**	7.25×10^{-1}	9.57×10^{-1}	8.96×10^{-1}	**2.47×10^{-3}**
	2	190,648,810–190,742,355	*PMS1*	4.15×10^{-1}	**4.93×10^{-6}**	7.25×10^{-1}	9.57×10^{-1}	8.96×10^{-1}	**2.47×10^{-3}**
Omni	16	89,762,764–89,768,121	*SPATA2L*	7.03×10^{-1}	**2.56×10^{-2}**	**4.96×10^{-6}**	**2.77×10^{-2}**	1.96×10^{-1}	**5.19×10^{-6}**
Fisher	21	42,513,426–42,519,991	*LINC00323*	5.21×10^{-1}	**7.41×10^{-3}**	**1.11×10^{-5}**	3.02×10^{-1}	5.87×10^{-2}	**7.54×10^{-6}**

No regions were genome-wide significant after Bonferroni adjustment.
Bold values indicate suggestive association in the discovery phase or nominal significant association in the replication phase.

TABLE 8 | Number of eQTLs located within identified regions in skin tissues exposed or not exposed to sun.

Population	Chro	Regions	Gene	Number of eQTLs within region	
				Skin not exposed to sun	Skin exposed to sun
Illumina	1	21,069,170–21,113,181	*HP1BP3*	0	0
	3	48,445,260–48,471,460	*PLXNB1*	2	2
	3	49,506,135–49,573,051	*DAG1*	3	3
	17	5,015,229–5,017,677	*LOC101928000*	0	2
	17	5,019,732–5,078,326	*USP6*	1	1
HumanCore	12	56,512,003–56,516,280	*ZC3H10*	0	1
	12	56,521,985–56,538,460	*ESYT1*	2	1
	12	56,546,203–56,551,771	*MYL6B*	2	0
	12	56,660,641–56,664,750	*COQ10A*	2	4
	12	57,623,355–57,628,718	*SHMT2*	2	0
	12	57,628,685–57,634,475	*NDUFA4L2*	0	0
	12	57,637,237–57,644,976	*STAC3*	0	2
	12	576,47,546–57,824,788	*R3HDM2*	2	4
	12	57,828,467–57,845,845	*INHBC*	0	0
Illumina	9	71,650,478–71,715,094	*FXN*	1	2
Onco	17	33,737,940–33,760,195	*SLFN12*	3	4
HumanCore	7	45,763,385–45,808,617	*SEPT7P2*	3	1
HumanCore	12	56,631,590–56,652,143	*ANKRD52*	3	3

with vitiligo in a recent study (Cai et al., 2021), and the inverse relationship between vitiligo and NMSC was suggested in many research (Paradisi et al., 2014; Rodrigues, 2017; Wu et al., 2018; Wen et al., 2020).

A number of other methods were also available for region-based association test. For example, we and others have proposed a generalized genetic random field (GGRF) method for testing the association between a set of variants and a disease phenotype (Li et al., 2014). The proposed GGRF is a similarity-based method. It maps subjects to a Euclidean space using on their genotypes as coordinates, so that subjects who are close to each other in space would have similar phenotype if there is a gene-phenotype association (Li et al., 2014). GGRF used a Wald-type of test statistic and may achieve improved power over SKAT under various disease scenario. However, fastSKAT used a score test and is more computationally efficient with the approximation by random projection. In this study, we have used fastSKAT for analysis and we showed in **Appendix**, GGRF would be equivalent to SKAT if a generalized score test is used.

Our study must be considered in the light of certain limitations. First, none of the association was consistently replicated in all populations. This is partly due to the heterogeneous nature of rare variants and their low allele frequencies across populations. Multiple rare mutations within the same gene can independently influence the disease (i.e., allelic heterogeneity), and rare variants in different genes can also be involved in related pathways underlying complex human diseases (i.e., locus heterogeneity) (McClellan and King, 2010). Second, due to the nature of gene-based analysis, it is not straightforward to ascertain the causal SNPs or estimate their effect on cSCC risk. We also have not considered intergenic variants that were not within the gene regions (Mo et al., 2015). Third, the existing findings based on region-based association have been limited. For example, the eQTL variants available in GTEx database were mainly identified via single-locus analysis.

Additional functional analysis is needed to validate the identified regions in the future. Forth, we are also aware that the results are subject to the strengths and limitations of fastSKAT due to its assumptions and implementation. For example, we have used a weight function that is inversely correlated with the MAF of each variant (i.e., probability density of beta distribution, default option of fastSKAT). It is often helpful to incorporate functional annotation of the variants to upweight those with potentially stronger effect on the disease (Kumar et al., 2009; Lee et al., 2015; Quick et al., 2019). Further, extensions of SKAT, such as SKAT-O, were able to effectively combine the test statistics of SKAT and burden test (Lee et al., 2012), which may have improved power when the causal variants have the same direction of effects. We have adopted fastSKAT mainly because of the computational advantage for studies with a very large number of subjects and variants. It can also be helpful to improve the power in other scenarios when SKAT-O becomes feasible for extremely large studies. Fifth, no genomic region was identified by common variants analysis after Bonferroni adjustment. It is partly because the weight function adopted gave more weight to variants with low MAF and regions with common variants receiving less weight may not be able to identify. Furthermore, region-based test would be less powerful when there are a few susceptible loci with effects in this region and the total number of tested SNPs is large.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: GWAS data has not been publicly available. Further information including the procedures to obtain and access data from the Nurses' Health Studies and Health Professionals Follow-up Study is described at https://www.nurseshealthstudy.org/researchers (contact email: nhsaccess@channing.harvard.edu) and https://sites.sph.harvard.edu/hpfs/for-collaborators/. The expression quantitative trait loci (eQTL) database are openly available from the Genotype-Tissue Expression (GTEx) project at https://www.gtexportal.org/home/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MH and ML conceived and designed the analysis. JH and AQ collected the data. MH, CL, XL, AQ, JH, and ML contributed data and analysis tools and wrote the manuscript. MH, CL, and ML performed the analysis. All authors have read and approved the manuscript.

FUNDING

This study was supported, in part, by the National Heart, Lung and Blood Institute under award number K01HL140333 (ML), the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development under award number R03HD092854 (ML), and the National Cancer Institute under award number UM1CA186107, P01CA87969, R01CA49449, U01CA176726, R01CA67262, and U01CA167552. The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of Health.

ACKNOWLEDGMENTS

We would like to thank the participants and staff of the NHS, the NHS II and the HPFS, for their valuable contributions, as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. We also want to thank Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States for data sharing. We assume full responsibility for analyses and interpretation of these data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.657499/full#supplementary-material

REFERENCES

An, R., Liu, J., He, J., Wang, F., Zhang, Q., and Yu, Q. (2019). PDE3A inhibitor anagrelide activates death signaling pathway genes and synergizes with cell death-inducing cytokines to selectively inhibit cancer cell growth. *Am. J. Cancer Res.* 9, 1905–1921.

Asgari, M. M., Wang, W., Ioannidis, N. M., Itnyre, J., Hoffmann, T., Jorgenson, E., et al. (2016). Identification of susceptibility loci for cutaneous squamous cell carcinoma. *J. Investig. Dermatol.* 136, 930–937. doi: 10.1016/j.jid.2016.01.013

Asgari, M. M., Warton, E. M., and Whittemore, A. S. (2015). Family history of skin cancer is associated with increased risk of cutaneous squamous cell carcinoma. *Dermatol. Surg.* 41, 481–486. doi: 10.1097/dss.0000000000000292

Boos, D. D. (1992). On generalized score tests. *Am. Stat.* 46, 327–333.

Cai, M., Yuan, T., Huang, H., Gui, L., Zhang, L., Meng, Z., et al. (2021). Integrative analysis of omics data reveals regulatory network of CDK10 in vitiligo risk. *Front. Genet.* 12:634553. doi: 10.3389/fgene.2021.634553

Chahal, H. S., Lin, Y., Ransohoff, K. J., Hinds, D. A., Wu, W., Dai, H. J., et al. (2016). Genome-wide association study identifies novel susceptibility loci for cutaneous

squamous cell carcinoma. *Nat. Commun.* 7:12048. doi: 10.1038/ncomms12048

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. doi: 10.1186/s13742-015-0047-8

Chitsazzadeh, V., Coarfa, C., Drummond, J. A., Nguyen, T., Joseph, A., Chilukuri, S., et al. (2016). Cross-species identification of genomic drivers of squamous cell carcinoma development across preneoplastic intermediates. *Nat. Commun.* 7:12601. doi: 10.1038/ncomms12601

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi: 10.1038/ng.3656

Duffy, D. L., Zhu, G., Li, X., Sanna, M., Iles, M. M., Jacobs, L. C., et al. (2018). Novel pleiotropic risk loci for melanoma and nevus density implicate multiple biological pathways. *Nat. Commun.* 9:4774. doi: 10.1038/s41467-018-06649-5

Dutta, B., Yan, R., Lim, S. K., Tam, J. P., and Sze, S. K. (2014). Quantitative profiling of chromatome dynamics reveals a novel role for HP1BP3 in hypoxia-induced oncogenesis. *Mol. Cell. Proteom.* 13, 3236–3249. doi: 10.1074/mcp.M114.038232

GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* 53, 217–288. doi: 10.1137/090771806

Hussain, S. K., Sundquist, J., and Hemminki, K. (2009). The effect of having an affected parent or sibling on invasive and in situ skin cancer risk in Sweden. *J. Investig. Dermatol.* 129, 2142–2147. doi: 10.1038/jid.2009.31

Karagas, M. R., Greenberg, E. R., Spencer, S. K., Stukel, T. A., and Mott, L. A. (1999). Increase in incidence rates of basal cell and squamous cell skin cancer in New Hampshire, USA. New Hampshire skin cancer study group. *Int. J. Cancer* 81, 555–559. doi: 10.1002/(sici)1097-0215(19990517)81:4<555::aid-ijc9<3.0.co;2-r

Katsoulidis, E., Mavrommatis, E., Woodard, J., Shields, M. A., Sassano, A., Carayol, N., et al. (2010). Role of interferon {alpha} (IFN{alpha})-inducible Schlafen-5 in regulation of anchorage-independent growth and invasion of malignant melanoma cells. *J. Biol. Chem.* 285, 40333–40341. doi: 10.1074/jbc.M110.151076

Kindlon ARQaN (2009-2019). *Bedtools [Computer Software]*.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081. doi: 10.1038/nprot.2009.86

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., et al. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* 47, 955–961. doi: 10.1038/ng.3331

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237. doi: 10.1016/j.ajhg.2012.06.007

Lewis, T. A., de Waal, L., Wu, X., Youngsaye, W., Wengner, A., Kopitz, C., et al. (2019). Optimization of PDE3A modulators for SLFN12-dependent cancer cell killing. *ACS Med. Chem. Lett.* 10, 1537–1542. doi: 10.1021/acsmedchemlett.9b00360

Li, M., He, Z., Zhang, M., Zhan, X., Wei, C., Elston, R. C., et al. (2014). A generalized genetic random field method for the genetic association analysis of sequencing data. *Genet. Epidemiol.* 38, 242–253. doi: 10.1002/gepi.21790

Liang, K.-Y., and Zeger, S. L. (1989). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

Lindström, S., Loomis, S., Turman, C., Huang, H., Huang, J., Aschard, H., et al. (2017). A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts. *PLoS One* 12:e0173997. doi: 10.1371/journal.pone.0173997

Lumley, T., Brody, J., Peloso, G., Morrison, A., and Rice, K. (2018). FastSKAT: sequence kernel association tests for very large sets of markers. *Genet. Epidemiol.* 42, 516–527. doi: 10.1002/gepi.22136

McClellan, J., and King, M. C. (2010). Genetic heterogeneity in human disease. *Cell* 141, 210–217. doi: 10.1016/j.cell.2010.03.032

Mo, X. B., Lu, X., Zhang, Y. H., Zhang, Z. L., Deng, F. Y., and Lei, S. F. (2015). Gene-based association analysis identified novel genes associated with bone mineral density. *PLoS One* 10:e0121811. doi: 10.1371/journal.pone.0121811

Motaparthi, K., Kapil, J. P., and Velazquez, E. F. (2017). Cutaneous squamous cell carcinoma: review of the eighth edition of the American joint committee on cancer staging guidelines, prognostic factors, and histopathologic variants. *Adv. Anat. Pathol.* 24, 171–194. doi: 10.1097/pap.0000000000000157

Muzic, J. G., Schmitt, A. R., Wright, A. C., Alniemi, D. T., Zubair, A. S., Olazagasti Lourido, J. M., et al. (2017). Incidence and trends of basal cell carcinoma and cutaneous squamous cell carcinoma: a population-based study in Olmsted County, Minnesota, 2000 to 2010. *Mayo Clin. Proc.* 92, 890–898. doi: 10.1016/j.mayocp.2017.02.015

Nan, H., Xu, M., Kraft, P., Qureshi, A. A., Chen, C., Guo, Q., et al. (2011). Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma. *Hum. Mol. Genet.* 20, 3718–3724. doi: 10.1093/hmg/ddr287

Nettiksimmons, J., Tranah, G., Evans, D. S., Yokoyama, J. S., and Yaffe, K. (2016). Gene-based aggregate SNP associations between candidate AD genes and cognitive decline. *Age (Dordrecht, Netherlands)* 38:41. doi: 10.1007/s11357-016-9885-2

Nguyen, K. D., Han, J., Li, T., and Qureshi, A. A. (2014). Invasive cutaneous squamous cell carcinoma incidence in US health care workers. *Arch. Dermatol. Res.* 306, 555–560. doi: 10.1007/s00403-014-1469-3

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* 10:e1004234. doi: 10.1371/journal.pgen.1004234

Paradisi, A., Tabolli, S., Didona, B., Sobrino, L., Russo, N., and Abeni, D. (2014). Markedly reduced incidence of melanoma and nonmelanoma skin cancer in a nonconcurrent cohort of 10,040 patients with vitiligo. *J. Am. Acad. Dermatol.* 71, 1110–1116. doi: 10.1016/j.jaad.2014.07.050

Parekh, V., and Seykora, J. T. (2017). Cutaneous squamous cell carcinoma. *Clin. Lab. Med.* 37, 503–525. doi: 10.1016/j.cll.2017.06.003

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Human Genet.* 81, 559–575. doi: 10.1086/519795

Que, S. K. T., Zwald, F. O., and Schmults, C. D. (2018b). Cutaneous squamous cell carcinoma: management of advanced and high-stage tumors. *J. Am. Acad. Dermatol.* 78, 249–261. doi: 10.1016/j.jaad.2017.08.058

Que, S. K. T., Zwald, F. O., and Schmults, C. D. (2018a). Cutaneous squamous cell carcinoma: incidence, risk factors, diagnosis, and staging. *J. Am. Acad. Dermatol.* 78, 237–247. doi: 10.1016/j.jaad.2017.08.059

Quick, C., Wen, X., Abecasis, G., Boehnke, M., and Kang, H. M. (2019). Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis. *bioRxiv* [Preprint]. doi: 10.1101/732404 732404,

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxf. Engl.)* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rodrigues, M. (2017). Skin cancer risk (Nonmelanoma skin cancers/Melanoma) in vitiligo patients. *Dermatol. Clin.* 35, 129–134. doi: 10.1016/j.det.2016.11.003

Sarin, K. Y., Lin, Y., Daneshjou, R., Ziyatdinov, A., Thorleifsson, G., Rubin, A., et al. (2020). Genome-wide meta-analysis identifies eight new susceptibility loci for cutaneous squamous cell carcinoma. *Nat. Commun.* 11:820. doi: 10.1038/s41467-020-14594-5

Siiskonen, S. J., Zhang, M., Li, W. Q., Liang, L., Kraft, P., Nijsten, T., et al. (2016). A genome-wide association study of cutaneous squamous cell carcinoma among European descendants. *Cancer Epidemiol. Biomark. Prevent.* 25, 714–720. doi: 10.1158/1055-9965

Tong, L., Tayo, B., Yang, J., and Cooper, R. S. (2011). Comparison of SNP-based and gene-based association studies in detecting rare variants using unrelated individuals. *BMC Proc.* 5 Suppl. 9(Suppl. 9):S41. doi: 10.1186/1753-6561-5-s9-s41

Tropp, J. A. (2011). Improved analysis of the subsampled randomized hadamard transform. *Adv. Adapt. Data Anal.* 03, 115–116.

Waldman, A., and Schmults, C. (2019). Cutaneous squamous cell carcinoma. *Hematol. Oncol. Clin. North Am.* 33, 1–12. doi: 10.1016/j.hoc.2018.08.001

Wang, J., Xie, G. F., He, Y., Deng, L., Long, Y. K., Yang, X. H., et al. (2019). Interfering expression of chimeric transcript SEPT7P2-PSPH promotes cell proliferation in patients with nasopharyngeal carcinoma. *J. Oncol.* 2019, 1654724. doi: 10.1155/2019/1654724

Wen, Y., Wu, X., Peng, H., Li, C., Jiang, Y., Liang, H., et al. (2020). Cancer risks in patients with vitiligo: a Mendelian randomization study. *J Cancer Res. Clin. Oncol.* 146, 1933–1940. doi: 10.1007/s00432-020-03245-3

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., et al. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* 86, 929–942. doi: 10.1016/j.ajhg.2010.05.002

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Wu, W., Amos, C. I., Lee, J. E., Wei, Q., Sarin, K. Y., and Han, J. (2018). Inverse relationship between vitiligo-related genes and skin cancer risk. *J. Investig. Dermatol.* 138, 2072–2075. doi: 10.1016/j.jid.2018.03.1511

Zhang, X., Basile, A. O., Pendergrass, S. A., and Ritchie, M. D. (2019). Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics* 20:46. doi: 10.1186/s12859-018-2591-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

In our study, a fastSKAT method was applied to test the association between each genomic region and disease outcome. A number of other methods were also available for region-based association test. For example, we and others have proposed a generalized genetic random field (GGRF) method for testing the association between a set of variants and a disease phenotype (Li et al., 2014), and compared its performance to that of SKAT. We described below that GGRF would have similar test statistic with SKAT if a generalized score test is used for inference.

Suppose the study include a total of N subjects, each with K variants in a region and M covariates. Let Y, G, X denotes the phenotype ($N = 1$), genotype ($N = K$), and covariates ($N = M$) matrix, respectively. The GGRF adopts a conditional autoregression model as:

$$E\left(Y \mid Y_{-}\right) = \mu\gamma S\left(Y - \mu\right), ;$$

Where the i-th element of Y_{-} denotes the phenotype of all other subjects other than i-th subject, $\mu = f(X\beta)$ is used for covariants adjustment, and S is a matrix for pairwise genetic similarity among N subjects. To test the genotype-phenotype association ($H_0 : \gamma = 0$), a generalized score test can be used (Liang and Zeger, 1989), so that:

$$U_{\gamma}\left(\beta, \gamma\right) = \frac{\partial E\left(Y \mid Y_{-}\right)^{T}}{\partial\gamma}\left\{Y - E\left(Y \mid Y_{-}\right)\right\} = \left(Y - \mu\right)^{T} S\left\{I - \gamma S\right\}\left(Y - \mu\right) = 0;$$

A generalized score statistic can thus be defined as (Boos, 1992)

$$Q = U_{\gamma}\left(\widehat{\beta}, 0\right) = \left(Y - \widehat{\mu}\right)' S\left(Y - \widehat{\mu}\right);$$

where $\widehat{\beta}$ is estimated under the null hypothesis that $\gamma = 0$ via a generalized linear model. The score statistic $\frac{1}{m}Q$ takes the same format with that of SKAT, and follows asymptotically a mixture of Chi-square distributions (Wu et al., 2011).

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fgene.2021.688871

High-Dimensional Mediation Analysis With Confounders in Survival Models

*Zhangsheng Yu[1,2,3], Yidan Cui[1,2], Ting Wei[1,2], Yanran Ma[1,2] and Chengwen Luo[1,2]**

[1] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2] SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China, [3] Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Mediation analysis is a common statistical method for investigating the mechanism of environmental exposures on health outcomes. Previous studies have extended mediation models with a single mediator to high-dimensional mediators selection. It is often assumed that there are no confounders that influence the relations among the exposure, mediator, and outcome. This is not realistic for the observational studies. To accommodate the potential confounders, we propose a concise and efficient high-dimensional mediation analysis procedure using the propensity score for adjustment. Results from simulation studies demonstrate the proposed procedure has good performance in mediator selection and effect estimation compared with methods that ignore all confounders. Of note, as the sample size increases, the performance of variable selection and mediation effect estimation is as well as the results shown in the method which include all confounders as covariates in the mediation model. By applying this procedure to a TCGA lung cancer data set, we find that lung cancer patients who had serious smoking history have increased the risk of death *via* the methylation markers cg21926276 and cg20707991 with significant hazard ratios of 1.2093 (95% CI: 1.2019–1.2167) and 1.1388 (95% CI: 1.1339–1.1438), respectively.

Keywords: high-dimensional mediators, confounders, survival model, mediation analysis, propensity score

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong, Hong Kong

Reviewed by:
Lin Hou,
Tsinghua University, China
Xihao Li,
Harvard University, United States

***Correspondence:**
Chengwen Luo
luochengwen@sjtu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *31 March 2021*
Accepted: *07 June 2021*
Published: *28 June 2021*

Citation:
Yu Z, Cui Y, Wei T, Ma Y and Luo C (2021) High-Dimensional Mediation Analysis With Confounders in Survival Models. Front. Genet. 12:688871. doi: 10.3389/fgene.2021.688871

INTRODUCTION

Mediation analysis was firstly used to deal with the causal chain of events as the primary exposure has an effect on the outcome through affecting one or more mediators in psychological studies, and gradually extended to sociological and biomedical researches (Baron and Kenny, 1986; MacKinnon et al., 2002; Preacher and Hayes, 2008; Biesanz et al., 2010; Huan et al., 2016). Of note, the mediators are usually measured after the intervention, but before the main outcome of interest. Mediation effect is often assessed through a regression-based analysis procedure by decomposing the total effect that describes the relationship between the exposure and the outcome variable into direct effect and indirect effect (Baron and Kenny, 1986; MacKinnon et al., 2007). In the past couple of decades, the topic of mediation analysis has received a great deal of attention, particularly in the area of causal inference (Robins and Greenland, 1992; Ten Have et al., 2007; Albert, 2008; Sobel, 2008; VanderWeele, 2009; Pearl, 2014). Researches in mediation analysis have been generalized from the case of a single mediator to multiple mediators (Albert and Nelson, 2011; Zhang and Wang, 2013; VanderWeele and Vansteelandt, 2014; Daniel et al., 2015), even to the case of high-dimensional mediators (Huang and Pan, 2016; Zhang et al., 2016; Zhao and Luo, 2016; Chén et al., 2018; Sohn and Li, 2019; van Kesteren and Oberski, 2019; Zhao et al., 2020). Recently, much progress

has been made in extensive of mediation methods to survival models (Lange and Hansen, 2011; VanderWeele, 2011; Wang and Zhang, 2011; Huang and Yang, 2017).

The regression-based or structural equation modeling approach is commonly used to assess mediation effect. This approach assumes that there are no confounders influencing the relationships among exposure and mediator, mediator and outcome, and exposure and outcome. Randomization to levels of the exposure guarantees that there are no confounders that influence both the relation of exposure-mediator and exposure-outcome. However, the assumption that individuals are randomly assigned to exposure, especially for research about smoking and lung cancer, is difficult to achieve.

Propensity score method can be used to solve such a problem with a non-randomized exposure which usually appears in observational studies (Rosenbaum and Rubin, 1983). Previous studies have focused on mediation analysis with confounders in the case of a single mediator. For example, Valente et al. (2017) introduced confounders in mediation analysis and described how to address confounders with design-based techniques and analysis-based approaches. Coffman (2011) proposed to use the calculated propensity score to adjust for confounders between the mediator and the outcomes. However, methods for high-dimensional mediation selection adjusting for confounders, especially for survival outcome, are still yet to be developed.

For example, in a lung cancer study, it is showed that smoking increases the risk of lung cancer patients' progression to death through DNA methylation markers (Luo et al., 2020). However, as an observational (or non-randomized) study, it is unrealistic for a subject to be randomly assigned to the exposure, as moral and ethical factors, in the research of how smoking affects the lung cancer patients' risk of progression to death mediated by DNA methylations. Therefore, the relationship among smoking status, DNA methylations, and overall survival may be confounded by baseline characteristics, such as age, gender, and other physical health indicators. However, high-dimensional mediation analysis for survival analysis subject to confounders is still to be developed.

In this paper, we study mediator selection and indirect effect estimation *via* high-dimensional mediation analysis in survival models with confounders. For observational studies, as the exposure is not randomly assigned, we propose to use the propensity score approach to adjust confounding effects. The key ideas are as follows. Firstly, we adjust for baseline confounders based on the calculated propensity score which serves as a covariate in the mediation models. Secondly, we reduce the dimension of potential mediators from ultra high-dimensional to moderate (i.e., one that is less than the sample size) using sure independence screening (SIS) method (Fan and Lv, 2008). Thirdly, we conduct variable selection *via* Cox proportional hazards model with the minimax concave penalty (MCP) (Zhang, 2010). Finally, we carry out the Sobel and joint significance test for mediation effect.

The rest of the paper proceeds as follows. In the next part, we introduce the notations and models, definition of propensity score, and develop the proposed procedure. Then, we provide simulation studies to evaluate the performance of our proposed procedure and a real data application to analyze the mediation effects of high-dimensional DNA methylation markers on the causal effect of smoking on lung cancer in an epigenome-wide study. Finally, we conclude the paper through discussing limitations and other feasibilities.

STATISTICAL METHOD

Notations and Models

For individual i, $i = 1, 2, \cdots, n$, we let D_i denote the time from onset to an event (death) and C_i be the potential censoring time. The observed survival time is $T_i = \min(D_i, C_i)$, and the failure indicator is $\delta_i = I(D_i \leq C_i)$, where $I(\cdot)$ is an indicator function. Let X_i be the exposure (smoking status, i.e., smoker or non-smoker), $M_i = (M_{1i}, M_{2i}, \cdots, M_{pi})^T$ be a p-dimensional continuous mediator vector (including all the methylation information), $p \gg n$. In observational studies, the assumption that no confounders influence the relation of exposure-mediator, mediator-outcome, and exposure-outcome is violated. Let $Z = (Z_1, \cdots, Z_m)^T$ denotes for the baseline confounders. **Figure 1** illustrates how confounders Z influence the relation of $X - M$, $M - Y$, and $X - Y$.

For survival outcome (Cox, 1972), the high-dimensional mediation models with confounders can be expressed as follows,

$$\lambda_i(t) = \lambda_0(t) \exp\left\{\gamma^* X_i + \beta^T M_i + \varphi^T Z_i\right\}, \tag{1}$$

$$M_{ki} = c_k + \alpha_k X_i + \phi_k^T Z_i + e_{ki},\ k = 1, 2, \cdots, p, \tag{2}$$

where Eq. (1) is the Cox proportional hazards model which describes the relationship between the exposure X, mediators M and the time-to-event variable; Eq. (2) characterizes how the exposure variables influence the mediators; $\lambda_0(t)$ is the baseline hazard function; γ^* is the direct effect of the exposure on the outcome; $\beta = (\beta_1, \cdots, \beta_p)^T$ is the coefficient vector relating the mediators to the outcome adjusting for the effect of exposure and confounders; $\varphi = (\varphi_1, \cdots, \varphi_m)^T$ is the coefficient vector relating the confounders to the outcome;

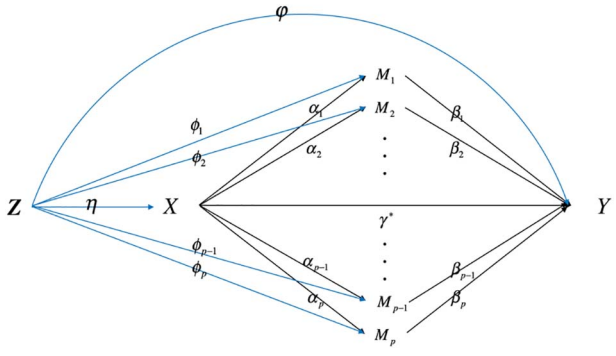


FIGURE 1 | The directed acyclic graph describes high-dimensional mediation with confounders affecting the relation among exposure, mediator, and outcome.

$\alpha = (\alpha_1, \cdots, \alpha_p)^T$ is the coefficient vector relating the exposure to the mediators; $\phi_k = (\phi_{k1}, \cdots, \phi_{km})^T$ is the coefficient vector relating the confounders to the mediator; c_k is the intercept term; $e_{ki} \sim N(0, \sigma^2)$ is the residual.

Propensity Score

The propensity score is proposed to help remove the selection bias result from potential confounders of X (Rosenbaum and Rubin, 1982). The propensity score is defined as the probability that an individual i, $i = 1, \cdots, n$ be allocated to the treatment group often estimated using logistic regression models, $\pi_i = Pr(X_i = 1 \mid Z_{1i}, \cdots, Z_{mi})$, given measured confounders $Z = (Z_1, \cdots, Z_m)^T$. This method is often used to minimize the influence of observed baseline covariates on the exposure. There are many propensity-based techniques for estimating average causal effect, including sub-classification (Rosenbaum and Rubin, 1984), matching (Rosenbaum and Rubin, 1985), and inverse propensity weighting (Robins et al., 1995). In this article, we focus on incorporating the calculated propensity score as the covariate to adjusting the confounding effects.

According to Rosenbaum and Rubin (1984), the propensity score is assessed by using baseline measured confounders as covariates in a logistic regression model with treatment status as the outcome as following

$$\text{logit}\left(P\left(X_i = 1\right)\right) = \theta_0 + \theta_1 Z_{1i} + \cdots + \theta_m Z_{mi},$$

where $\theta = (\theta_1, \cdots, \theta_m)^T$ denotes the coefficients of confounders on the exposure, and θ_0 denotes the intercept. Hence, the propensity score, π_i, the probability to be assigned to the intervention group can be expressed as

$$\pi_i = \frac{1}{1 + \exp\left(-\left(\theta_0 + \theta_1 Z_{1i} + \cdots + \theta_m Z_{mi}\right)\right)}.$$

The superiorities of propensity score over the classical regression adjustment method have been described elsewhere for the non-mediation model (Schafer and Joseph, 2008; VanderWeele, 2010). Briefly, propensity score approaches allow the inclusion of a large scale of confounders through reducing the potential covariates into a single numerical summary. More importantly, the comparison between subjects in treatment group and control group who have the same propensity score equals the comparison of control conditions with randomly assigned (Rosenbaum and Rubin, 1982).

Methodology

Since the assumption of no confounders affecting the relation among exposure, mediator and outcome is violated in observational researches, we propose a new method using the propensity score as a covariate in the high-dimensional mediation model as follows,

$$\lambda_i(t) = \lambda_0(t) \exp\left\{\gamma^* X_i + \beta_1 M_{1i} + \cdots \beta_p M_{pi} + \widetilde{\varphi}\pi_i\right\}, \quad (3)$$

$$M_{ki} = c_k + \alpha_k X_i + e_{ki} + \tilde{\phi}_k \pi_i,\ k = 1, 2, \cdots, p, \quad (4)$$

where π_i is the covariate of calculated propensity score; $\widetilde{\varphi}$ is the effect of the covariate on the outcome; $\tilde{\phi}_k$ is the effect of the covariate on the mediator. We will compare this with the method of adjusting all confounders as covariates and the method of ignoring confounders.

The goal of variable selection is to identify $S = \left\{k: \widehat{\alpha}_k \widehat{\beta}_k \neq 0\right\}$, which are the significant mediators between the exposure and the outcome when the number of potential mediators p is much larger than the sample size n, and the traditional statistics methods for Cox regression analysis fail to work (Luo et al., 2020). Besides, there are confounders influence the relationship of exposure, mediators, and outcome. To solve this problem, we propose the following procedure for high-dimensional mediation analysis with confounders in survival models. The overall workflow is as follows (**Figure 2**):

Step 0: We first construct the propensity score of confounders through a logistic regression model of exposure vs. baseline confounders, and use it as a covariate in the mediation models.

Step 1: For $k = 1, \cdots, p$, we select a subset $S_1 = \left\{k: 1 \leq k \leq p\right\}$ of size $d = \lceil 2n/\log(n) \rceil$ based on SIS method, where $\lceil \cdot \rceil$ is the ceiling function (Fan and Lv, 2008). For the mediators in S_1 are among the top d strongest P-values for the response variable. SIS procedure has been a general technique to reduce dimensionality from high to a small scale that is below the sample size. Here we use $d = \lceil 2n/\log(n) \rceil$ instead of $d = \lceil n/\log(n) \rceil$ to increase the probability for identifying important mediators, considering that both α_k and β_k have to be selected as nonzero to ensure a specific mediator to be selected.

Step 2: Among all the screened mediators M_k, $k \in S_1$ from Step 1, we further identify the subset $S_2 = \left\{k: \widehat{\beta}_k \neq 0\right\}$ *via* MCP-based Cox model. We obtain mediators M_k through the penalized log-partial likelihood optimization

$$\hat{\beta} = \text{argmax}_\beta \left\{ l_n(\beta) - \sum_{k=1}^{p} P_\lambda(\beta_k) \right\},\ k \in S_1,$$

where $l_n(\beta) = \sum_{i=1}^{n} \delta_i \left\{P_i^{\mathrm{T}} Q - \log[\sum_{l \in R_i} exp(P_l^T Q)]\right\}$ with the at-risk set $R_i = \left\{l: T_l \geq T_i\right\}$, $P_i = (X_i, \pi_i, M_{1i}, \cdots, M_{ki}, \cdots)^T$, and $Q = (\gamma^*, \widetilde{\varphi}, \beta_1, \cdots, \beta_k, \cdots)^{\mathrm{T}}$; $P'_\lambda(\beta_k) = \frac{(a\lambda - |\beta_k|)_+}{a\lambda}$ with shape parameter $a > 1$. Breheny and Huang (2011) implemented the MCP procedure with the R package *ncvreg*.

Step 3: For $k \in S_2$, a variable M_k is considered as a mediator between the exposure and outcome only if the indirect effect is significant. Here, we considered two methods to test the mediation effects, including the Sobel test (i.e., product method; Sobel, 1982) and the joint significant test.

Followed with the Sobel test for indirect effect, we have the P-value for testing the null hypothesis $H_0: \alpha_k \beta_k = 0$ of no indirect effect

$$P_{raw,\,k} = 2\left\{1 - \phi\left(\frac{|\widehat{\alpha}_k \widehat{\beta}_k|}{\widehat{\sigma}_{\alpha_k \beta_k}}\right)\right\},$$

where $\widehat{\sigma}_{\alpha_k \beta_k}$ is the estimate of the Sobel standard error (SE) (Sobel, 1982); $\widehat{\alpha}_k$ is the ordinary least squares estimator for α_k;

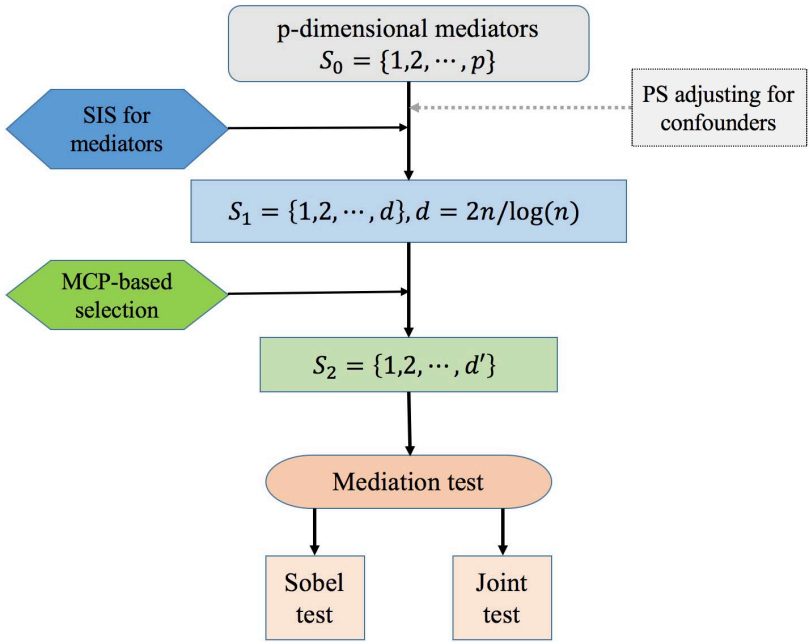


FIGURE 2 | Overall workflow for high-dimensional mediation analysis. The workflow includes the main processes: (0) adjusting for confounders based on the propensity score method; (1) using SIS technique for preliminary screening; (2) conducting MCP-based variable selection; (3) testing for mediation effects. (0–3) is correspond to Step0–Step3 in the methodology.

$\widehat{\beta}_k$ is the estimate of β_k, by refitting regression Eq. (3) with the mediators obtained in step 2.

The joint significant test for indirect effect is based on the path-specific (i.e., $X \to M$ and $M \to Y$) P-values (MacKinnon et al., 2002) and does not provide an estimate. The P-value for testing $H_0: \alpha_k = 0$ is given as

$$P_{raw,\,\alpha_k} = 2\left\{1 - \phi\left(\frac{|\widehat{\alpha}_k|}{\widehat{\sigma}_{\alpha_k}}\right)\right\},$$

and the P-value for testing $H_0: \beta_k = 0$ is

$$P_{raw,\,\beta_k} = 2\left\{1 - \phi\left(\frac{|\widehat{\beta}_k|}{\widehat{\sigma}_{\beta_k}}\right)\right\}.$$

Thus, the P-value for the joint significance test is defined as

$$P_{raw,\,k} = \max\left(P_{raw,\,\alpha_k},\ P_{raw,\,\beta_k}\right).$$

We have the revised P-value *via* the Bonferroni's method in order to adjust for multiple comparisons

$$P_k = \min\left\{P_{raw,k} \cdot |S_2|,\ 1\right\},$$

where $|S_2|$ is the number of elements in set S_2. Hence, we can reject the null hypothesis of no IE_k if $P_k < 0.05$, and conclude that the variable M_k is the significant mediator between the exposure and outcome.

Remark 1: Luo et al. (2020) proposed a compositional mediation framework to identify biomarkers which mediate the influence of smoking on lung cancer survival with high-dimensional candidates. They used a regression-based approach, which relies on the assumptions that there are no confounders that influence the relations between exposure and mediator, and exposure and outcome. This assumption holds if subjects are randomly assigned to levels of exposure, but generally random assignment is not possible in observational studies. We propose the use of propensity scores to adjust for confounders in high-dimensional mediation analysis in survival models.

Remark 2: Our method has three advantages. First, different from Luo et al. (2020), our approach is a simultaneous inference for high-dimensional mediation analysis with multiple confounders in survival models implemented with a propensity score method. The propensity score method can help remove the selection bias that may result when subjects are not randomly assigned to levels of exposure in observational studies. Second, compared with regression adjustment approach that include the confounders in mediation model directly, our method is more concise since we focus on incorporating the logit propensity score as a covariate in the mediation analysis. An advantage of propensity scores is that they reduce multiple potential confounders into a single numerical summary. Third, our method has a substantial improvement over method that does not include propensity scores.

SIMULATION STUDIES

In this section, we evaluate the performance of the proposed mediator selection and mediation effect estimation method through simulation studies. In order to investigate how the sample size impacts the performance, three sample size

levels ($N = 300$, $N = 500$, $N = 1,000$) are presented with potential mediators number $p = 10,000$. For each scenario, 500 replications of simulated data sets are conducted. Besides, we also consider two censoring rate settings of 15 and 30%.

For each subject i, $i = 1, 2, \cdots, N$:

1) we consider 10 confounders $Z = (Z_1, \cdots, Z_{10})^T$ affecting the relationship of X, M, and Y, where $Z_1, \cdots, Z_5$ are independently generated from the Bernoulli distribution with $Pr(Z_m = 1) = 0.3$, $m = 1, 2, \cdots, 5$ and $Z_6, \cdots, Z_{10}$ are generated from the multivariate normal distribution $N(0,)$ with a covariance matrix $\Sigma = (\sigma_{ij})_{5\times5}$, $\sigma_{ii} = 1$, $i = 1, \cdots, 5$ and $\sigma_{ij} = 0.3$, $i \neq j$;
2) we generate exposure X as a Bernoulli distributed variable $X_i \sim Bernoulli(P)$, where $P = 1/\left[1e^{-(\theta^T Z)}\right]$, and $\theta = (\theta_1, \cdots, \theta_{10})^T = (0.2, 0.3, 0.3, 0.5, 0.6, 0.2, 0.3, 0.3, 0.5, 0.6)^T$ denote for the coefficients of confounders Z on X. The 10 confounders have varying influences on the exposure. For example, the coefficients of Z_1 and Z_6 are much smaller than Z_5 and Z_{10};
3) we generate the mediator $M_{ki} = c_k + \alpha_k X_i + \phi_{k1} Z_1 + \cdots + \phi_{k10} Z_{10} + e_{ki}, k = 1, 2, \cdots, p$, where c_k is generated from the uniform distribution $U(0, 1)$; $(\alpha_1, \cdots, \alpha_8)^T = (0.5, 0.6, 0.5, 0.6, 0.5, 0.5, 0, 0)^T$ and the rest elements of α equals zero; $\phi_k = (\phi_{k1}, \cdots, \phi_{k10})^T = (0.3, 0, 0.4, 0.2, 0.5, 0, 0.4, 0.2, 0.5, 0.3)^T$ denote the effects of Z on mediator M_k; e_k is generated from the standard normal distribution $N(0, 1)$; the correlation between mediators basically falls between 0.5 and 0.6 which is close to the real data;
4) the death time D_i is generated as exponential distribution with the hazard function $\lambda_i(t \mid X_i, M_i) = \lambda_0(t) \exp\{\gamma X_i + \beta_1 M_{1i} + \cdots + \beta_p M_{pi} + \varphi_1 Z_1 + \cdots + \varphi_{10} Z_{10}\}$, where $\lambda_0(t)$ equals 0.5; γ equals 0.5; the first eight elements of β be $(\beta_1, \cdots, \beta_8)^T = (0.6, 0.6, 0.5, 0.5, 0, 0, 0.5, 0.5)^T$ and the rest elements of β equals zero; $\varphi = (\varphi_1, \cdots, \varphi_{10})^T = (0, 0.2, 0.2, 0.3, 0.2, 0.3, 0, 0.2, 0.3, 0.2)^T$ denote the effects of Z on Y;
5) the censoring time is generated through $C_i \sim U(0, c_0)$ with constant c_0 chosen so that we can control the percentage of censored subjects.

To summarize, only the first four mediators have significant mediation effects, which satisfy the condition of $\alpha_k \beta_k \neq 0$. In this part, we conduct a comparison of our proposed method with the other two approaches, including models ignoring confounders (Naïve approach) and models adjusting all 10 confounders as covariates (Z approach). We use the proposed procedure to identify significant mediators and estimate mediation effects, where the proposed approach uses the logit propensity score estimated through logistic regression as the covariate to adjust for confounding effects. Through the simulation studies, we want to demonstrate that propensity score methods can be used to adjust for confounding in the high-dimensional mediation selection and estimation.

TABLE 1 | Accuracy of mediator selection (p = 10,000, with 500 replications).

Cen = 15%	Methods	Sobel test			Joint test		
		TPR	FP	FDP	TPR	FP	FDP
N = 300	PS	0.6025	0	0	0.6890	0.0100	0.0025
	Naïve	0.6580	4.4780	0.4982	0.6580	5.0860	0.5721
	Z	0.6670	0	0	0.7800	0.0240	0.0060
N = 500	PS	0.9610	0.0020	0.0004	0.9690	0.0040	0.0008
	Naïve	0.9425	3.6400	0.4745	0.9425	4.3420	0.5168
	Z	0.9565	0.0020	0.0004	0.9695	0.0260	0.0056
N = 1,000	PS	1	0.0100	0.0020	1	0.0100	0.0020
	Naïve	0.9995	3.3420	0.4401	0.9995	3.6460	0.4593
	Z	1	0.0100	0.0020	1	0.0280	0.0056
Cen = 30%	**Methods**	**TPR**	**FP**	**FDP**	**TPR**	**FP**	**FDP**
N = 300	PS	0.5370	0.0020	0.0005	0.6565	0.0080	0.0021
	Naïve	0.6370	3.6060	0.5469	0.6370	5.3460	0.6474
	Z	0.5825	0	0	0.7450	0.0220	0.0058
N = 500	PS	0.9505	0.0020	0.0004	0.9650	0.0080	0.0017
	Naïve	0.9235	3.5900	0.4778	0.9235	4.3940	0.5285
	Z	0.9460	0	0	0.9695	0.0340	0.0069
N = 1,000	PS	1	0.0080	0.0016	1	0.0100	0.0020
	Naïve	0.9995	3.6160	0.4581	0.9995	3.9020	0.4756
	Z	1	0.0040	0.0008	1	0.0260	0.0052

PS, method of using the propensity score as the covariate; Naïve, method of ignoring confounders; Z approach, method of using confounders as covariates.
TPR, true positive rates; FP, false positive; FDP, false discovery proportion (=V/R, where V is the number of false discoveries and R is the number of total discoveries); all the three measures are the average value over 500 times.

TABLE 2 | Estimation of log hazard mediation effects: $\alpha_k\beta_k$ (Cen = 15%).

Cen = 15%	***N* = 300**			***N* = 500**			***N* = 1,000**		
(α_k, β_k)	**PS**	**Naïve**	**Z**	**PS**	**Naïve**	**Z**	**PS**	**Naïve**	**Z**
(0.5, 0.6) = 0.30 (MSE)	0.2895 (0.0088)	0.7712 (0.2484)	0.2925 (0.0100)	0.2960 (0.0049)	0.8473 (0.3156)	0.3117 (0.0062)	0.3077 (0.0026)	0.8545 (0.3145)	0.3151 (0.0026)
(0.6, 0.6) = 0.36 (MSE)	0.3501 (0.0096)	0.8333 (0.2547)	0.3574 (0.0127)	0.3559 (0.0058)	0.8932 (0.3024)	0.3765 (0.0075)	0.3682 (0.0030)	0.9111 (0.3122)	0.3728 (0.0031)
(0.5, 0.5) = 0.25 (MSE)	0.2426 (0.0061)	0.6614 (0.1901)	0.2680 (0.0082)	0.2499 (0.0037)	0.7032 (0.2192)	0.2626 (0.0048)	0.2576 (0.0019)	0.7075 (0.2161)	0.2592 (0.0018)
(0.6, 0.5) = 0.30 (MSE)	0.2907 (0.0073)	0.7034 (0.1885)	0.3146 (0.0102)	0.3045 (0.0044)	0.7513 (0.2191)	0.3228 (0.0060)	0.3120 (0.0021)	0.7564 (0.2149)	0.3130 (0.0021)
(0.5, 0) = 0 (MSE)	– (–)	0.3002 (0.0902)	– (–)	– (–)	– (–)	– (–)	– (–)	0.1246 (0.0155)	– (–)
(0.5, 0) = 0 (MSE)	– (–)	– (–)	– (–)	0.0724 (0.0052)	0.1789 (0.0320)	– (–)	– (–)	0.1588 (0.0253)	0.0688 (0.0047)
(0, 0.5) = 0 (MSE)	0.0037 (0.0048)	0.4406 (0.2075)	0.0201 (0.0061)	0.0040 (0.0026)	0.4727 (0.2305)	0.0079 (0.0028)	0.0031 (0.0015)	0.4769 (0.2309)	0.0031 (0.0014)
(0, 0.5) = 0 (MSE)	0.0026 (0.0051)	0.4479 (0.2143)	0.0102 (0.0059)	0.0022 (0.0029)	0.4711 (0.2289)	0.0006 (0.0032)	0.0017 (0.0016)	0.4746 (0.2293)	0.0020 (0.0016)
(0, 0) = 0 (MSE)	– (–)	– (–)	– (–)	– (–)	0.1624 (0.0263)	0.0247 (0.0006)	– (–)	0.0882 (0.0078)	– (–)
(0, 0) = 0 (MSE)	– (–)	– (–)	– (–)	– (–)	– (–)	0.0135 (0.0002)	– (–)	– (–)	– (–)

PS, method of using the propensity score as the covariate; Naïve, method of ignoring confounders; Z approach, method of using confounders as covariates.
MSE, mean square error; –, means the not available value.

TABLE 3 | Estimation of log hazard mediation effects: $\alpha_k\beta_k$ (Cen = 30%).

Cen = 30%	***N* = 300**			***N* = 500**			***N* = 1,000**		
(α_k, β_k)	**PS**	**Naïve**	**Z**	**PS**	**Naïve**	**Z**	**PS**	**Naïve**	**Z**
(0.5, 0.6) = 0.30 (MSE)	0.2793 (0.0096)	0.7589 (0.2415)	0.2982 (0.0108)	0.2932 (0.0053)	0.8362 (0.3085)	0.3139 (0.0070)	0.3081 (0.0028)	0.8597 (0.3214)	0.3197 (0.0034)
(0.6, 0.6) = 0.36 (MSE)	0.3431 (0.0113)	0.8289 (0.2546)	0.3666 (0.0146)	0.3528 (0.0059)	0.8851 (0.2978)	0.3819 (0.0084)	0.3689 (0.0032)	0.9179 (0.3213)	0.3839 (0.0040)
(0.5, 0.5) = 0.25 (MSE)	0.2377 (0.0069)	0.6745 (0.2054)	0.2697 (0.0098)	0.2480 (0.0040)	0.6983 (0.2176)	0.2649 (0.0057)	0.2571 (0.0019)	0.7099 (0.2199)	0.2618 (0.0022)
(0.6, 0.5) = 0.30 (MSE)	0.2803 (0.0084)	0.7082 (0.2021)	0.3241 (0.0130)	0.3018 (0.0046)	0.7449 (0.2162)	0.3226 (0.0067)	0.3121 (0.0023)	0.7626 (0.2225)	0.3184 (0.0029)
(0.5, 0) = 0 (MSE)	– (–)	– (–)	– (–)	– (–)	– (–)	0.0818 (0.0067)	– (–)	– (–)	0.0390 (0.0015)
(0.5, 0) = 0 (MSE)	– (–)	– (–)	– (–)	– (–)	– (–)	0.0925 (0.0085)	– (–)	0.1301 (0.0169)	0.0730 (0.0053)
(0, 0.5) = 0 (MSE)	0.0056 (0.0046)	0.4388 (0.2087)	0.0080 (0.0061)	0.0043 (0.0026)	0.4657 (0.2259)	0.0043 (0.0029)	0.0034 (0.0015)	0.4798 (0.2346)	0.0034 (0.0015)
(0, 0.5) = 0 (MSE)	0.0016 (0.0051)	0.4394 (0.2084)	0.0031 (0.0061)	0.0015 (0.0029)	0.4631 (0.2232)	0.0021 (0.0033)	0.0018 (0.0016)	0.4743 (0.2297)	0.0021 (0.0016)
(0, 0) = 0 (MSE)	– (–)	– (–)	0.0391 (0.0015)	– (–)	– (–)	– (–)	– (–)	0.0977 (0.0095)	0.0062 (0.0001)
(0, 0) = 0 (MSE)	– (–)	– (–)	0.0147 (0.0002)	– (–)	– (–)	– (–)	– (–)	– (–)	0.0008 (0.0000)

PS, method of using the propensity score as the covariate; Naïve, method of ignoring confounders; Z approach, method of using confounders as covariates.
MSE, mean square error; –, means the not available value.

Simulation results are presented in **Tables 1–3**. **Table 1** evaluates the performance of mediator selection of the proposed approach in comparison to the other two approaches using the true positive rate (TPR), the number of false positive (FP), and false discovery proportion (FDP) of selection after the significance test for mediation effects based on the joint and the Sobel methods. The TPR of the proposed propensity score approach is lower than the Z approach when the sample size is 300, but performs similarly to the Z approach as the sample size increases. And the proposed method has lower FP and FDP rates than the Z approach. The Naïve approach has lower TPR and higher FP and FDP rates, indicating the deficiency in identifying significant mediators due to confounding effects. Take sample size 500 as an example, the FP and FDP rates based on the joint test are 0.004 and 0.0008 for the proposed approach; 0.026 and 0.0056 for the Z approach; and 4.342 and 0.5168 for the Naïve approach. Selection results based on the joint test are similar. Besides, as the censoring rate increases, the TPR rates decrease, especially for the lower sample size. Similar results can be seen for the setting with a 30% censoring rate.

Tables 2, **3** show the estimation of mediation effects with censoring rate by 15 and 30%, respectively. The bias of the indirect effect estimator using the PS approach is very small. The Naïve approach is biased severely. It is important to note that the proposed method even has slightly better performance than the Z approach including all confounders as covariates in the estimation of indirect effects.

In summary, the results demonstrate that the bias of the mediation effect estimator of our proposed methods for high-dimensional mediation analysis using the calculated propensity score to adjust confounding influence is nearly unbiased. Besides, with the increase of sample size, the ability in mediator selection including TPR, the number of FP, and FDP shows good performance as well as the Z approach. The Naïve approach ignoring the confounders produces a severe bias in both mediator selection and mediation effects estimation. Compared with the classical regression method for mediation analysis with confounders, the procedure we proposed is more concise and efficient.

REAL DATA ANALYSIS

As we know, smoking is an important risk factor for lung cancer, one of the deadliest cancer worldwide (Herbst et al., 2008). With the development of sequencing technology, both Illumina Infinium HumanMethylation27 and HumanMethylation450 are widely used platforms that allow measuring high-dimensional DNA methylation levels of roughly 27 and 450 k respectively(Bibikova et al., 2011). As the individual level phenotype and genotype data are available, researchers have indicated that methylation markers are acting as mediators between smoking and lung function or lung cancer patient's overall survival (Zhang et al., 2016; Luo et al., 2020). The TCGA (The Cancer Genome Atlas) lung cancer cohort study had been used for mediation analysis to identify the methylation markers (Luo et al., 2020). However, the assumption that samples are randomly assigned to the smoking or non-smoking group is violated. Hence, it is of great importance to adjust for confounding effects when conducting high-dimensional mediation analysis.

We apply the proposed method using the calculated propensity score as a covariate in high-dimensional mediation analysis with survival outcome to a lung cancer dataset including lung squamous cell carcinoma and lung adenocarcinoma. There are 1,299 lung cancer patients aged 33–90 years and 907 of them had DNA methylation profile measured using the Illumina Infinium HumanMethylation 450 platform. DNA methylation values were recorded for each array probe in each sample *via* BeadStudio software. A total of 365,307 probes were included in the analysis.

To identify the potential methylation mediators between the tobacco smoking and the overall survival, we apply the high-dimensional mediator model with smoking status assessed at their initial diagnosis (smoker/non-smoker) as the exposure variable, DNA methylation measured concurrently as the high-dimensional mediators, and the survival time as the outcome variable. The overall survival time is defined as the number of days from the initial diagnosis to the death or the last follow-up date. Subjects with no observed time, exposure, and other covariates are excluded; there are 696 patients with 269 deaths left. Covariates including age at initial diagnosis, gender, and radiotherapy (yes/no) are considered.

We first adjust for the baseline confounders including age, gender, and radiotherapy using the calculated propensity score. Due to the fact that the relationships between methylation and the outcome are much stronger than those between exposure and methylation in the analysis data set, we add top $d = 2n/\log(n)$ CpGs using SIS method based on the path from smoking to the methylation in order to improve the probability to recognize significant mediators. Then, we run a variable selection on the CpGs screened in the above step. Finally, we carry out the significance test for the mediation effects.

The analysis results are presented in **Table 4**. We identify CpGs mediating the relationship between smoking and the overall survival of lung cancer patients with Bonferroni's adjusted $P < 0.05$. Since smoking generally increases the risk of progression to death and reduces the overall survival of lung cancer patients with the total effect of 1.3436 (95% CI: 1.0377–1.7400), we focus on the mediators with the log-hazard indirect effect $\alpha\beta = 0$ (smoking increases the mortality). Our method finds two CpGs (cg21926276 and cg20707991) mediating the relationship of smoking and risk of progression to death, while methods including all confounders as covariates and methods ignoring confounders only find cg20707991 to be a significant mediator. The methylation site cg21926276 has been reported as a mediator of smoking and the risk of progression to death (Luo et al., 2020). All the two genes in which methylation sites locate are associated with lung cancer or tumor growth in previous studies. For example, the gene H19 (cg21926276 locate) is related to both lung cancer and tumor growth, methylation of which has been thought of as a sensitive marker of tobacco history (Bouwland-Both et al., 2015; Matouk et al., 2015). The gene PTPRN2 (cg20707991 locate) is also associated with lung cancer

TABLE 4 | Summary of selected CpGs with estimators ($\widehat{\alpha}\widehat{\beta} > 0$) and *P*-values for significant mediators.

Methods	CpGs	Chromosome	Gene	$\widehat{\alpha}$	$\widehat{\beta}$	*P*(Sobel)	*P*(Joint)
Proposed	cg21926276	chr11	H19	–0.06	–3.21	6.69e–03	1.75e-04
	cg20707991	chr7	PTPRN2	–0.06	–2.12	5.36e–02	1.28e–02
Z	cg20707991	chr7	PTPRN2	–0.06	–2.40	2.49e–03	4.82e–05
Naïve	cg20707991	chr7	PTPRN2	–0.06	–1.01	2.58e–02	1.61e–02

and survival of cancer patients (Anglim et al., 2008; Wielscher et al., 2015). Besides confirming the previously reported genes, cg20707991 is identified as a novel marker for the survival of lung cancer patients.

The CpGs are the DNA methylation sites. Chromosomes and Genes are where the CpGs locate. $\widehat{\alpha}$ is the estimation of the effect of exposure on methylation. $\widehat{\beta}$ is the estimation of the effect of methylation on the risk of progression to death. *P*(Sobel) is the Sobel test *P*-values and *P*(Joint) is the joint test *P*-values, which are corrected by Bonferroni's method.

Based on the above analysis, compared with non-smokers, the risk of death for those smokers is 1.3436 (95% CI: 1.0377–1.7400). Mediation analysis using Cox proportional hazards model discovers that the effect of having serious smoking history on the increased risk of progression to death is mediated through methylation markers including cg21926276 and cg20707991; the hazard ratio for each mediator is 1.2093 (95% CI: 1.2019–1.2167) and 1.1388 (95% CI: 1.1339–1.1438), respectively. Interventions can be explored on these markers to improve medical care for the detection and treatment of lung cancer among smokers.

To sum up, through the mediation analysis of smoking, DNA methylation, and the survival time of the lung cancer patients, we found two CpGs mediating the smoking and the mortality. Our findings not only were in line with previous studies which found that the gene that CpGs locate were important biomarkers for lung cancer, but also uncovered the mediation role of the markers connecting the smoking exposure and the survival time.

DISCUSSION

The motivation of this study is that the assumption of no confounders affecting the relationship of exposure, mediators and outcome in the classical mediation model is difficult to be satisfied with observational studies. Hence, how to adjust these confounders is an important and practical question. The propensity score method can summarize a large scale of confounders into a single value which is more concise than the methods with a regression adjustment for all the potential confounders. Thus, motivated by the above facts, we develop a new method that using the propensity score as a covariate to adjust for confounding effects in high-dimensional mediation models.

In this article, we focus on how to adjust for confounding influences when the exposure is not randomly assigned in observational studies. We propose a new model for high-dimensional mediation analysis using propensity score methods to adjust for confounding effects. To identify the significant mediators from high-dimensional potential candidate variables, we mainly combine the sure screening technique, MCP-based penalty, and Sobel and joint methods for significance tests. We evaluate the performance of the proposed procedure *via* several simulation studies and a real data application.

Compared with the mediation analysis which includes all the confounders as covariates, our proposed approach for high-dimensional mediation analysis using the calculated propensity score to adjust confounding influence would be an improvement in mediator selection and indirect effect estimation. The simulation results also show that the proposed method can obtain a nearly unbiased estimation for indirect effects. It is also interesting to note that if confounders are omitted from the model, then the estimates for mediation effects will be severely biased. In conclusion, we suggest using the calculated propensity score to adjust for confounders among the exposure, mediators, and the outcome when evaluating mediation.

As mentioned previously, propensity score methods have many other applications, such as matching, weighting, and sub-classification. It is of interest to explore the performance of high-dimensional mediation selection and estimators using propensity score weighting. Also, the propensity score in our current approach is only valid for single exposure. Analysis approach for the high-dimensional mediators with more than two exposure status is still to be developed. The present simulation results do not address the cases that confounders only affect mediators and the outcome. It is of future interest to developed methods involving estimating propensity score for high-dimensional mediators. Of note, the Sobel test and the joint significant test we used are conservative, which paves the way for developing a more powerful test method, such as the Divide-Aggregate Composite null Test (DACT; Liu et al., 2021). The DACT method is especially useful for the composite null hypothesis of no mediation effect in large-scale genome-wide epigenetic studies. It is desirable to consider such a powerful test method for mediation effects in the future research.

DATA AVAILABILITY STATEMENT

The TCGA (The Cancer Genome Atlas) lung cancer data we used in our real data analysis can be found in (https://xenabrowser.net/) without limitation. Our procedure is implemented using the R tool. The corresponding R code can be found at https://github.com/luo-chengwen/HIMAsurvival-PS.

AUTHOR CONTRIBUTIONS

CL and ZY implemented the method, drafted the manuscript, conceived the idea, designed the study, and implemented the code. YC, TW, and YM were involved in the data analysis. All authors read and approved the final manuscript.

FUNDING

The study was supported by the following fundings: Three-year Plan of Shanghai Public Health System Construction (fundingID: GWV-10.1-XK05) and Shanghai Commission of Science and Technology (fundingID: 21ZR1436300).

REFERENCES

Albert, J. M. (2008). Mediation analysisvia potential outcomes models. *Stat. Med.* 27, 1282–1304. doi: 10.1002/sim.3016

Albert, J. M., and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics* 67, 1028–1038. doi: 10.1111/j.1541-0420.2010.01547.x

Anglim, P., Galler, J., Koss, M., Hagen, J. A., Turla, S., Campan, M., et al. (2008). Identification of a panel of sensitive and specific DNA methylation markers for squamous cell lung cancer. *Mol. Cancer* 7:62. doi: 10.1186/1476-4598-7-62

Baron, R. M., and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and stastical consideration. *J. Pers. Soc. Psychol.* 51, 1173–1182.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* 98, 288–295. doi: 10.1016/j.ygeno.2011.07.007

Biesanz, J. C., Falk, C. F., and Savalei, V. (2010). Assessing mediational models: testing and interval estimation for indirect effects. *Multivariate Behav. Res.* 45, 661–701. doi: 10.1080/00273171.2010.498292

Bouwland-Both, M., Van Mil, N., Tolhoek, C., Stolk, L., Eilers, P., Verbiest, M., et al. (2015). Prenatal parental tobacco smoking, gene specific DNA methylation, and newborns size: the generation R study. *Clin. Epigenetics* 7:83.

Breheny, P., and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* 5, 232–253.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* 19, 121–136. doi: 10.1093/biostatistics/kxx027

Coffman, D. L. (2011). Estimating causal effects in mediation analysis using propensity scores. *Struct. Equ. Modeling* 18, 357–369. doi: 10.1080/10705511.2011.582001

Cox, D. (1972). Regression models and life tables. *J. R. Stat. Soc.* 34, 187–220.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* 71, 1–14. doi: 10.1111/biom.12248

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc.* 70, 849–911. doi: 10.1111/j.1467-9868.2008.00674.x

Herbst, R. S., Heymach, J. V., and Lippman, S. M. (2008). Lung cancer. *N. Engl. J. Med.* 359, 1367–1380.

Huan, T., Joehanes, R., Schurmann, C., Schramm, K., Pilling, L. C., Peters, M. J., et al. (2016). A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. *Hum. Mol. Genet.* 25:ddw288. doi: 10.1093/hmg/ddw288

Huang, Y.-T., and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* 72, 402–413. doi: 10.1111/biom.12421

Huang, Y. T., and Yang, H. I. (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology* 28, 370–378. doi: 10.1097/ede.0000000000000651

Lange, T., and Hansen, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology* 22, 575–581. doi: 10.1097/ede.0b013e31821c680c

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A., and Lin, X. (2021). Large-scale hypothesis testing for causal mediation effects with applications in genomewide epigenetic studies. *J. Am. Stat. Assoc.* doi: 10.1080/01621459.2021.1914634

Luo, C., Fa, B., Yan, Y., Wang, Y., Zhou, Y., Zhang, Y., et al. (2020). High-dimensional mediation analysis in survival models. *PLoS Comput. Biol.* 16:e1007768. doi: 10.1371/journal.pcbi.1007768

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol.* 58, 593–614.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* 7, 83–104. doi: 10.1037/1082-989x.7.1.83

Matouk, I. J., Halle, D., Gilon, M., and Hochberg, A. (2015). The non-coding RNAs of the H19-IGF2 imprinted loci: a focus on biological roles and therapeutic potential in lung cancer. *J. Transl. Med.* 13:113.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychol. Methods* 19, 459–481. doi: 10.1037/a0036434

Preacher, K. J., and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* 40, 879–891. doi: 10.3758/brm.40.3.879

Robins, J. M., Andrea, R., and Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* 90, 106–121. doi: 10.1080/01621459.1995.10476493

Robins, J. M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3, 143–155. doi: 10.1097/00001648-199203000-00013

Rosenbaum, P., and Rubin, D. (1982). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc.* 45, 212–218. doi: 10.1111/j.2517-6161.1983.tb01242.x

Rosenbaum, P., and Rubin, D. (1983). The central role of the propensity scores in observational studies for causal effects. *Biometrika* 70, 41–55. doi: 10.1093/biomet/70.1.41

Rosenbaum, P., and Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79, 516–524. doi: 10.1080/01621459.1984.10478078

Rosenbaum, P., and Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* 39, 33–38. doi: 10.1080/00031305.1985.10479383

Schafer, J. L., and Joseph, K. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol. Methods* 13, 279–313. doi: 10.1037/a0014268

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.* 13, 290–312. doi: 10.2307/270723

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *J. Educ. Behav. Stat.* 33, 230–251. doi: 10.3102/1076998607307239

Sohn, M. B., and Li, H. (2019). Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* 13, 661–681.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics* 63, 926–934. doi: 10.1111/j.1541-0420.2007.00766.x

Valente, M. J., Pelham, W. E., Smyth, H., and Mackinnon, D. P. (2017). Confounding in statistical mediation analysis: what it is and how to address it. *J. Couns. Psychol.* 64, 659–671. doi: 10.1037/cou0000242

van Kesteren, E.-J., and Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Struct. Equ. Modeling* 26, 710–723. doi: 10.1080/10705511.2019.1588124

VanderWeele, T. (2010). The use of propensity score methods in psychiatric research. *Int. J. Methods Psychiatr. Res.* 15, 95–103. doi: 10.1002/mpr.183

VanderWeele, T., and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiol. Methods* 2, 95–115.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* 20, 18–26. doi: 10.1097/ede.0b013e31818f69ce

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology* 22, 582–585. doi: 10.1097/ede.0b013e31821db37e

Wang, L., and Zhang, Z. (2011). Estimating and testing mediation effects with censored data. *Struct. Equ. Modeling* 18, 18–34. doi: 10.1080/10705511.2011.534324

Wielscher, M., Vierlinger, K., Kegler, U., Ziesche, R., Gsur, A., and Weinhäusel, A. (2015). Diagnostic performance of plasma DNA methylation profiles in lung cancer, pulmonary fibrosis and COPD. *EBioMedicine* 2, 929–936. doi: 10.1016/j.ebiom.2015.06.025

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 38, 894–942.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* 32, 3150–3154. doi: 10.1093/bioinformatics/btw351

Zhang, Z., and Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika* 78, 154–184. doi: 10.1007/s11336-012-9301-5

Zhao, Y., Lindquist, M., and Caffo, B. (2020). Sparse principal component based high-dimensional mediation analysis. *Comput. Stat. Data Anal.* 142:106835. doi: 10.1016/j.csda.2019.106835

Zhao, Y., and Luo, X. (2016). Pathway lasso: estimate and select sparse mediation pathways with high-dimensional mediators. *arXiv* [Preprint]. Available online at: https://ui.adsabs.harvard.edu/abs/2016arXiv160307749Z (accessed March 01, 2016).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fgene.2021.656826

Meta-Analyzing Multiple Omics Data With Robust Variable Selection

Zongliang Hu[1], *Yan Zhou*[1]* *and Tiejun Tong*[2]*

[1] College of Mathematics and Statistics, Shenzhen University, Shenzhen, China, [2] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

High-throughput omics data are becoming more and more popular in various areas of science. Given that many publicly available datasets address the same questions, researchers have applied meta-analysis to synthesize multiple datasets to achieve more reliable results for model estimation and prediction. Due to the high dimensionality of omics data, it is also desirable to incorporate variable selection into meta-analysis. Existing meta-analyzing variable selection methods are often sensitive to the presence of outliers, and may lead to missed detections of relevant covariates, especially for lasso-type penalties. In this paper, we develop a robust variable selection algorithm for meta-analyzing high-dimensional datasets based on logistic regression. We first search an outlier-free subset from each dataset by borrowing information across the datasets with repeatedly use of the least trimmed squared estimates for the logistic model and together with a hierarchical bi-level variable selection technique. We then refine a reweighting step to further improve the efficiency after obtaining a reliable non-outlier subset. Simulation studies and real data analysis show that our new method can provide more reliable results than the existing meta-analysis methods in the presence of outliers.

Keywords: heterogeneity, logistic regression, meta-analysis, robust estimation, variable selection

OPEN ACCESS

Edited by:
Jiebiao Wang,
University of Pittsburgh, United States

Reviewed by:
Cen Wu,
Kansas State University, United States
Duo Jiang,
Oregon State University, United States

***Correspondence:**
Yan Zhou
zhouy1016@szu.edu.cn
Tiejun Tong
tongt@hkbu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 January 2021*
Accepted: *24 May 2021*
Published: *05 July 2021*

Citation:
Hu Z, Zhou Y and Tong T (2021) Meta-Analyzing Multiple Omics Data With Robust Variable Selection. Front. Genet. 12:656826. doi: 10.3389/fgene.2021.656826

1. INTRODUCTION

With the advances in biological sciences, omics data have been playing an important role in many different fields of research. A typical example of such data includes gene expression data targeting for the identification of important genes that are related to disease status or clinical outcomes (Zhao et al., 2015). Nevertheless, as biological experiments are often measured with a relatively small number of samples, many identified genes are in fact very sensitive to mild data perturbations and thus lack of reliability. From another perspective, since many publicly available datasets have addressed the same scientific problems, one may consider to integrate multiple sources of data to borrow information across the studies and so improve the model interpretation and boost the statistical power (Glass, 1976; Wu et al., 2019). As an example, the integration analysis of genomic data from multiple studies has discovered new loci that are related to diseases including childhood obesity, colorectal cancer, and Crohn's disease (Houlston et al., 2008).

Meta-analysis is an efficient tool for integrating the scientific results from multiple studies. The classical meta-analysis methods are mainly based on the summary statistics including the p-values (Li and Tseng, 2011; Zhang et al., 2020) and/or the effect sizes (Choi et al., 2003; Chang et al., 2013). Recently, He et al. (2016) proposed a sparse method for meta-analyzing high-dimensional regression coefficients, which is based solely on the estimates of coefficients from multiple studies. When raw data from multiple studies are available, as recommended by

Tang and Song (2016), a retreat to the classical meta-analysis methods is often necessary. Specifically, under such circumstances, it becomes possible to jointly assess the effect of selected covariates at the study and group levels, which can incorporate heterogeneous effects from different studies so as to outperform the classical meta-analysis with better estimation accuracy (George, 2019).

Due to the high-dimensionality of omics data, the number of genes is often larger than the sample size. Incorporation of variable selection into raw data analysis has been one hot topic in statistics. For example, Zhou and Zhu (2010) proposed a bi-level variable selection method for selecting important genes, which not only removes unimportant groups efficiently but also maintains the flexibility of selecting variables within the group. When the heterogeneity exists between multiple studies, however, the important genes may only be remarkable in some studies but not in others. In view of this, Li et al. (2014) further extended the bi-level variable selection to heterogeneous high-dimensional multiple datasets. They treated the coefficients of each covariate from all datasets as groups, and performed the simultaneously variable selection both on the group and within the group. For other existing variable selection methods including, for example, group Bridge, composite MCP, and group exponential lasso that can be extended to meta-analyzing multiple studies, one may refer to Zhao et al. (2015), Kim et al. (2017), and Rashid et al. (2020).

Despite the huge popularity of variable selection methods in meta-analysis, little attention has been paid to the extension of these methods to handle outliers in high-dimensional data (Chi and Scott, 2014). For biological data, it is not uncommon that the tissue samples are mislabeled or contaminated (Wu et al., 2019). Outliers may strongly influence the accuracy of parameter estimation and variable selection, and as shown in Alfons et al. (2013), even one single outlier has the potential to make the selected variables based on the lasso penalty completely unreliable. This motivates us to consider the robust alternatives, especially when integrating the multiple datasets collected from different platforms and laboratories. Needless to say, robust estimation has a long history under the classical paradigm where the sample size is large and the dimension is small, see, for example, Yohai (1987), Hadi and Simonoff (1993), and Bianco and Yohai (1996). In particular, Rousseeuw and Leroy (1987) proposed a least trimmed squares estimator (LTS), which was shown to have a high breakdown point and was further improved by the well-designed fast algorithm (Fast-LTS) in Rousseeuw and Driessen (2006). More recently, Alfons et al. (2013) and Yang et al. (2018) extended LTS to high-dimensional data with the alternating minimization algorithm. Ren et al. (2019) investigated a robust variable selection for continuous censored data, where the least absolute deviation loss was adopted to accommodate heavy-tailed data. For a review of recent developments on robust regression and variable selection methods, one may refer to Wu and Ma (2015) and Sun et al. (2020).

We note, however, that the aforementioned robust methods have all been focused on a single study. Moreover, most of the existing methods are based on robust loss functions that aim to deal with heavy-tailed continuous data; see, for example, the least absolute deviation and check loss functions (Wu and Ma, 2015). In recent public biological database (e.g., Gene Expression Omnibus database), many datasets are collected from case-control studies with binary phenotypes. Therefore, the commonly used robust loss functions may not be directly applicable to this scenario. In this paper, inspired by the idea of the LTS estimator and the bi-level lasso variable selection (Zhou and Zhu, 2010; Li et al., 2014), we propose a two-step procedure for the robust variable selection that can be applied to meta-analyzing multiple case–control studies. In the first step, we search a clean index subset for each study based on the Fast-LTS algorithm and the bi-level variable selection technique. In the second step, we further refine a reweighting rule to enhance the estimation efficiency and the accuracy of variable selection. The key idea in this step is to identify outliers according to the current model obtained in the first step and to assign a small or zero weight for outliers. Our new robust meta-analysis method has two main advantages: (1) the Fast-LTS algorithm guarantees the convergence of the selected clean subsets; (2) the bi-level variable selection not only identifies important covariates with the strength of multiple datasets, but also maintains the flexibility of variable selection between the datasets to account for the data heterogeneity. Consequently, in the presence of outliers, our proposed method can provide better parameter estimation and also identify more accurate informative covariates than the existing strategies, especially when the dimension is large.

The rest of this paper is organized as follows. In section 2, we describe the model setting and develop the new algorithm for our two-step robust meta-analysis method. The selection of tuning parameters involved in the algorithm is also discussed. In section 3, we conduct simulation studies to assess the performance of the our robust estimation in meta-analyzing multiple datasets. We further apply the new method to robustly analyze a real data example in section 4. Finally, we conclude the paper with some future work in section 5, and provide the technical results in the **Appendix**.

2. METHODS AND ALGORITHM

In this section, we first formulate the model in section 2.1, then propose a two-step robust meta-analysis method in section 2.2, and finally, we present the selection of tuning parameters in section 2.3.

2.1. Data and Models

Suppose there are M independent studies, and each study contains n_k subjects for $k = 1, \ldots, M$. Let also $\mathcal{D}_k = \{(\boldsymbol{x}_{ki}, y_{ki}), i = 1, \ldots, n_k\}$ be the raw data, where $y_{ki} \in \{0, 1\}$ is a binary response variable and $\boldsymbol{x}_{ki} = (x_{ki,1}, \ldots, x_{ki,p})^T \in \mathcal{R}^p$ is the covariate vector. Throughout this paper, we assume that the dimension p is common for all the studies. To link y_{ki} to $\boldsymbol{x}_{ki}$, we consider the logistic model with

$$\pi_{ki} = P(y_{ki} = 1|\boldsymbol{x}_{ki}) = \frac{\exp(\beta_{k0} + \boldsymbol{x}_{ki}^T \boldsymbol{\beta}_k)}{1 + \exp(\beta_{k0} + \boldsymbol{x}_{ki}^T \boldsymbol{\beta}_k)}, \tag{2.1}$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^T \in \mathcal{R}^p$ is the unknown coefficient vector for the kth study that captures the effect of each covariate. Since the intercept β_{k0} can be readily handled, without loss of generality, we will suppress it for convenience. To model the heterogeneity between the studies, we allow $\boldsymbol{\beta}_k$ to vary with k. For omics data, as mentioned earlier, the number of covariates p is often much larger than the sample size n, and meanwhile only a small proportion of covariates will be related to the response variable. We divide the covariates into two disjoint sets: the informative set $I_{k1} = \{j = 1, \ldots, p : \beta_{kj} \neq 0\}$ and the noninformative set $I_{k1} = \{j = 1, \ldots, p : \beta_{kj} = 0\}$ for $k = 1, \ldots, M$. Our main goals are to identify the informative sets and to estimate the coefficients of the informative covariates.

Note that each covariate has M coefficients across the studies. When the M datasets come from studies that focus on the same biological questions, the M coefficients may share some common information. This makes it possible to integrate information across multiple datasets and make simultaneous coefficient estimation and variable selection. On the other side, however, outliers and data contamination have been widely observed in the predictors and responses, and as a consequence, they will yield the lasso-type penalties largely unreliable.

2.2. Robust Meta-Analysis Method

In this section, we propose a new two-step procedure for robustly meta-analyzing multiple omics data.

2.2.1. Simultaneous Estimation

Let $H_k \subseteq \{1, 2, \ldots, n_k\}$ be a subset of the indexes from the kth study with the cardinality $|H_k| = h_k$ for $k = 1, \ldots, M$, and $\mathcal{H} = \{H_1, \ldots, H_M\}$ be a subset of the indexes for the M studies. Then by following Zhou and Zhu (2010) and Li et al. (2014), we define the objective function as

$$Q(\mathcal{H}, \boldsymbol{\beta}) = \sum_{k=1}^{M} \sum_{i \in H_k} d(\boldsymbol{x}_{ki}^T \boldsymbol{\beta}_k, y_{ki}) + \lambda \sum_{j=1}^{p} \left(\sum_{k=1}^{M} |\beta_{kj}| \right)^{1/2}, \quad (2.2)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_M^T)^T$ is the stack of the coefficient vectors, and

$$d(\boldsymbol{x}_{ki}^T \boldsymbol{\beta}, y_{ki}) = -y_{ki} \log \pi_{ki} - (1 - y_{ki}) \log(1 - \pi_{ki}) \quad (2.3)$$

is the deviance. When the set $\mathcal{H}$ is outlier-free, minimizing the objective function (2.2) gives the robust and sparse estimator for the coefficients as

$$\widehat{\boldsymbol{\beta}}_{\mathcal{H}} = (\widehat{\boldsymbol{\beta}}_1^T, \ldots, \widehat{\boldsymbol{\beta}}_M^T)^T = \arg\min_{\boldsymbol{\beta}} Q(\mathcal{H}, \boldsymbol{\beta}),$$

where $\widehat{\boldsymbol{\beta}}_k = (\hat{\beta}_{11}, \ldots, \hat{\beta}_{1p})^T$ is the estimate of the coefficient vector in the kth study.

Note that the square root penalty (or $L_{1/2}$ norm) in (2.2) treats $\beta_{1j}, \ldots, \beta_{Mj}$ as a group for each covariate j, and conducts a group-type variable selection. In addition, the L_1 norm used inside the square root penalty can perform a study-specific variable selection that shrinks the small coefficients to zero and keeps only the large coefficients (Tsybakov and Vande, 2005). Then, in total, the penalty term in (2.2) essentially plays a role for the bi-level variable selection; that is, it cannot only borrow common information across the studies, but also take into account the data heterogeneity and maintain the flexibility of parameter estimation between the studies. From this perspective, with the penalty term in (2.2), the optimization procedure actually borrows the strength across the M studies and is quite different from performing a separate variable selection in each individual study (Li et al., 2014).

In practice, to determine a set that can well approximate the outlier-free set $\mathcal{H}$, it will involve iteratively optimizing the objective function (2.2). Note also that the square root penalty in (2.2) is not a convex function and has a complex nonlinear form. To solve the problem, we first give a simpler and equivalent version for the optimization.

THEOREM 1. *Let $\beta_{kj} = \alpha_j \gamma_{kj}$ for $k = 1, \ldots, M$ and $j = 1, \ldots, p$. Let also $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kp})^T$. Consider the following objective function:*

$$Q_1(\mathcal{H}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{k=1}^{M} \sum_{i \in H_k} d(\boldsymbol{x}_{ki} \boldsymbol{\beta}_k, y_{ki}) + \sum_{j=1}^{p} |\alpha_j| + \lambda_1 \sum_{k=1}^{M} \sum_{j=1}^{p} |\gamma_{kj}|, \quad (2.4)$$

where $\mathcal{H}$ is a set of indexes as in (2.2) and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_M^T)^T$. For the minimization problems of (2.2) and (2.4) with tuning parameter $\lambda_1 = \lambda^2/4$, (a) if $(\widehat{\boldsymbol{\alpha}}_{\mathcal{H}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{H}})$ is a solution of (2.4), then $\widehat{\boldsymbol{\beta}}_{\mathcal{H}}$ with $\hat{\beta}_{kj} = \widehat{\alpha}_j \widehat{\gamma}_{kj}$ is a solution of (2.2); and (b) if $\widehat{\boldsymbol{\beta}}_{\mathcal{H}}$ is a solution of (2.2), then there exists a solution $(\widehat{\boldsymbol{\alpha}}_{\mathcal{H}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{H}})$ of (2.4) such that $\hat{\beta}_{kj} = \widehat{\alpha}_j \widehat{\gamma}_{kj}$.

The proof of Theorem 1 is given in **Appendix A**. This theorem further verifies that the penalty term in (2.2) performs a bi-level variable selection. By a decomposition of β_{kj}, the parameter α_j controls the sparsity of the jth group $\beta_{1j}, \ldots, \beta_{Mj}$, and γ_{kj} reflects the sparsity within the jth group. If α_j is shrunk to zero, all $\beta_{kj}, k = 1, \ldots, M$ in the jth group will be zero. Since the objective function (2.4) only has two lasso penalties without interaction, Zhou and Zhu (2010) and Li et al. (2014) applied the lasso algorithm to solve $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, iteratively. Moreover, they have also implemented this algorithm by the "glmnet" in the R software.

Next, to find an approximate outlier-free subset for the M studies, we propose to combine the bi-level variable selection technique with Fast-LTS (Rousseeuw and Driessen, 2006; Alfons et al., 2013). We first introduce a definition that will be useful for the searching algorithm.

DEFINITION 1. *Let $\widehat{\boldsymbol{\beta}}_{\mathcal{H}} = (\widehat{\boldsymbol{\beta}}_{1,\mathcal{H}}^T, \ldots, \widehat{\boldsymbol{\beta}}_{M,\mathcal{H}}^T)^T$ be the estimate of $\boldsymbol{\beta}$ based on the set $\mathcal{H} = \{H_1, \ldots, H_M\}$. Then, an approximate clean subset for the kth study based on $\mathcal{H}$ is given as*

$$\widetilde{H}_k | \mathcal{H} = \arg\min_{G \in \widetilde{\mathcal{G}}_k} \sum_{i \in G} d(\boldsymbol{x}_{ki}^T \widehat{\boldsymbol{\beta}}_{k,\mathcal{H}}, y_{ki}), \quad (2.5)$$

where $\widetilde{\mathcal{G}}_k = \{G : G \subseteq \{1, \ldots, n_k\} \text{ and } |G| = h_k\}$. Furthermore, an approximate clean subset for the M studies based on $\mathcal{H}$ is defined as $\widetilde{\mathcal{H}} | \mathcal{H} = \{\widetilde{H}_1, \ldots, \widetilde{H}_M\}$.

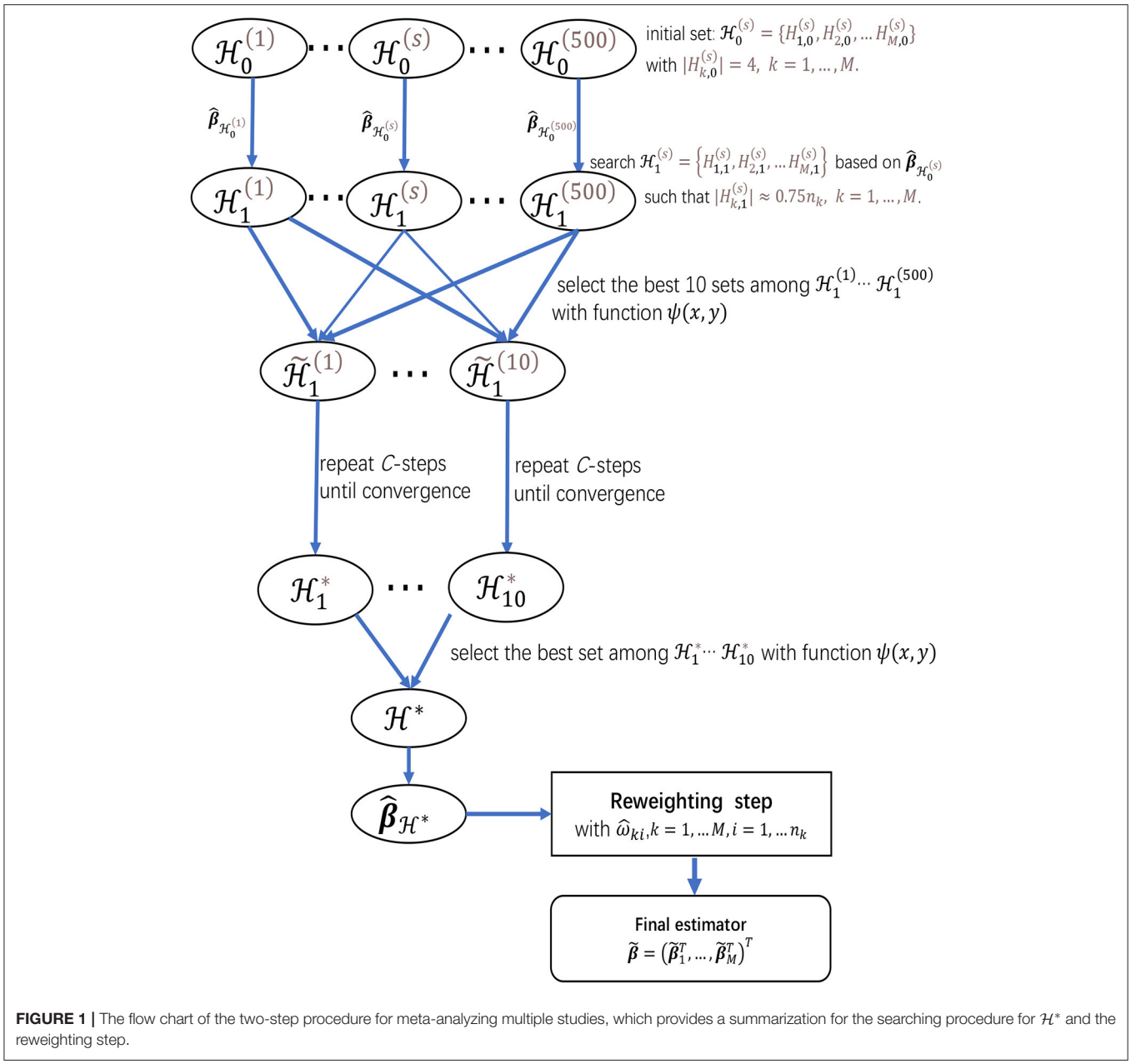


FIGURE 1 | The flow chart of the two-step procedure for meta-analyzing multiple studies, which provides a summarization for the searching procedure for $\mathcal{H}^*$ and the reweighting step.

Accordingly, let $\mathcal{H}_0$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{H}_0}$ be the initial subset for the studies and the corresponding estimate of $\boldsymbol{\beta}$, respectively. By using (2.5) recursively, we can obtain the approximate clean subset for the kth study in the tth iteration, denoted as $H_{k,t}$. Consequently, the approximate clean subset for all studies in the tth iteration is given as $\mathcal{H}_t = \{H_{1,t}, \ldots, H_{M,t}\}$. A similar procedure was also adopted in Rousseeuw and Driessen (2006) and Alfons et al. (2013); that is, selecting a subset with minimal deviance may gradually exclude outlier samples.

THEOREM 2. *For any given initial set $\mathcal{H}_0$, by recursively applying (2.5), we have*

$$Q(\mathcal{H}_{t+1}, \widehat{\boldsymbol{\beta}}_{\mathcal{H}_{t+1}}) \leq Q(\mathcal{H}_t, \widehat{\boldsymbol{\beta}}_{\mathcal{H}_t}).$$

This theorem, with the proof in **Appendix A**, shows that the objective function decreases in each iteration. Since there are only a finite number of index subsets of the collected observations, we can obtain a decreasing finite-length sequence, e.g., $Q_1 \geq Q_2 \geq \cdots \geq Q_{t_M}$ with $Q_t = Q(\mathcal{H}_t, \widehat{\boldsymbol{\beta}}_{\mathcal{H}_t})$, this shows that a convergence can be reached after a finite number of iterations (Rousseeuw and Driessen, 2006; Alfons et al., 2013). For convenience, we refer to the searching procedure in (2.5) as the concentration step (C-step). Note that the selected subset after convergence of the C-step is related to the initial subset; to alleviate this problem, we perform this searching procedure with several different initial subsets as in Alfons et al. (2013). Throughout this paper, we consider 500 initial

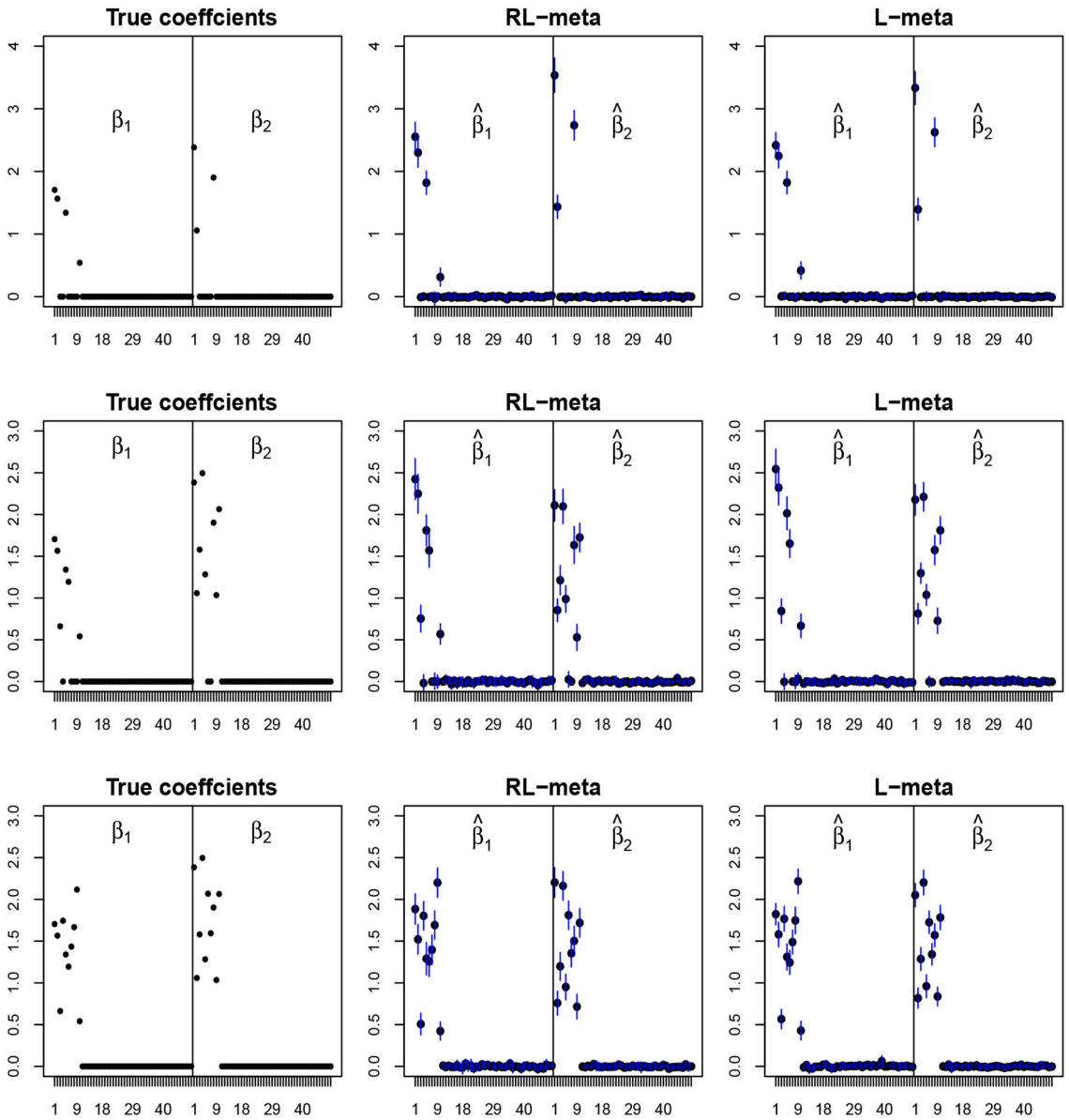


FIGURE 2 | The coefficient estimates with clean data for $M = 2$ and $(n, p) = (100, 50)$. The blue points and lines represent the mean values and the interval estimates of coefficients over 100 simulations. Rows from top to bottom correspond to $\pi_0 = 0.2, 0.5, 0.9$, respectively.

sets as $\mathcal{H}_0^{(s)} = \{H_{1,0}^{(s)}, \ldots, H_{M,0}^{(s)}\}$ for $s = 1, \ldots, 500$, where $H_{k,0}^{(s)}$ is the initial subset for the kth study. To construct $H_{k,0}^{(s)}$, we adopt a similar procedure as in Kurnaz et al. (2018), where the indexes of four observations from the kth study are randomly selected, two from each of the groups. This construction method leads to a high possibility of having no outliers in the initial subsets.

Assume that $\mathcal{H}_s^* = \{H_{s,1}^*, \ldots, H_{s,M}^*\}$ is the converged approximate clean subset based on $\mathcal{H}_0^{(s)}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{H}_s^*} = (\widehat{\boldsymbol{\beta}}_{1,\mathcal{H}_s^*}, \ldots, \widehat{\boldsymbol{\beta}}_{M,\mathcal{H}_s^*})^T$ is the resulting coefficient estimate. Then for the kth study, the index of the best clean subset among $H_{1,k}^*, \ldots, H_{500,k}^*$ can be given as

$$s_k^* = \arg \min_{s \in \{1,\ldots,500\}} \sum_{i \in H_{s,k}^*} \psi(\boldsymbol{x}_{ki}^T \widehat{\boldsymbol{\beta}}_{k,\mathcal{H}_s^*}, y_{ki}) \text{ for } k = 1, \ldots, M,$$

where $\psi(x, y = 0) = \phi(x)$, $\psi(x, y = 1) = \phi(-x)$, and $\phi(x)$ is given in Definition A1 of the **Appendix**. As mentioned in Bianco and Yohai (1996) and Crous and Haesbroeck (2003), the function $\psi(x, y)$ provides a robust loss measure for binary classification, which assigns a nearly zero score to the points with correct classification and a high score to the points with misclassification. Hence, the best clean subset for the kth study indicates the lowest

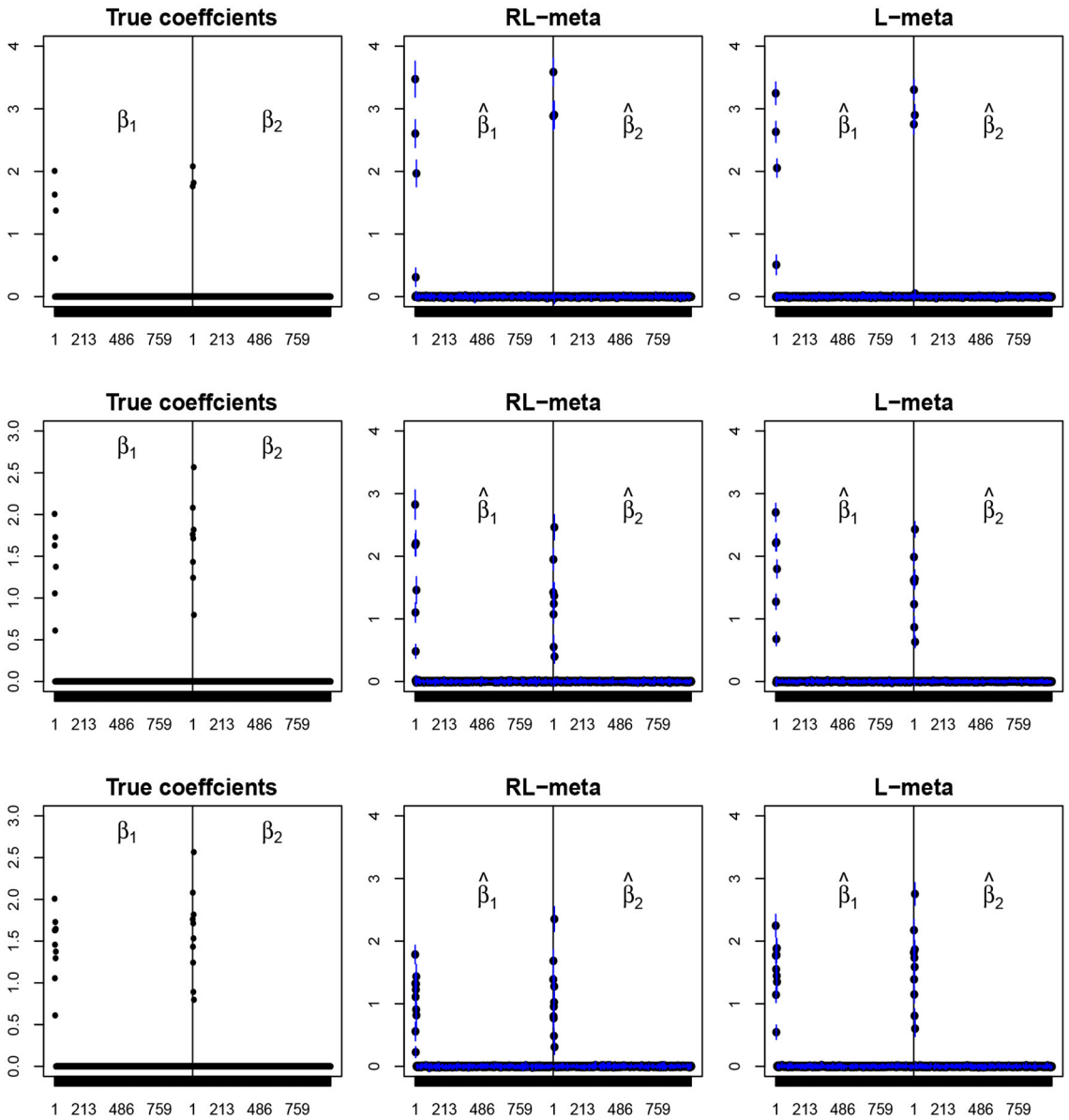


FIGURE 3 | The coefficient estimates with clean data for $M = 2$ and $(n, p) = (150, 1,000)$. The blue points and lines represent the estimated values and the interval estimates of coefficients over 100 simulations. Rows from top to bottom correspond to $\pi = 0.2, 0.5, 0.9$, respectively.

classification loss among all those identified clean subsets for this study. Finally, the best clean set for the M studies is given as $\mathcal{H}^* = \{H^*_{s^*_1,1}, \ldots, H^*_{s^*_M,M}\}$.

Also, in view of the heavy computation in the C-step on each of the 500 initial subsets. As alternative, we propose an alternative to perform two C-steps and find the best 10 subsets for the M studies as initial subsets. The rest searching procedure is similar as above paradigm. To summarize, we have the new algorithm as follows.

1. Let $\mathcal{H}^{(s)}_o = \{H^{(s)}_{1,o}, \ldots, H^{(s)}_{M,o}\}$ be the initial sets for $s = 1, 2, \ldots, 500$.
2. Let $\mathcal{H} = \mathcal{H}^{(s)}_o$ and compute the estimator for $\boldsymbol{\beta}$ by minimizing (2.4), denoted as $\widehat{\boldsymbol{\beta}}_{\mathcal{H}^{(s)}_o} = (\widehat{\boldsymbol{\beta}}^T_{1,\mathcal{H}^{(s)}_o}, \ldots, \widehat{\boldsymbol{\beta}}^T_{M,\mathcal{H}^{(s)}_o})^T$.
3. Search the approximate clean subset for the kth study as

$$H^{(s)}_{k,1} = \arg\min_{G \in \widetilde{\mathcal{G}}_k} \sum_{i \in G} d(\boldsymbol{x}^T_{ki} \widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^{(s)}_o}, y_{ki}),$$

where $\widetilde{\mathcal{G}}_k$ is the index set as in (2.5). The approximate clean subset for the M studies is $\mathcal{H}^{(s)}_1 = \{H^{(s)}_{1,1}, \ldots, H^{(s)}_{M,1}\}$.

TABLE 1 | Results for low-dimensional data with clean data.

			$(n,p) = (100, 50)$		
$M = 2$		RL-meta	L-meta	RL-each	L-each
$\pi = 0.2$	Precision	0.689 (0.019)	**0.714** (0.018)	0.623 (0.020)	0.563 (0.013)
	Recall	0.910 (0.007)	**0.934** (0.007)	0.762 (0.016)	0.909 (0.008)
	F_1	0.769 (0.012)	**0.795** (0.012)	0.653 (0.010)	0.684 (0.011)
	RMSE	0.303 (0.011)	0.272 (0.010)	0.224 (0.003)	**0.202** (0.005)
$\pi = 0.5$	Precision	0.861 (0.008)	**0.879** (0.007)	0.806 (0.015)	0.708 (0.012)
	Recall	0.950 (0.006)	**0.976** (0.040)	0.358 (0.018)	0.763 (0.018)
	F_1	0.820 (0.008)	**0.835** (0.007)	0.566 (0.010)	0.710 (0.008)
	RMSE	0.295 (0.005)	**0.253** (0.004)	0.578 (0.010)	0.456 (0.008)
$\pi = 0.9$	Precision	0.861 (0.008)	**0.879** (0.007)	0.806 (0.015)	0.708 (0.012)
	Recall	0.959 (0.006)	**0.978** (0.040)	0.358 (0.018)	0.727 (0.016)
	F_1	0.900 (0.005)	**0.920** (0.045)	0.457 (0.015)	0.069 (0.010)
	RMSE	0.302 (0.005)	**0.259** (0.034)	0.706 (0.010)	0.643 (0.013)
$M = 8$		RL-meta	L-meta	RL-each	L-each
$\pi = 0.2$	Precision	0.721 (0.009)	**0.729** (0.010)	0.597 (0.008)	0.531 (0.007)
	Recall	0.949 (0.004)	**0.953** (0.004)	0.767 (0.011)	0.942 (0.003)
	F_1	0.815 (0.005)	**0.821** (0.005)	0.665 (0.006)	0.676 (0.006)
	RMSE	0.138 (0.002)	**0.110** (0.001)	0.179 (0.002)	0.141 (0.001)
$\pi = 0.5$	Precision	**0.706** (0.007)	0.691 (0.006)	0.658 (0.007)	0.588 (0.006)
	Recall	0.987 (0.002)	**0.992** (0.001)	0.626 (0.011)	0.928 (0.004)
	F_1	**0.820** (0.005)	0.812 (0.004)	0.634 (0.006)	0.718 (0.005)
	RMSE	0.189 (0.003)	**0.176** (0.004)	0.306 (0.005)	0.266 (0.003)
$\pi = 0.9$	Precision	0.889(0.005)	**0.905** (0.005)	0.754 (0.008)	0.671 (0.006)
	Recall	0.992 (0.001)	**0.995** (0.007)	0.469 (0.002)	0.774 (0.011)
	F_1	0.936 (0.003)	**0.947** (0.003)	0.578 (0.007)	0.712 (0.005)
	RMSE	0.209 (0.002)	**0.187** (0.001)	0.634 (0.005)	0.565 (0.011)

The presented values are the means of Precision, Recall, F_1, and RMSE with standard errors in parentheses, respectively, averaged over 100 simulations. The bold values for Presicion, Recall, and F1 score are the highest values, and the bold value for RMSE is the lowest value.

4. Repeat Step 2 on $\mathcal{H} = \mathcal{H}_1^{(s)}$. Let also $\widehat{\boldsymbol{\beta}}_{\mathcal{H}_1^{(s)}} = (\widehat{\boldsymbol{\beta}}_{1,\mathcal{H}_1^{(s)}}^T, \ldots, \widehat{\boldsymbol{\beta}}_{M,\mathcal{H}_1^{(s)}}^T)^T$ be the corresponding coefficient estimate.
5. For $H_{k,1}^{(s)} \in \mathcal{H}_1^{(s)}$ with $s = 1, \ldots, 500$ and $k = 1, \ldots, M$, let

$$e_{ks} = \sum_{i \in H_{k,1}^{(s)}} \psi(\boldsymbol{x}_{ki}^T \widehat{\boldsymbol{\beta}}_{k,\mathcal{H}_1^{(s)}}, y_{ki}).$$

Search a subset of indexes such that $\{\pi_{k,1}, \ldots, \pi_{k,10}\} \subset \{1, \ldots, 500\}$ with $e_{k,\pi_{k,1}} \leq \ldots \leq e_{k,\pi_{k,10}}$. The best 10 sets among $\mathcal{H}_1^{(1)}, \ldots, \mathcal{H}_1^{(500)}$ are given as $\widetilde{\mathcal{H}}_1^{(s)} = \{H_{1,1}^{(\pi_{1,s})}, \ldots, H_{M,1}^{(\pi_{M,s})}\}$ for $s = 1, \ldots, 10$.

6. Let $\mathcal{H} = \widetilde{\mathcal{H}}_1^{(s)}$ be the initial set for $s = 1, \ldots, 10$, respectively, and repeat Steps 2–3 for a total of t times until convergence such that $||\widehat{\boldsymbol{\beta}}_{\mathcal{H}_t^{(s)}} - \widehat{\boldsymbol{\beta}}_{\mathcal{H}_{t-1}^{(s)}}||_2 \leq \epsilon$, where $||\cdot||_2$ is the Euclidean norm and ϵ is a pre-specified small constant. The converged approximate clean subset and the coefficient estimate for all M studies are denoted as $\mathcal{H}_s^* = \{H_{s,1}^*, \ldots, H_{s,M}^*\}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{H}_s^*} = (\widehat{\boldsymbol{\beta}}_{1,\mathcal{H}_s^*}^T, \ldots, \widehat{\boldsymbol{\beta}}_{M,\mathcal{H}_s^*}^T)^T$, respectively.
7. For $H_{s,k}^* \in \mathcal{H}_s^*$ with $s = 1, \ldots, 10$ and $k = 1, \ldots, M$, let

$$s_k^* = \arg \min_{s \in \{1,\ldots,10\}} \sum_{i \in H_{s,k}^*} \psi(\boldsymbol{x}_{ki}^T \widehat{\boldsymbol{\beta}}_{k,\mathcal{H}_s^*}, y_{ki}).$$

The best clean subset for all M studies is given as $\mathcal{H}^* = \{H_{s_1^*,1}^*, \ldots, H_{s_M^*,M}^*\}$, and the corresponding estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{\mathcal{H}^*} = (\widehat{\boldsymbol{\beta}}_{1,\mathcal{H}^*}^T, \ldots, \widehat{\boldsymbol{\beta}}_{M,\mathcal{H}^*}^T)^T$.

Finally, we observe that in the first several C-steps, the algorithm for minimizing (2.4) may not stable. For this, we may restrict that $\alpha_1 = \cdots = \alpha_p = 1$.

2.2.2. Reweighting Step

Note that the LTS-type estimator only uses a subset of data and may suffer from a low efficiency. To further improve the model estimation, Kurnaz et al. (2018) proposed a reweighting step such that the identified outliers based on the current estimated model will be assigned with a small or zero weight. For our robust meta-analysis method, we adopt a similar reweighting procedure, which is based on the Pearson residuals $\widehat{r}_{ki} = (y_{ki} - \widehat{\pi}_{ki})/\sqrt{\widehat{\pi}_{ki}(1 - \widehat{\pi}_{ki})}$, where $\widehat{\pi}_{ki} = \exp(\boldsymbol{x}_{ki}^T \widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*})/[1 + \exp(\boldsymbol{x}_{ki}^T \widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*})]$ is the conditional probability of the logistic model.

TABLE 2 | Results for high-dimensional data with clean data.

		$(n,p) = (150, 1,000)$			
$M = 2$		**RL-meta**	**L-meta**	**RL-each**	**L-each**
$\pi = 0.2$	Precision	0.536 (0.015)	0.489 (0.014)	0.633 (0.024)	**0.807** (0.015)
	Recall	0.900 (0.007)	**0.934** (0.007)	0.740 (0.016)	0.878 (0.006)
	F_1	0.658 (0.011)	0.632 (0.012)	0.642 (0.013)	**0.833** (0.010)
	RMSE*	0.276 (0.047)	0.239 (0.071)	**0.208** (0.054)	0.226 (0.014)
$\pi = 0.5$	Precision	**0.747** (0.012)	0.665 (0.011)	0.716 (0.021)	0.787 (0.011)
	Recall	0.965 (0.005)	**0.974** (0.010)	0.321 (0.012)	0.684 (0.013)
	F_1	**0.835** (0.007)	0.785 (0.007)	0.417 (0.011)	0.721 (0.009)
	RMSE*	0.253 (0.040)	**0.157** (0.089)	0.447 (0.091)	0.452(0.102)
$\pi = 0.9$	Precision	0.722 (0.012)	0.759 (0.011)	0.762 (0.025)	**0.790** (0.013)
	Recall	0.850 (0.010)	**0.975** (0.004)	0.178 (0.008)	0.448 (0.015)
	F_1	0.778 (0.010)	**0.849** (0.007)	0.274 (0.010)	0.548 (0.012)
	RMSE*	0.261 (0.074)	**0.196** (0.014)	0.489 (0.075)	0.446 (0.071)
$M = 8$		**RL-meta**	**L-meta**	**RL-each**	**L-each**
$\pi = 0.2$	Precision	0.694 (0.015)	0.709 (0.012)	0.573 (0.014)	**0.783** (0.008)
	Recall	**0.957** (0.004)	0.840(0.002)	0.630 (0.008)	0.863 (0.004)
	F_1	0.796 (0.007)	0.811 (0.008)	0.587 (0.007)	**0.818** (0.004)
	RMSE*	0.321 (0.068)	**0.311** (0.075)	0.509 (0.063)	0.451 (0.045)
$\pi = 0.5$	Precision	0.687 (0.005)	**0.688** (0.005)	0.664 (0.014)	0.769 (0.007)
	Recall	0.988 (0.002)	**0.994** (0.001)	0.395 (0.008)	0.784 (0.006)
	F_1	0.809 (0.004)	**0.812** (0.003)	0.483 (0.006)	0.774 (0.005)
	RMSE*	0.416 (0.064)	**0.316** (0.087)	0.447 (0.076)	0.435 (0.081)
$\pi = 0.9$	Precision	0.942 (0.002)	**0.952** (0.001)	0.712 (0.012)	0.760 (0.006)
	Recall	**0.988** (0.001)	0.964 (0.003)	0.198 (0.004)	0.473 (0.008)
	F_1	**0.965** (0.002)	0.958 (0.002)	0.305 (0.004)	0.579 (0.007)
	RMSE*	0.574 (0.072)	**0.475** (0.031)	0.639 (0.064)	0.622 (0.105)

The presented values are the means of Precision, Recall, F_1, and RMSE with standard errors in parentheses, respectively, averaged over 100 simulations. RMSE*= 10× RMSE. The bold values for Presicion, Recall, and F1 score are the highest values, and the bold value for RMSE is the lowest value.*

Since r_{ki} is a standardized statistic and is approximately normally distributed, the weights for the observations are given as

$$\widehat{w}_{ki} = \begin{cases} 0, & |\widehat{r}_{ki}| > \Phi^{-1}(1-\delta), \\ 1, & |\widehat{r}_{ki}| \leq \Phi^{-1}(1-\delta), \end{cases} \text{ for } k = 1, \ldots, M \text{ and } i = 1, \ldots, n_k, \quad (2.6)$$

where Φ is the cumulative distribution function of the standard normal distribution. Throughout this paper, we select $\delta = 0.0125$ as in Alfons et al. (2013) and Kurnaz et al. (2018) such that 2.5% of the observations from the standard normal distribution are considered as outliers. The reweighed objective function is

$$Q_w(\boldsymbol{\beta}) = \sum_{k=1}^{M} \sum_{i=1}^{n_k} \widehat{w}_{ki} d(\boldsymbol{x}_{ki}^T \boldsymbol{\beta}_k, y_{ki}) + \lambda \sum_{j=1}^{p} \left(\sum_{k=1}^{M} |\beta_{kj}| \right)^{1/2}, \quad (2.7)$$

Consequently, the robust estimator for meta-analyzing multiple studies is given as

$$\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1^T, \ldots, \widetilde{\boldsymbol{\beta}}_M^T)^T = \arg\min_{\boldsymbol{\beta}} Q_w(\boldsymbol{\beta}),$$

where $\widetilde{\boldsymbol{\beta}}_k$ is the estimate of the coefficient vector for the kth study. Obviously, when the data do not have outliers or only have a small proportion of outliers, the reweighing procedure uses more observations and hence may improve the estimation accuracy.

Finally, to give more insights on the algorithms in sections 2.2.1 and 2.2.2, we present a flow chart of the two-step procedure for meta-analyzing multiple studies in **Figure 1**, which provides a summary for the searching procedure for $\mathcal{H}^*$ and the reweighting step.

2.3. Selection of the Tuning Parameters

In section 2.2, we need to pre-specify the cardinalities $h_1, \ldots, h_M$ before searching the approximate clean subset $\widetilde{\mathcal{H}}^*$. If some studies contain a large proportion of outliers, then the cardinalities of the selected subsets from the studies ought to be small, e.g., $h_k \approx n_k/2$, and vice versa. In practice, if we do not have prior knowledge for the number of outliers, we recommend to use $h_k \approx 0.75 n_k$ as adopted in Rousseeuw and Driessen (2006), Alfons et al. (2013), and Kurnaz et al. (2018).

Note that the optimization problems in (2.2) and (2.7) can be rewritten as (2.4), and hence we only need to select the tuning parameter in (2.4). Various data-driven techniques have been well developed in the literature including, for example, the cross-validation, the generalized cross-validation, and the Bayesian information criterion (BIC) procedures. We adopt the BIC to

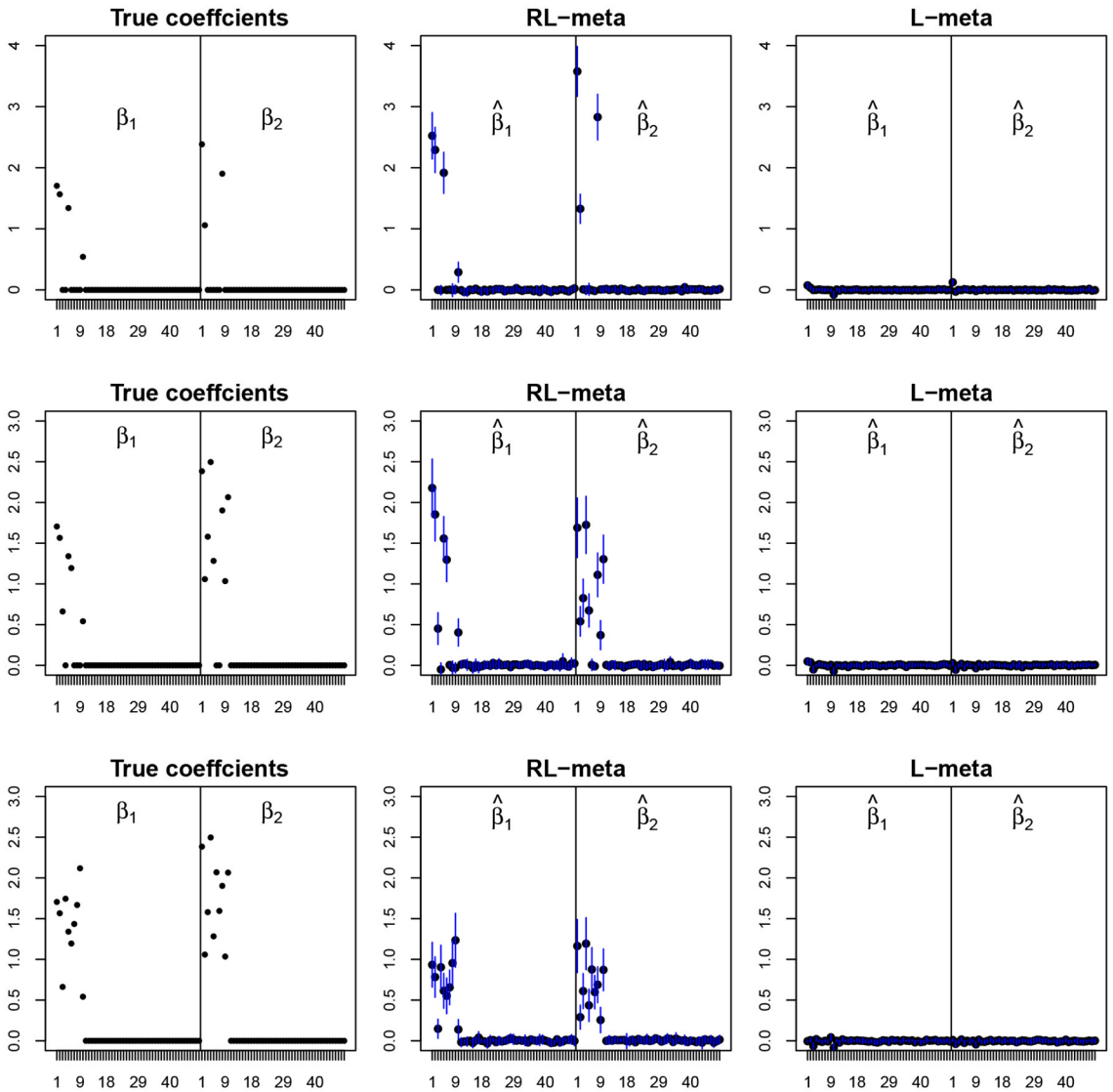


FIGURE 4 | The coefficient estimates with contamination data for $M = 2$ and $(n, p) = (100, 50)$. The blue points and lines represent the mean values and the interval estimates of coefficients over 100 simulations. Rows from top to bottom correspond to $\pi_0 = 0.2, 0.5, 0.9$, respectively.

select the tuning parameter as recommended in Alfons et al. (2013). Specifically, we compute the BIC after obtaining $\mathcal{H}^*$ in the C-steps, which is given as

$$\mathrm{BIC}(\lambda_1) = \sum_{k=1}^{M} \left\{ -2\mathcal{L}_k(\widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*}, \mathcal{H}^*) + \mathrm{df}(\widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*}) \log(h_k) \right\}, \quad (2.8)$$

where $\mathcal{L}_k(\widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*}, \mathcal{H}^*) = \sum_{i=1}^{H_k} d(\boldsymbol{x}_{ki}\widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*}, y_{ki})$ with $H_k \in \mathcal{H}^*$, and $\mathrm{df}(\widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*})$ is the number of non-zero components in $\widehat{\boldsymbol{\beta}}_{k,\mathcal{H}^*}$. A similar procedure is also performed in the reweighting procedure to select the tuning parameter.

3. NUMERICAL STUDIES

In this section, we conduct simulations to evaluate the numerical performance of our new robust lasso-type meta-analysis method (RL-meta) and compare it with some existing methods. Specifically, we consider the state-of-the-art methods from Li et al. (2014), Alfons et al. (2013), and Friedman et al. (2010). Noting that the latter two methods perform the variable selection on each study individually, hence for convenience, we refer to them as L-meta, RL-each, and L-each, respectively.

Let TP, FP, and FN be the number of true positives, false positives, and false negatives, respectively. Then to evaluate the performance of these methods, we consider four criteria including Precision = TP/(TP+FP), Recall = TP/(TP+FN), the F_1

TABLE 3 | Results for low-dimensional data with contamination.

		$(n,p) = (100, 50)$			
$M = 2$		**RL-meta**	**L-meta**	**RL-each**	**L-each**
$\pi = 0.2$	Precision	0.635 (0.019)	0.248 (0.027)	0.338 (0.007)	**0.764** (0.032)
	Recall	**0.854** (0.016)	0.184 (0.019)	0.848 (0.010)	0.138 (0.011)
	F_1	**0.728** (0.014)	0.362 (0.013)	0.480 (0.009)	0.301 (0.010)
	RMSE	0.376 (0.015)	0.449 (0.023)	**0.207** (0.009)	0.456 (0.017)
$\pi = 0.5$	Precision	0.664 (0.017)	0.537 (0.001)	0.374 (0.015)	**0.898** (0.018)
	Recall	**0.768** (0.026)	0.113 (0.026)	0.634 (0.023)	0.135 (0.010)
	F_1	**0.701** (0.020)	0.252 (0.013)	0.465 (0.017)	0.243 (0.010)
	RMSE	**0.470** (0.021)	0.725 (0.024)	0.615 (0.019)	0.725 (0.026)
$\pi = 0.9$	Precision	0.715 (0.021)	0.340 (0.029)	0.391 (0.015)	**0.921** (0.012)
	Recall	**0.636** (0.030)	0.089 (0.008)	0.518 (0.021)	0.077 (0.005)
	F_1	**0.658** (0.025)	0.188 (0.010)	0.445 (0.017)	0.159 (0.006)
	RMSE	**0.608** (0.027)	0.817 (0.031)	0.707 (0.026)	0.813 (0.034)
$M = 8$		**RL-meta**	**L-meta**	**RL-each**	**L-each**
$\pi = 0.2$	Precision	**0.682** (0.009)	0.568 (0.043)	0.291 (0.005)	0.592 (0.032)
	Recall	**0.899** (0.010)	0.105(0.010)	0.842 (0.011)	0.051 (0.003)
	F_1	**0.770** (0.012)	0.251 (0.005)	0.432 (0.006)	0.104 (0.004)
	RMSE	**0.216** (0.014)	0.289(0.021)	0.219 (0.018)	0.927 (0.013)
$\pi = 0.5$	Precision	0.691 (0.010)	0.128 (0.017)	0.358 (0.010)	**0.840** (0.018)
	Recall	**0.923** (0.011)	0.042 (0.005)	0.724 (0.012)	0.102 (0.003)
	F_1	**0.787** (0.005)	0.118 (0.007)	0.479 (0.008)	0.181 (0.005)
	RMSE	**0.251** (0.016)	0.473 (0.021)	0.352 (0.019)	0.480 (0.023)
$\pi = 0.9$	Precision	0.828 (0.010)	0.238 (0.028)	0.358 (0.007)	**0.939** (0.011)
	Recall	**0.866** (0.010)	0.053 (0.006)	0.519 (0.012)	0.052 (0.002)
	F_1	**0.839** (0.013)	0.161 (0.008)	0.423 (0.008)	0.098 (0.004)
	RMSE	**0.516** (0.027)	0.701(0.029)	0.680 (0.015)	0.699 (0.030)

The presented values are the means of Precision, Recall, F_1 score, and RMSE with standard errors in parentheses, respectively, averaged over 100 simulations. The bold values for Presicion, Recall, and F1 score are the highest values, and the bold value for RMSE is the lowest value.

score (F_1), and the root mean squared error (RMSE), where

$$F_1 = \frac{2 \times \text{Presicion} \times \text{Recall}}{\text{Presicion} + \text{Recall}} \quad \text{and}$$

$$\text{RMSE} = \Big(\frac{1}{M}\sum_{k=1}^{M}\sum_{j=1}^{p}(\hat{\beta}_{kj} - \beta_{kj})^2\Big)^{1/2}.$$

Note that Precision, Recall, and F_1 all range from 0 to 1, which measure the accuracy of variable selection with a larger value being preferred. As an addition, RMSE measures the loss of coefficient estimation, for which a small value is favorable.

3.1. Clean Data

In the first simulation, we consider clean data with no outliers. Specifically, we generate M studies and each has n observations. The covariate vector $\boldsymbol{x}_{ki} = (x_{ki1}, \ldots, x_{kip})^T$ are randomly sampled from the multivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_p)$ for $k = 1, \ldots, M$ and $i = 1, \ldots, n$, where $\boldsymbol{I}_p$ is the identity matrix. Then the response variables are generated as $y_{ki} = 1$ if $\boldsymbol{x}_{ki}\boldsymbol{\beta}_k + \varepsilon_{ki} > 0$, otherwise, $y_{ki} = 0$, where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^T$, ε_{ki} is the independent noise sampled from $N(0, 1)$. To access the performance of our RL-meta under different levels of heterogeneity, we let $\beta_{kj} = z_{kj}b_{kj}$ for $k = 1, \ldots, M$ and $j = 1, \ldots, 10$ and $\beta_{kj} = 0$ for $j = 11, \ldots, p$, where z_{kj} is randomly sampled from a Bernoulli distribution with the success rate π_0 and b_{kj} is randomly sampled from $N(1.5, 0.5^2)$. This means that only the first 10 covariates may be possibly related to the response variable in each study, it is informative with probability π_0 and uninformative with probability $1 - \pi_0$. We consider π_0 = 0.2, 0.5, or 0.9 to represent three levels of heterogeneity between the studies. In addition, we also consider $(n = 100, p = 50)$ or $(n = 150, p = 1,000)$ as a low or large dimension, respectively, and the numbers of studies as $M = 2$ or 8.

To visualize the coefficient estimation for more insights, we plot the average values of the estimates for each coefficient with the confidence intervals (mean $\pm 3\times$ standard error) for $M = 2$ studies (see **Figures 2**, **3**). To save space, we move the plots of L-each and RL-each to **Appendix B** (see **Figures S1**, **S2**). When there is no outlier, RL-meta and L-meta both provide good estimates for the coefficients, where they are close to the true coefficients especially with a low dimension (e.g., $p = 50$). Figure A1 shows that RL-each and L-each can provide an accurate variable selection, whereas the estimates for nonzero coefficients tend to be smaller than the true coefficients, especially when the dimension is larger than the sample size. This phenomenon was also observed in Zhao and Yu (2006), where they showed that the convex penalty often shrinks the estimates of the nonzero

TABLE 4 | Results for high-dimensional data with contamination data.

		$(n,p) = (150, 1,000)$			
$M = 2$		**RL-meta**	**L-meta**	**RL-each**	**L-each**
$\pi = 0.2$	Precision	0.510 (0.017)	0.076 (0.006)	**0.878** (0.009)	0.273 (0.043)
	Recall	**0.873** (0.009)	0.324 (0.014)	0.261 (0.005)	0.018 (0.005)
	F_1	**0.613** (0.013)	0.125 (0.007)	0.202 (0.003)	0.246 (0.001)
	RMSE*	0.298 (0.032)	0.379 (0.045)	**0.161** (0.008)	0.316(0.032)
$\pi = 0.5$	Precision	**0.463** (0.024)	0.104 (0.009)	0.139 (0.006)	0.314 (0.044)
	Recall	**0.685** (0.027)	0.082 (0.004)	0.553 (0.020)	0.015 (0.003)
	F_1	**0.563** (0.023)	0.082 (0.005)	0.226 (0.009)	0.136 (0.002)
	RMSE*	**0.387** (0.069)	0.548 (0.093)	0.425(0.060)	0.500 (0.073)
$\pi = 0.9$	Precision	0.414 (0.036)	0.029 (0.004)	0.108 (0.007)	**0.583** (0.046)
	Recall	**0.531** (0.030)	0.055 (0.006)	0.348 (0.023)	0.022 (0.003)
	F_1	**0.499** (0.025)	0.070 (0.003)	0.181 (0.010)	0.107 (0.003)
	RMSE*	**0.435** (0.070)	0.548 (0.071)	0.489 (0.083)	0.561 (0.071)
$M = 8$		**RL-meta**	**L-meta**	**RL-each**	**L-each**
$\pi = 0.2$	Precision	**0.660** (0.011)	0.053 (0.001)	0.120 (0.002)	0.374 (0.033)
	Recall	**0.915** (0.007)	0.664(0.009)	0.810 (0.008)	0.040 (0.004)
	F_1	**0.760** (0.007)	0.097 (0.001)	0.208 (0.003)	0.106 (0.004)
	RMSE*	**0.210** (0.011)	0.374 (0.039)	0.227 (0.091)	0.292 (0.018)
$\pi = 0.5$	Precision	**0.660** (0.007)	0.044 (0.002)	0.145 (0.003)	0.383 (0.034)
	Recall	**0.906** (0.008)	0.323 (0.013)	0.675 (0.011)	0.019 (0.002)
	F_1	**0.760** (0.005)	0.077 (0.003)	0.237 (0.005)	0.059 (0.003)
	RMSE*	**0.316** (0.098)	0.547 (0.107)	0.519 (0.095)	0.469 (0.026)
$\pi = 0.9$	Precision	**0.752** (0.011)	0.048 (0.002)	0.087 (0.002)	0.584 (0.024)
	Recall	**0.767** (0.021)	0.219 (0.011)	0.433 (0.010)	0.032 (0.002)
	F_1	**0.753** (0.015)	0.080 (0.004)	0.144 (0.003)	0.062 (0.003)
	RMSE*	**0.358** (0.097)	0.614(0.085)	0.503 (0.082)	0.471 (0.074)

The presented values are the means of Precision, Recall, F_1 score, and RMSE with standard errors in parentheses, respectively, averaged over 100 simulations. The bold values for Presicion, Recall, and F1 score are the highest values, and the bold value for RMSE is the lowest value.

coefficients too heavily. In contrast, since our RL-meta and L-meta both use a nonconvex regularization, they are able to reduce the estimation biases.

Tables 1, **2** show the scores of the measure criteria for each scenario with clean data. Based on the evaluation criteria, L-meta exhibits superiority than the other three methods, which has a higher Precision, Recall, and F_1 in most settings. RL-meta is nearly as good as L-meta and is much better than L-each and RL-each. For example, when the dimension and the number of informative covariates tend to large, L-each and RL-each both exhibit an inflated RMSE, whereas RL-meta and L-meat still keep a low RMSE. This verifies that borrowing information across the studies can improve the estimation accuracy, especially when the dimension is large and the sample size is small.

3.2. Contamination Data

In the second simulation, we consider contamination data with outliers. We randomly select m_0 observations from each study and add outliers to those informative covariates. More specifically, for the observations with $y_{ki} = 0$ (or $y_{ki} = 1$), the informative covariates are replaced by values randomly sampled from $N(5, 1)$. To avoid high-leverage points, those observations are assigned an opposite class label. That is, $y_{ki} = 1 - \delta(\boldsymbol{x}_{ki}\boldsymbol{\beta}_k > 0)$, where $\delta(\cdot)$ is an indicator function. Throughout this section, we fix $m_0 = 10$, and the other parameter are the same as those in section 3.1.

Figures 4, **5** plot the mean values of the estimates for each coefficient with the confidence intervals (mean $\pm 3\times$ standard error) for $M = 2$ studies under contamination data. To save space, we also move the plots of L-each and RL-each to **Appendix B** (see **Figures S3**, **S4**). From those figures, it is evident that RL-meta outperforms the other three methods in the presence of outliers. In particular, RL-meta and L-meta are able to select more informative covariates, whereas, L-meta and L-each both miss most informative variables, especially when the dimension is large. As we mentioned in the Introduction, this may due to the fact that classical lasso-type variable selection is sensitive to outliers and has a high-break down point.

Tables 3, **4** show the scores of the four measure criteria for each scenario under contamination data. RL-meta and RL-each both exhibit a higher Precision and Recall and a smaller RMSE than L-meta and L-each, especially when the number of informative covariates is large. This indicates that the two robust methods are able to identify more informative covariates and also keep a low false discovery rate when presenting outliers. When the number of studies and the number of informative

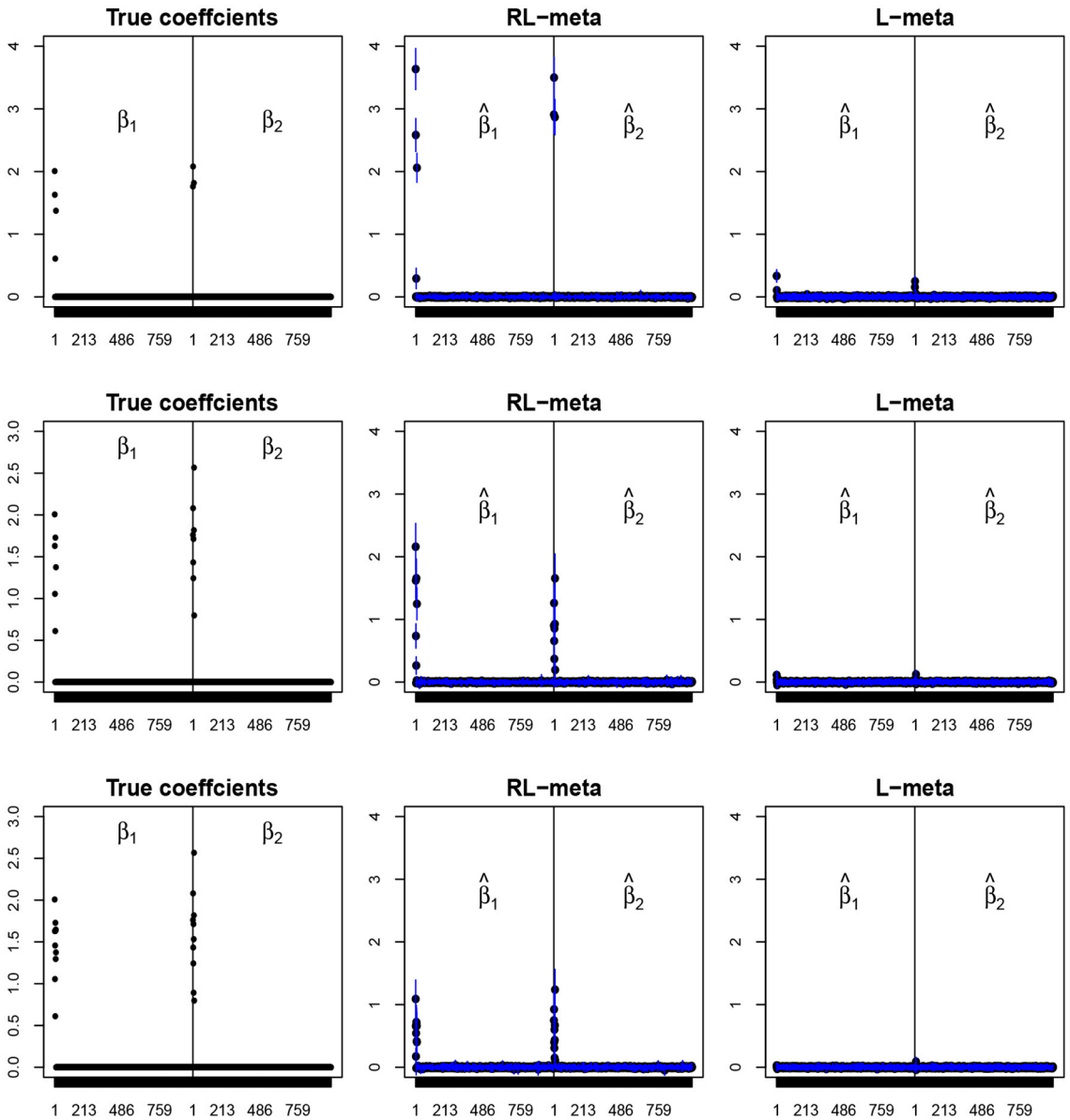


FIGURE 5 | The coefficient estimates with contamination data for $M = 2$ and $(n, p) = (150, 1, 000)$. The blue points and lines represent the estimated values and the interval estimates of coefficients over 100 simulations. Rows from top to bottom correspond to $\pi = 0.2, 0.5, 0.9$, respectively.

variable are both small (e.g., $M = 2$ and $\pi = 0.2$), we note that RL-each has a smaller RMSE than RL-meta, which exhibits a good coefficient estimation. One possible reason is that when the number of studies and informative covariates is very small, there no or little information can be borrowed to improve the estimation accuracy. As the number of studies and/or the number of informative variable tend to large, our RL-meta consistently has the best performance among the three methods including L-meta, RL-each, and L-each. This again verifies that borrowing information across similar studies can significantly improve parameter estimation and the accuracy of variable selection (Liu et al., 2011).

4. REAL DATA APPLICATION

In this section, we consider three publicly available lung cancer datasets from GEO (https://www.ncbi.nlm.nih.gov/gds/). The first data are the gene expression signature of cigarette smoking (GSE10072), which contains the gene expression levels of 49 normal and 58 tumor tissues from 28 current smokers, 26 former smokers, and 20 never smokers, and each sample has 22,283 genes. The second data are the early stage non-small-cell lung cancer (GSE19188), which contains the gene expression levels of 65 adjacent normal and 91 tumor tissues, and each sample has 54,675 genes. The third data are the non-smoking

TABLE 5 | Gene selections of RL-meta, L-meta, RL-each, and L-each in three lung cancer studies.

	GSE10072	GSE19188	GSE19804
RL-meta	**AF007147**	**AF007147**, ACSL4, GRIA1	
	SVEP1	**SVEP1**, ***EHD1***, ***LGALSL***	***EHD1***, ***LGALSL***
	COL10A1	COL10A1, ***FUT2***	COL10A1, ***FUT2***
L-meta	COL10A1	ACSL4, COL10A1	COL10A1
	AF007147	**AF007147**	FUT2
	LINC01140	**LINC01140**	
RL-each	CA4, CD36	AGER, CA4	CA4, SGCG
	SPP1, GPM6A	GDF10, GAPDH	MME
	FAM107A, FCN3	FAM189A2, LRRC36	
L-each	PDE2A	AOX1, AF007147, ACADL	ALDH18A1, COL10A1
	SPP1	GAPDH, PAFAH1B3	GOLM1, MME
	TNXA	LRRC36, LINC00341	EFNA4, SORD
		HIST1H2BD, PPBP	SPOCK2, HN1L
		CCL23, FCN3	

The bolded and italic bolded names are overlapped identified genes between GSE10072 and GSE19188 and between GSE19188 and GSE19804, respectively. The underlined names are overlapped identified genes across the three studies.

TABLE 6 | Gcta, L-meta, RL-each, and L-each in three lung cancer studies.

	GSE10072	GSE19188	GSE19804
RL-meta	COL10A1	COL10A1, ACSL4	COL10A1
	SPOCK2	***FUT2***, GRIA1, TYRP1	***FUT2***
	LINC01140	***EHD1***, AF007147	***EHD1***
L-meta	MIF	***CFP***, MIF	***CFP***, MIF, GOLM1
	KCNJ8	KCNJ8	KCNJ8, COL10A1
RL-each	CA4	AGER, CA4	CA4, SGCG
	GPM6A	LRRC36, GDF10	SH3GL3, MASP1
			COL10A1, SPP1
L-each	FAM107A	AGER, GAPDH	COL10A1
	SPP1	PAFAH1B3, NEK2	GOLM1
		MIF, HIST1H2BD	SPOCK2

The bolded and italic bolded names are overlapped identified genes between GSE10072 and GSE19188 and between GSE19188 and GSE19804, respectively. The underlined names are overlapped identified genes across the three studies.

female lung cancer in Taiwan (GSE19804), which contains the gene expression levels of 60 normal and 60 tumor tissues, and each sample has 54,675 genes. Consequently, there are 13,515 common genes shared between these three datasets. We map the probes of the raw data to gene names by annotation packages in Bioconductor. Also as per Hui et al. (2020) and Alfons et al. (2013), if multiple probes match a same gene, we compute the median values of these probes as the expression values for this gene and do the normalization for the raw expression data by the "affy" R package. Let $|t_{kj}|$ be the absolute value of standardized mean difference for the jth gene in the kth dataset and $T_j = \max(|t_{1j}|, |t_{2j}|, |t_{3j}|)$. We select the 1,000 genes with the largest values of T_j for $j = 1, \ldots, 13,515$, and then perform the variable selection for the three datasets based on RL-meta, L-meta, RL-each, and L-each, respectively.

Figure 6 shows the density plots for each of the selected 1,000 genes in the GSE10072, GSE19198, and GSE19804, respectively. The expression levels of some genes in GSE10072 and GSE19198 exhibit heavy-tailed distributions and may present outliers. **Table 5** shows the detected informative genes by RL-meta, L-meta, RL-each, and L-each in the three lung cancer studies. We observe that RL-meta detects 7 more genes than L-meta, and both of the methods identify one common gene "COL10A1" between the three studies. In addition, RL-meta detects four overlapped informative genes in GSE19188 and GSE19804, whereas L-meta only detects 1 overlapped gene. Since GSE19188 and GSE19804 are both from the same Affymetrix Platform (U133 Plus 2.0), it is expected that RL-meta has a higher detection power and is also more reproducible than L-meta. Finally, RL-each and L-each detect 15 and 22 genes, respectively. Nevertheless, these two methods identify very different genes across the three studies and so may lack of reproducibility.

To further compare the performance of the four methods, we also consider to create outliers for the three datasets. Specifically,

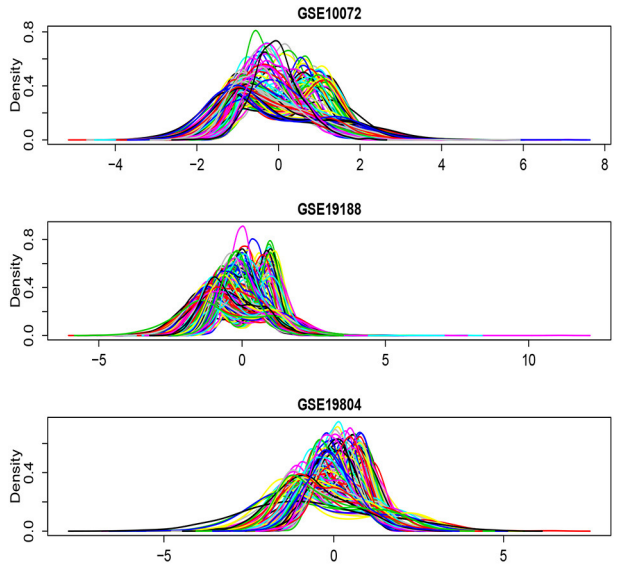


FIGURE 6 | Density plots for each of the selected 1,000 genes in the GSE10072, GSE19198, and GSE19804, respectively.

we select the first eight samples from each of the three datasets, and then add a number 10 to the expression levels of those informative genes. In order to generate outliers instead of leverage points, we assign the labels of those samples to their opposite class. **Table 6** shows the identified informative genes with RL-meta, L-meta, RL-each, and L-each in the artificially created three datasets. L-meta and L-each both identify quite different genes between the artificially created datasets and the original datasets, whereas RL-meta and RL-each identify more common genes between the artificial created datasets and the original datasets. In addition, RL-meta detects four overlapped informative genes in artificially created GSE19188 and GSE19804, whereas L-meta only detects one overlapped gene. As we discussed in the analysis of original datasets, GSE19188 and GSE19804 are both from the same Affymetrix Platform, and hence it is expected that RL-meta is more reproducible than L-meta. To conclude, RL-meta is more robust and tends to be more powerful when outliers present in the datasets.

5. DISCUSSION

In this paper, we propose a robust method for meta-analyzing multiple studies with high-dimensional data. Our method is based on a two-step procedure including a search step for a clean subset from each study and a reweighting scheme to improve the estimation efficiency. In particular, we incorporate the bi-level variable selection technique into both of the two steps. Our new robust method has the capacity to capture the sparsity at both the study and group levels so as to better integrating the structural information that can enhance the parameter estimation and variable selection. Simulation studies demonstrate that, in the presence of outliers, our proposed method can provide better parameter estimation and also identify informative covariates more accurately than the existing strategies, especially when the dimension is large. We also provide a comparison of computational cost for RL-meta, RL-each, L-meta, and L-each in Table A1. We note that RL-meta and RL-each suffer from a heavy computational burden. The main reason is that the two robust methods need to perform *C*-steps with different starting subsets, and hence the number of iterations is considerably higher than the classical lasso-based methods.

In addition, the lasso-based variable selection methods usually suffer from a low power when some covariates are highly correlated. As an alternative, Zou and Hastie (2005) and Tibshirani et al. (2005) proposed the elastic net and the fused lasso penalty to handle correlations among covariates. Following this direction, our RL-meta may further be improved by incorporating the correlated covariates. Specifically, with the hierarchical reparameterization in Theorem 1, one possible extension of (2.4) can be given as:

$$Q_{net}(\mathcal{H}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{k=1}^{M} \sum_{i \in H_k} d(\boldsymbol{x}_{ki}\boldsymbol{\beta}_k, y_{ki}) + \sum_{j=1}^{p} (|\alpha_j| + |\alpha_j|^2) + \lambda_1 \sum_{k=1}^{M} \sum_{j=1}^{p} (|\gamma_{kj}| + |\gamma_{kj}|^2).$$

We leave this problem for further theoretical and numerical studies.

Finally, we note that Bayesian meta-analysis is also a popular approach for the integration of multiple studies. Recently, Cai et al. (2020) proposed a Bayesian variable selection approach for joint modeling multiple datasets. They developed a hierarchical spike-and-slab prior (a Bayesian version of the bi-level lasso penalty) to borrow information across the studies, which is shown to have a higher power for detecting informative single nucleotide polymorphisms in genome-wide association studies (GWAS). In addition, Pickrell (2014) proposed a Bayesian hierarchical model for GWAS data by borrowing information from functional genomic studies. As a future work, it would be worthwhile to develop such Bayesian methods for robustly meta-analyzing multiple datasets and make a comparison with the RL-meta and L-meta methods introduced in the current paper.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be downloaded from the link: https://www.ncbi.nlm.nih.gov/gds/.

AUTHOR CONTRIBUTIONS

ZH developed the study and performed the simulation studies and the real data analysis. YZ and TT initiated the study and provided helpful discussions. All authors contributed to the article and approved the final version.

FUNDING

ZH's research was supported by the National Natural Science Foundation of China (No. 12001378), the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515110449), and the Natural Science Foundation of Gangdong Province (No. 2020A1515010372). YZ's research was supported by the National Natural Science Foundation of China (Grant Nos. 12071305, 11871390, and 11871411), the Natural Science Foundation of Guangdong Province of China under grant 2020B1515310008, the Project of Educational Commission of Guangdong Province of China under grant 2019KZDZX1007. TT's research was supported by the National Natural Science Foundation of China (No. 1207010822), the General Research Fund (No. HKBU12303918), and the Initiation Grant for Faculty Niche Research Areas (Nos. RC-IG-FNRA/17-18/13, RC-FNRA-IG/20-21/SCI/03) of Hong Kong Baptist University.

ACKNOWLEDGMENTS

The authors sincerely thank the editor, the associate editor, and the two reviewers for their constructive comments that have led to a substantial improvement of this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.656826/full#supplementary-material

REFERENCES

Alfons, A., Croux, C., and Sarah, G. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat*. 7, 226–248. doi: 10.1214/12-AOAS575

Bianco, A. M., and Yohai, V. J. (1996). *Robust Estimation in the Logistic Regression Model*. New York, NY: Springer. doi: 10.1007/978-1-4612-2380-1_2

Cai, M. X., Dai, M. W., Ming, J. S., Peng, H., Liu, J., and Yang, C. (2020). BIVAS: A scalable Bayesian method for bi-level variable selection with applications. *J. Comput. Graph. Stat*. 29, 40–52. doi: 10.1080/10618600.2019.1624365

Chang, L. C., Lin, U., Sibille, E., and Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics* 14:368. doi: 10.1186/1471-2105-14-368

Chi, E. C., and Scott, D. W. (2014). Robust parametric classification and variable selection by a minimum distance criterion. *J. Comput. Graph. Stat*. 23, 111–128. doi: 10.1080/10618600.2012.737296

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19, 84–90. doi: 10.1093/bioinformatics/btg1010

Crous, C., and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Comput. Stat. Data Anal*. 44, 273–295. doi: 10.1016/S0167-9473(03)00042-2

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*. 33, 1–22. doi: 10.18637/jss.v033.i01

George, K. A. (2019). Individual participant data meta-analysis explained. *J. Pediatr*. 207, 265–266. doi: 10.1016/j.jpeds.2018.12.046

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educ. Res*. 5, 3–8. doi: 10.3102/0013189X005010003

Hadi, A. S., and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *J. Am. Stat. Assoc*. 88, 1264–1272. doi: 10.1080/01621459.1993.10476407

He, Q., Zhang, H. H., Avery, C. L., and Lin, D. Y. (2016). Sparse meta-analysis with high-dimensional data. *Biostatistics* 2, 205–220. doi: 10.1093/biostatistics/kxv038

Houlston, R. S., Webb, E., Broderick, P., and et al. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet*. 40, 1426–1435. doi: 10.1038/ng.262

Hui, Z., Li, S. J., Zhang, H., Yang, Z. Y., and Liang, Y. (2020). Meta-analysis based on nonconvex regularization. *Sci. Rep*. 10:5755. doi: 10.1038/s41598-020-62473-2

Kim, S. H., Jhong, J. H., Lee, J. J., and Koo, J. Y. (2017). Meta-analytic support vector machine for integrating multiple omics data. *BioData Mining* 10, 18–32. doi: 10.1186/s13040-017-0128-6

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometr. Intell. Lab. Syst*. 172, 211–222. doi: 10.1016/j.chemolab.2017.11.017

Li, J., and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat*. 5, 994–1019. doi: 10.1214/10-AOAS393

Li, Q., Wang, S., Huang, C., Yu, M., and Shao, J. (2014). Meta-analysis based variable selection for gene expression data. *Biometrics* 70, 872–880. doi: 10.1111/biom.12213

Liu, F., Dunson, D., and Zou, F. (2011). High-dimensional variable selection in meta-analysis for censored data. *Biometrics* 67, 504–512. doi: 10.1111/j.1541-0420.2010.01466.x

Pickrell, J. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet*. 94, 559–573. doi: 10.1016/j.ajhg.2014.03.004

Rashid, N. U., Li, Q., Yeh, J., and Ibrahim, J. G. (2020). Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *J. Am. Stat. Assoc*. 115, 1125–1138. doi: 10.1080/01621459.2019.1671197

Ren, J., Du, Y., Li, S., Ma, S., Jiang, Y., and Wu, C. (2019). Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genet. Epidemiol*. 43, 276–291. doi: 10.1002/gepi.22194

Rousseeuw, P. J., and Driessen, K. V. (2006). Computing LTS regression for large data sets. *Data Mining Knowl. Discov*. 12, 29–45. doi: 10.1007/s10618-005-0024-4

Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York, NY: John Wiley and Sons. doi: 10.1002/0471725382

Sun, Q., Zhou, W., and Fan, J. (2020). Adaptive Huber regression. *J. Am. Stat. Assoc*. 529, 254–265. doi: 10.1080/01621459.2018.1543124

Tang, L., and Song, P. X. K. (2016). Fused lasso approach in regression coefficients clustering-learning parameter heterogeneity in data integration. *J. Mach. Learn. Res*. 17, 1–23. Available online at: https://jmlr.org/papers/volume17/15-598/15-598.pdf

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B* 67, 91–108. doi: 10.1111/j.1467-9868.2005.00490.x

Tsybakov, A. B., and Vande, S. A. (2005). Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Stat*. 33, 1203–1224. doi: 10.1214/009053604000001066

Wu, C., and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. *Brief. Bioinformatics* 16, 873–883. doi: 10.1093/bib/bbu046

Wu, C., Zhou, F., Li, X., Jiang, Y., and Ma, S. (2019). A selective review of multi-level omics data integration using variable selection. *High-Through*. 8, 104–129. doi: 10.3390/ht8010004

Yang, E., Lozano, A. C., and Aravkin, A. (2018). A general family of trimmed estimators for robust high-dimensional data analysis. *Electron. J. Stat*. 12, 3519–3553. doi: 10.1214/18-EJS1470

Yohai, V. J. (1987). High breakdown point and high efficiency robust measures of scales. *Ann. Stat*. 15, 642–656. doi: 10.1214/aos/1176350366

Zhang, H., Tong, T., Landers, J., and Wu, Z. (2020). TFisher: a powerful truncation and weighting procedure for combining *p*-values. *Ann. Appl. Stat*. 14, 178–201. doi: 10.1214/19-AOAS1302

Zhao, P., and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res*. 7, 2541–2563. Available online at: https://jmlr.csail.mit.edu/papers/v7/zhao06a.html

Zhao, Q., Shi, X., Huang, J., Liu, J., Li, Y., and Ma, S. (2015). Integrative analysis of "-omics" data using penalty functions. *Wiley Interdisc. Rev. Comput. Stat*. 7, 99–108. doi: 10.1002/wics.1322

Zhou, N., and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Stat. Interface* 4, 54–69. doi: 10.4310/SII.2010.v3.n4.a13

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 20 September 2021
doi: 10.3389/fgene.2021.704538

CoMM-S^4: A Collaborative Mixed Model Using Summary-Level eQTL and GWAS Datasets in Transcriptome-Wide Association Studies

Yi Yang[†], *Kar-Fu Yeung*[†] *and Jin Liu* *

Centre for Quantitative Medicine, Program in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore

Motivation: Genome-wide association studies (GWAS) have achieved remarkable success in identifying SNP-trait associations in the last decade. However, it is challenging to identify the mechanisms that connect the genetic variants with complex traits as the majority of GWAS associations are in non-coding regions. Methods that integrate genomic and transcriptomic data allow us to investigate how genetic variants may affect a trait through their effect on gene expression. These include CoMM and CoMM-S^2, likelihood-ratio-based methods that integrate GWAS and eQTL studies to assess expression-trait association. However, their reliance on individual-level eQTL data render them inapplicable when only summary-level eQTL results, such as those from large-scale eQTL analyses, are available.

Result: We develop an efficient probabilistic model, CoMM-S^4, to explore the expression-trait association using summary-level eQTL and GWAS datasets. Compared with CoMM-S^2, which uses individual-level eQTL data, CoMM-S^4 requires only summary-level eQTL data. To test expression-trait association, an efficient variational Bayesian EM algorithm and a likelihood ratio test were constructed. We applied CoMM-S^4 to both simulated and real data. The simulation results demonstrate that CoMM-S^4 can perform as well as CoMM-S^2 and S-PrediXcan, and analyses using GWAS summary statistics from Biobank Japan and eQTL summary statistics from eQTLGen and GTEx suggest novel susceptibility loci for cardiovascular diseases and osteoporosis.

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

Keywords: summary statistics, genome-wide association studies, variational bayesian, parameter expanded expectation-maximization (PX-EM) algorithm, transcriptome-wide association studies

OPEN ACCESS

Edited by:
Jiebiao Wang,
University of Pittsburgh, United States

Reviewed by:
Xiangyu Luo,
Renmin University of China, China
Huwenbo Shi,
Harvard University, United States

***Correspondence:**
Jin Liu
jin.liu@duke-nus.edu.sg

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *03 May 2021*
Accepted: *03 September 2021*
Published: *20 September 2021*

Citation:
Yang Y, Yeung K-F and Liu J (2021) CoMM-S^4: A Collaborative Mixed Model Using Summary-Level eQTL and GWAS Datasets in Transcriptome-Wide Association Studies. Front. Genet. 12:704538. doi: 10.3389/fgene.2021.704538

1 INTRODUCTION

Genome-wide association studies (GWAS) have identified a large number of genetic risk variants associated with complex traits, with over 250,000 single nucleotide polymorphism (SNP)-trait associations tagged as significant in the NHGRI-EBI GWAS Catalog (Buniello et al., 2018). However, the specific biological mechanisms through which the identified genetic variants affect these traits

have yet to be elucidated. Genetic variants may influence complex traits by altering gene expression and, consequently, protein abundance. These genetic variants may be within the regulatory sequences or secondary motifs of the target gene (cis regulation), or may affect genes at larger genomic distances by modifying upstream regulators which interact with the *cis*-regulatory sequences (Williams et al., 2007).

Transcriptome-wide association studies (TWAS) aim to provide insights into the specific mechanisms through which variants affect traits. In TWAS, the gene expression of GWAS samples is predicted with the aid of an eQTL dataset; the predicted expression is then analysed for any association with the trait of interest. Unlike approaches that examine gene expression and genetic variants in a pairwise manner, TWAS consider the combinatory effects of all genetic variants within a pre-defined window of the target gene, hence it is especially effective at detecting novel susceptibility loci when multiple variants influence expression. TWAS have proved useful as a stepping stone to generate new insights to a range of complex traits, including schizophrenia (Gusev et al., 2018), glioma (Strunz et al., 2020), prostate cancer (Mancuso et al., 2018), and age-related macular degeneration (Atkins et al., 2019).

Existing TWAS methods can be categorised into two groups, depending on whether they use individual-level or summary-level GWAS data. PrediXcan (Gamazon et al., 2015) and CoMM (Yang et al., 2018) use individual-level GWAS data, while S-PrediXcan (Barbeira et al., 2018) and CoMM-S^2 (Yang et al., 2020) use summary-level GWAS data in conjunction with a matching reference panel to estimate linkage disequilibrium. Both CoMM and CoMM-S^2 account for the imputation uncertainty in the prediction step and thus are more powerful in identifying expression-trait associations than other methods. However, these methods are limited by the availability of individual-level transcriptome data, and they neglect the ready accessibility of summary-level eQTL datasets. Datasets of eQTL summary statistics are maintained by various consortia including the eQTLGen Consortium (Võsa et al., 2018) and the GTEx Consortium (The GTEx Consortium, 2020). The ability to integrate summary-level eQTL data and summary-level GWAS data would broaden the scope of studies to which TWAS can be applied.

Here we introduce a powerful strategy that integrates eQTL summary statistics (SNP-expression correlation), GWAS summary statistics (SNP-phenotype correlation), and linkage disequilibrium information from reference panels (SNP-SNP correlation) to assess the association between the *cis* component of expression and trait. We extend CoMM-S^2, a likelihood-based method which uses individual-level eQTL data to assess expression-trait association, and propose a probabilistic model, Collaborative Mixed Models using Summary Statistics from eQTL and GWAS (CoMM-S^4). Compared with CoMM-S^2, a major advantage of CoMM-S^4 is its ability to use summary-level eQTL data and integrate them with GWAS summary statistics. In CoMM-S^4, a joint likelihood is constructed using summary statistics from GWAS and eQTL studies, as well as SNP correlation information from reference panels representative of the GWAS and eQTL populations. We develop an efficient algorithm based on variational Bayes expectation-maximization and parameter expansion (PX-VBEM). To examine the expression-trait association, a likelihood ratio test is constructed.

The performance of CoMM-S^4 is assessed in simulated data, and is also applied to traits from the NFBC1966 cohort (Sabatti et al., 2009) and Biobank Japan (Ishigaki et al., 2020). The TWAS analysis using GWAS summary statistics from NFBC1966 and eQTL summary statistics from eQTLGen suggest novel susceptibility loci for lipid traits, glucose levels, insulin levels and C-reactive protein, when compared against known susceptibility loci in the GWAS Catalog (Buniello et al., 2018). Moreover, the TWAS analysis using GWAS summary statistics from Biobank Japan and eQTL summary statistics from eQTLGen and GTEx reiterate the importance of MHC molecules, interferon-gamma signalling and apoptosis for several autoimmune and infection-related traits (rheumatoid arthritis, Graves' disease, chronic hepatitis B and chronic hepatitis C), and suggest novel susceptibility loci for cardiovascular traits (congestive heart failure, ischemic stroke, peripheral artery disease) and osteoporosis.

2 MATERIALS AND METHODS

2.1 Notation

We denote the individual-level eQTL dataset for n_1 samples by $\{\mathbf{Y}, \mathbf{W}_1\}$, where $\mathbf{Y}$ is the gene expression matrix for g genes and $\mathbf{W}_1$ is the genotype matrix for m SNP positions. For the j-th gene, let $\mathbf{y}_j$ denote the gene expression vector, and $\mathbf{W}_{1j} \in \mathbb{R}^{n_1 \times m_j}$ denote the centered genotype matrix for the m_j SNPs within a pre-defined distance from the gene. In addition, we denote the individual-level GWAS dataset for n_2 samples by $\{\mathbf{z}, \mathbf{W}_2\}$, where $\mathbf{z}$ is the phenotype vector and $\mathbf{W}_2$ is the genotype matrix. Similarly, for the j-th gene, $\mathbf{W}_{2j} \in \mathbb{R}^{n_2 \times m_j}$ denotes the centered genotype matrix for the m_j SNPs within a pre-defined distance from the gene.

We have the summary statistics, in the form of z-scores, from the analysis of genetic variant-gene expression pairs in the eQTL dataset. We also have the summary statistics from single-variate analysis in the GWAS dataset. We denote the eQTL z-scores for the j-th gene by $\hat{\boldsymbol{\gamma}}_{1j} \in \mathbb{R}^{m_j}$, and the GWAS z-scores by $\hat{\boldsymbol{\gamma}}_{2j} \in \mathbb{R}^{m_j}$ ($j = 1, \ldots, g$). To model linkage disequilibrium (LD) in the eQTL and GWAS datasets, we require the SNP correlation matrices $\hat{\mathbf{R}}_{1j} \in \mathbb{R}^{m_j \times m_j}$ and $\hat{\mathbf{R}}_{2j} \in \mathbb{R}^{m_j \times m_j}$ ($j = 1, \ldots, g$) estimated using reference panels that correspond to the eQTL and GWAS populations respectively.

2.2 Model

The relationship between the j-th gene expression $\mathbf{y}_j$ and genotype $\mathbf{W}_{1j}$ is modelled as

$$\mathbf{y}_j = \mathbf{W}_{1j}\boldsymbol{\beta}_{1j} + \mathbf{e}_1, \tag{1}$$

where $\boldsymbol{\beta}_{1j} = [\beta_{1j,1}, \ldots, \beta_{1j,m_j}]^T$ is an m_j-vector of effect sizes, and $\mathbf{e}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_{e_1}^2 \mathbf{I})$ is an n_1-vector of independent noise. Similarly, the relationship between trait $\mathbf{z}$ and genotype $\mathbf{W}_{2j}$ is modelled as

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

where $\boldsymbol{\beta}_{2j} = [\beta_{2j,1}, \ldots, \beta_{2j,m_j}]^T$ is an m_j-vector of effect sizes, and $\mathbf{e}_2 \sim \mathcal{N}(\mathbf{0}, \sigma^2_{e_2}\mathbf{I})$ is an n_2-vector of independent noise. We further model the GWAS effect size as $\boldsymbol{\beta}_{2j} = \alpha_j\boldsymbol{\beta}_{1j}$, where α_j can be interpreted as the effect of gene expression on phenotype under the assumption of no horizontal pleiotropy. To perform a likelihood ratio test for the null hypothesis $\alpha_j = 0$, we first derive the form of the log-likelihood and develop an efficient algorithm to estimate its parameters.

Let $\hat{\boldsymbol{\gamma}}_{1j} = [\hat{\gamma}_{1j,1}, \ldots, \hat{\gamma}_{1j,m_j}]^T$ and $\hat{\boldsymbol{\gamma}}_{2j} = [\hat{\gamma}_{2j,1}, \ldots, \hat{\gamma}_{2j,m_j}]^T$ denote the z-scores for the eQTL and GWAS data, respectively. Let $\hat{\mathbf{s}}_{1j} = [\hat{s}_{1j,1}, \ldots, \hat{s}_{1j,m_j}]^T$ and $\hat{\mathbf{s}}_{2j} = [\hat{s}_{2j,1}, \ldots, \hat{s}_{2j,m_j}]^T$ denote the standard errors of the effect size estimators, $\hat{\boldsymbol{\beta}}_{1j}$ and $\hat{\boldsymbol{\beta}}_{2j}$, in the eQTL and GWAS analyses respectively. Using the approximated likelihood in regression with summary statistics (RSS) (Zhu and Stephens, 2017), the distribution for $\hat{\boldsymbol{\beta}}_{ij}$ can be written as $\hat{\boldsymbol{\beta}}_{ij}|\boldsymbol{\beta}_{ij}, \hat{\mathbf{R}}_{ij}, \hat{\mathbf{S}}_{ij} \sim \mathcal{N}(\hat{\mathbf{S}}_{ij}\hat{\mathbf{R}}_{ij}\hat{\mathbf{S}}_{ij}^{-1}\boldsymbol{\beta}_{ij}, \hat{\mathbf{S}}_{ij}\hat{\mathbf{R}}_{ij}\hat{\mathbf{S}}_{ij})$, where $\hat{\mathbf{S}}_{ij} = \mathrm{diag}(\hat{\mathbf{s}}_{ij})$ $(i = 1, 2)$. Details regarding this approximated distribution can also be found in related literature (Hormozdiari et al., 2014; Huang et al., 2021). In practice, we may observe only the z-scores for the summary statistics. In this case, the distribution of the eQTL z-scores $\hat{\boldsymbol{\gamma}}_{1j} = \hat{\mathbf{S}}_{1j}^{-1}\hat{\boldsymbol{\beta}}_{1j}$ can be written as

$$\hat{\boldsymbol{\gamma}}_{1j}|\boldsymbol{\gamma}_j, \hat{\mathbf{R}}_{1j} \sim \mathcal{N}(\hat{\mathbf{R}}_{1j}\boldsymbol{\gamma}_j, \hat{\mathbf{R}}_{1j}), \tag{3}$$

where $\boldsymbol{\gamma}_j = \hat{\mathbf{S}}_{1j}^{-1}\boldsymbol{\beta}_{1j}$. Similarly, the distribution of the GWAS z-scores $\hat{\boldsymbol{\gamma}}_{2j}$ can be approximated by

$$\hat{\boldsymbol{\gamma}}_{2j}|\boldsymbol{\gamma}_j, \hat{\mathbf{R}}_{2j} \sim \mathcal{N}(\alpha_j c_j\hat{\mathbf{R}}_{2j}\boldsymbol{\gamma}_j, \hat{\mathbf{R}}_{2j}), \tag{4}$$

where $c_j \approx \frac{\hat{\sigma}_{yj}}{\hat{\sigma}_z}\sqrt{\frac{n_2}{n_1}}$ when the summary statistics are generated using simple linear regression, $\hat{\sigma}_{yj}$ is the sample standard deviation for the expression of gene j, and $\hat{\sigma}_z$ is the sample standard deviation of the trait (details in **Supplementary Material**). Furthermore, a Gaussian prior is used for $\boldsymbol{\gamma}_j$,

$$\boldsymbol{\gamma}_j \sim \mathcal{N}(0, \sigma^2_{\gamma_j}\mathbf{I}_{m_j}), \tag{5}$$

and the complete-data likelihood can be written as

$$\Pr(\hat{\boldsymbol{\gamma}}_{1j}, \hat{\boldsymbol{\gamma}}_{2j}, \boldsymbol{\gamma}_j|\hat{\mathbf{R}}_{1j}, \hat{\mathbf{R}}_{2j}; \boldsymbol{\theta}) = \prod_{i=1}^{2}\mathrm{pr}(\hat{\boldsymbol{\gamma}}_{ij}|\boldsymbol{\gamma}_j, \hat{\mathbf{R}}_{ij})\mathrm{pr}(\boldsymbol{\gamma}_j), \tag{6}$$

where $\boldsymbol{\theta} = \{\sigma^2_{\gamma_j}, \alpha'_j\}$ is the collection of parameters and $\alpha'_j = \alpha_j c_j$.

We are primarily interested in the effect α_j of gene expression on trait. Notably, testing the hypothesis of whether $\alpha_j = 0$ is equivalent to testing whether $\alpha'_j = 0$, as c_j is a positive constant. The accuracy of the above distributional approximations depend on the sample sizes of the eQTL and GWAS datasets, as well as the number of SNPs/genes associated with the gene expression/phenotype. The larger the sample size and the higher the degree of polygenecity, the greater the estimation accuracy.

2.3 Parameter Expansion-Variational Bayes Expectation-Maximization Algorithm

An efficient algorithm is needed to estimate the parameters of the model. Although the EM algorithm is widely used and has a highly stable performance, it requires inverting the matrix $\hat{\mathbf{R}}_{1j}$ and $\hat{\mathbf{R}}_{2j}$ in each iteration. To speed up the computational process, we use Variational Bayes Expectation-Maximization (VBEM), augmented with parameter expansion (PX) (Liu et al., 1998). The parameter-expanded model is

$$\hat{\boldsymbol{\gamma}}_{1j}|\boldsymbol{\gamma}_j, \hat{\mathbf{R}}_{1j} \sim \mathcal{N}(\tau\hat{\mathbf{R}}_{1j}\boldsymbol{\gamma}_j, \hat{\mathbf{R}}_{1j}), \tag{7}$$

where the $\tau \in \mathbb{R}$ is the expanded parameter. The model parameters are $\boldsymbol{\theta} = \{\sigma^2_{\gamma_j}, \alpha_j, \tau\}$, and the expanded model reduces to the original one when $\tau = 1$. In VBEM, the marginal log-likelihood can be decomposed into the evidence lower bound (ELBO) and the Kullback-Liebler (KL) divergence between the variational and true posterior distribution of the latent variable $\boldsymbol{\gamma}_j$:

$$\log\Pr(\hat{\boldsymbol{\gamma}}_{1j}, \hat{\boldsymbol{\gamma}}_{2j}|\hat{\mathbf{R}}_{1j}, \hat{\mathbf{R}}_{2j}; \boldsymbol{\theta}) = \mathcal{L}(q) + \mathbb{KL}(q\|p), \tag{8}$$

where

$$\begin{aligned}\mathcal{L}(q) &= \int_{\gamma_j} q(\boldsymbol{\gamma}_j)\log\frac{Pr(\hat{\boldsymbol{\gamma}}_{1j}, \hat{\boldsymbol{\gamma}}_{2j}, \boldsymbol{\gamma}_j|\hat{\mathbf{R}}_{1j}, \hat{\mathbf{R}}_{2j}; \boldsymbol{\theta})}{q(\boldsymbol{\gamma}_j)}d\boldsymbol{\gamma}_j \\ \mathbb{KL}(q\|p) &= \int_{\gamma_j} q(\boldsymbol{\gamma}_j)\log\frac{q(\boldsymbol{\gamma}_j)}{p(\boldsymbol{\gamma}_j|\hat{\boldsymbol{\gamma}}_{1j}, \hat{\boldsymbol{\gamma}}_{2j}, \hat{\mathbf{R}}_{1j}, \hat{\mathbf{R}}_{2j}; \boldsymbol{\theta})}d\boldsymbol{\gamma}_j.\end{aligned} \tag{9}$$

We adopt the mean-field form of the variational posterior distribution

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

to speed up the computational process. The analytical form of the variational posterior distribution is obtained by minimizing the KL divergence, and the derived variational parameters are plugged back into the ELBO. The model parameters are then updated by setting the derivative of the ELBO with respect to the parameters equal to zero. By maximizing the ELBO with respect to the expanded parameter τ, we are able to further increase the ELBO and speed up the convergence process. Since the parameter-expanded model reduces to the original model when $\tau = 1$, the original model can be recovered by incorporating τ into the model parameters, as outlined in the **Supplementary Material**.

2.4 Likelihood Ratio Test to Evaluate Expression-Trait Association

We perform a likelihood ratio test for expression-trait association:

$$\mathcal{H}_0 : \alpha_j = 0 \qquad \mathcal{H}_a : \alpha_j \neq 0, \tag{11}$$

with the assumption of no horizontal pleiotropy. This is equivalent to testing

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

since $c_j \neq 0$. The test statistic for the j-th gene is

$$\Lambda_j = 2\left(\log\Pr\left(\hat{\boldsymbol{\gamma}}_{1j}, \hat{\boldsymbol{\gamma}}_{2j}|\hat{\mathbf{R}}_{1j}, \hat{\mathbf{R}}_{2j}; \hat{\boldsymbol{\theta}}^{\mathrm{ML}}\right) - \log\Pr\left(\hat{\boldsymbol{\gamma}}_{1j}, \hat{\boldsymbol{\gamma}}_{2j}|\hat{\mathbf{R}}_{1j}, \hat{\mathbf{R}}_{2j}; \hat{\boldsymbol{\theta}}_0^{\mathrm{ML}}\right)\right), \tag{13}$$

where $\hat{\boldsymbol{\theta}}_0^{\mathrm{ML}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{ML}}$ are parameter estimates obtained by maximizing the marginal likelihood under $\mathcal{H}_0$ and $\mathcal{H}_0 \cup \mathcal{H}_a$, respectively. The test statistic asymptotically follows the $\chi^2_{\mathrm{df}=1}$ under the null hypothesis (Van der Vaart, 2000), and the calculation of the marginal log-likelihood is detailed in the **Supplementary Material**. In practice, horizontal pleiotropy may be present, and the null hypothesis for CoMM-S^4 becomes "there is no expression-trait effect and no horizontal pleitoropy." As with other TWAS methods, horizontal pleiotropy could produce significant associations and inflation of test statistics (Gusev et al., 2016; Barbeira et al., 2018).

3 RESULTS

3.1 Simulation Studies

In the simulation studies, we primarily focus on a) comparing the likelihood ratio test statistics from CoMM-S^4 and CoMM-S^2, b) assessing the type-I error of CoMM-S^4 under the null hypothesis ($h_T^2 = 0$), and c) comparing the power of CoMM-S^4, CoMM-S^2 and S-PrediXcan.

3.1.1 Simulation Settings

When comparing the test statistic and type-I error of CoMM-S^4 with CoMM-S^2, the sample sizes of the eQTL and GWAS datasets are n_1 = 5, 000 and n_2 = 5, 000 respectively. In the power comparison with CoMM-S^2 and S-PrediXcan, the sample sizes are n_1 = 500 and n_2 = 10, 000 respectively. For all simulation scenarios, the sample size of the reference panels for the eQTL and GWAS datasets are n_3 = 400 and n_4 = 400 respectively.

A multivariate normal distribution with the covariance structure $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\rho))$ is used to generate a prototype of the genotype matrix, where the parameter $\rho \in \{0.2, 0.5, 0.8\}$ determines the strength of correlations among the SNPs. Subsequently, minor allele frequencies are generated from the uniform distribution $\mathcal{U}(0.05, 0.5)$. At each SNP position, the probability that an individual has 0, 1 or 2 minor alleles is calculated using the minor allele frequencies, assuming Hardy-Weinberg Equilibrium; individuals are assigned genotype values such that the desired genotype probabilities and minor allele frequencies are achieved.

We generate gene expression values according to $\mathbf{y}_j = \mathbf{W}_{1j}\boldsymbol{\gamma}_j + \mathbf{e}_1$, where $\mathbf{e}_1 \sim \mathcal{N}(0, \sigma^2_{e_1}\mathbf{I}_{n_1})$. The effect sizes γ_{jk} are generated from $\mathcal{N}(\mathbf{0}, \sigma^2_{\gamma_j})$ with probability π and set to 0 with probability $1 - \pi$, where π denotes the sparsity level and k indexes the genetic variants within a pre-defined window of gene j. To simulate distinct scenarios, we choose equally-spaced cellular heritability levels (h_C^2) of 0.01, 0.03, 0.05, 0.07, and 0.09, and sparsity levels of 0.1, 0.2, 0.3, 0.4, 0.5, and 1. Complex traits are generated according to $\mathbf{z} = \alpha_j\mathbf{W}_{2j}\boldsymbol{\gamma}_j + \mathbf{e}_2$ and the number of *cis*-SNPs is set to 100. The trait level heritability (h_T^2) is set to 0 under the null hypothesis and 0.001, 0.002, and 0.003 under the alternative hypothesis.

The corresponding summary statistics were generated by applying a simple linear regression to the individual-level eQTL and GWAS datasets. Further details on the simulation procedure are in the **Supplementary Material**.

3.1.2 Simulation Results

There is a high concordance between the likelihood ratio test statistics from CoMM-S^4 and CoMM-S^2, which suggests that eQTL summary statistics can generally provide comparable power as individual-level data. In the scatter plot of CoMM-S^4 and CoMM-S^2 test statistics, the R^2 value is greater than 80% and the simple linear regression slope ranges from 0.88 to 1 (**Figure 1** and **Supplementay Figures S1–S6**). Moreover, the QQ plots indicate that the observed p-values from CoMM-S^4 are close to the expected p-values under the null hypothesis of no expression-trait association (**Figure 2**, **Supplementary Figures S7–S9**), indicating good type-I error control.

The power of CoMM-S^4, CoMM-S^2 and S-PrediXcan is also evaluated in the following scenarios: i) the eQTL and GWAS populations have the same LD structure, ii) the eQTL and GWAS populations have different LD structures, and iii) the eQTL and GWAS populations have different LD structures and different gene expression architectures, i.e. the set of *cis*-SNPs for the two populations only partially overlap (**Figure 3**, **Supplementary Figures S10–S14**; simulation details in **Supplementary Material**).

Across the scenarios considered, the greatest gains in power were observed when the cellular heritability is low ($h_C^2 = 0.01$) and the trait heritability is high ($h_T^2 = 0.003$). When the eQTL and GWAS samples are drawn from the same population, there is 71% power for CoMM-S^4, compared with 30 and 16% power for S-PrediXcan (ridge) and S-PrediXcan (elastic net), respectively (sparsity = 0.1; **Figure 3**). When the eQTL and GWAS samples have distinct LD structures, there is 76% power for CoMM-S^4, compared with 38 and 15% power for S-PrediXcan (ridge) and S-PrediXcan (elastic net), respectively (sparsity = 0.1; Figure S13). When the eQTL and GWAS samples have distinct LD structures and different gene expression architectures, there is 67% power for CoMM-S^4, compared with 21 and 10% power for S-PrediXcan (ridge) and S-PrediXcan (elastic net), respectively (sparsity = 0.1; **Supplementary Figure S14**).

When the cellular heritability is large ($h_C^2 = 0.09$) and the gene expression architecture is the same in both the eQTL and GWAS datasets, the power of CoMM-S^4 is comparable to S-PrediXcan (**Figure 3**; **Supplementary Figures S10–S13**). However, when the eQTL and GWAS samples have distinct LD structures and different gene expression architectures, CoMM-S^4 shows some improvement in power over S-PrediXcan: there is 61% power for CoMM-S^4, compared with 39 and 48% power for S-PrediXcan (ridge) and S-PrediXcan (elastic net), respectively ($h_T^2 = 0.003$, sparsity = 0.1; **Supplementary Figure S14**).

3.2 Real Data Analysis

3.2.1 NFBC1966 Cohort

In the real data analysis, we apply CoMM-S^4 to the NFBC1966 dataset (Sabatti et al., 2009). The NFBC1966 dataset contains phenotype data for the following ten traits: body mass index (BMI), systolic blood pressure (SysBP), diastolic blood pressure (DiaBP), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG), total cholesterol (TC), insulin levels, glucose levels and C-reactive protein (CRP). The summary statistics were generated by applying simple linear regression to individual-

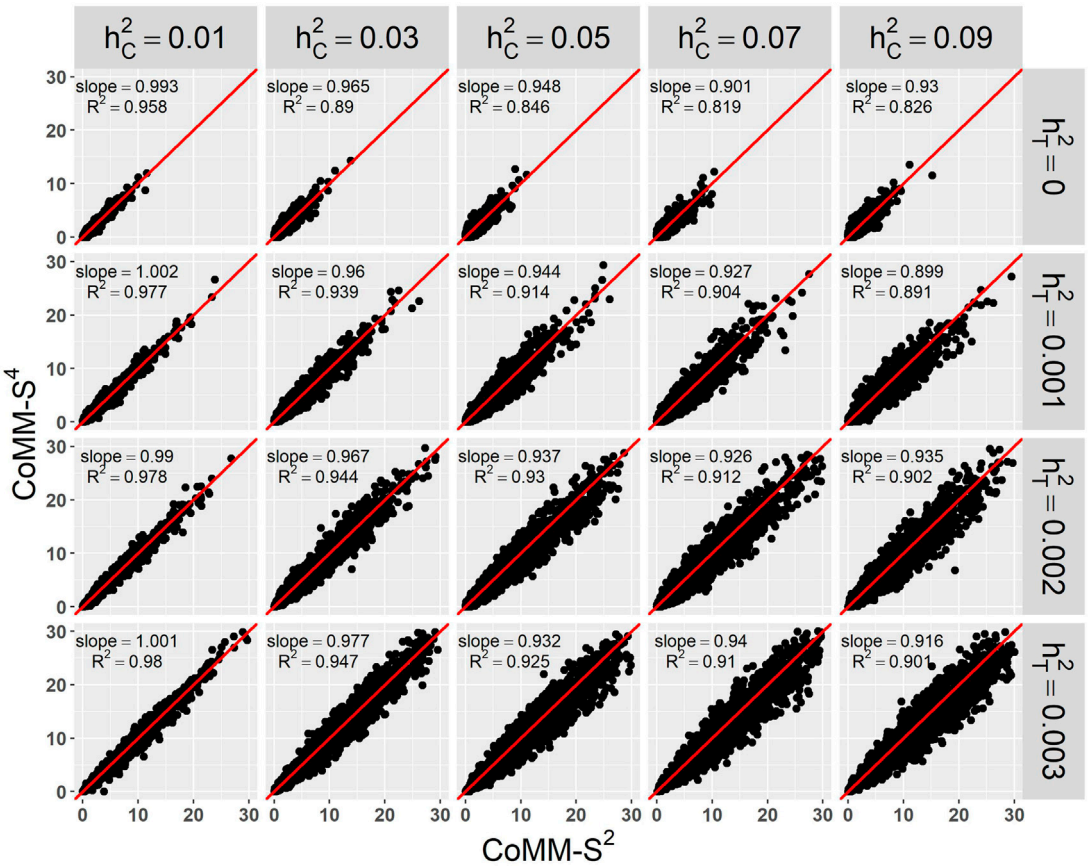


FIGURE 1 | The scatter plot of CoMM-S^4 vs. CoMM-S^2, the model setting is n_1 = 5,000, n_2 = 5,000, n_3 = 400, n_4 = 400, m_j = 100, ρ = 0.5, π = 0.2, the number of replication is 2,000.

level NFBC1966 using plink (Purcell et al., 2007). Summary statistics of *cis*-eQTLs from eQTLGen Consortium (Võsa et al., 2018) were used. In addition, linkage disequilibrium for the eQTL and GWAS datasets was estimated using the 1,000 Genomes dataset (The 1000 Genomes Project Consortium, 2015) and 400 NFBC subsamples, respectively.

The genomic inflation factor is between 0.91 and 1.09, and the number of significant genes (p-value $< 5 \times 10^{-6}$) identified by CoMM-S^4 is between 0 and 64 (**Table 1**). For the trait HDL, CoMM-S^4 identified 61 genes, of which 20 are reported to be associated with HDL in NHGRI-EBI GWAS Catalog (Buniello et al., 2018). For the trait LDL, CoMM-S^4 detected 64 genes, of which 13 are reported in GWAS Catalog. The corresponding QQ plots for these ten traits are illustrated in **Supplementary Figure S15**.

3.2.2 Biobank Japan

We apply CoMM-S^4 to GWAS summary statistics from Biobank Japan (BBJ) (Ishigaki et al., 2020). We considered two autoimmune traits (Graves' disease, rheumatoid arthritis), four cardiovascular traits (cerebral aneurysm, congestive heart failure, ischemic stroke, peripheral artery disease), two infection-related traits (chronic hepatitis B, chronic hepatitis C) and osteoporosis. The TWAS analysis is performed using whole-blood *cis*-eQTL summary statistics from two studies, eQTLGen (Võsa et al., 2018) and GTEx (v8) (The GTEx Consortium, 2020), to assess the robustness of TWAS results to choice of eQTL dataset. The GTEx and eQTLGen datasets contain association results for 19,599 and 19,176 genes respectively, of which 16,692 genes are in common. Linkage disequilibrium corresponding to the GWAS and eQTL datasets were estimated using Japanese and European samples from the 1,000 Genomes Project (The 1000 Genomes Project Consortium, 2015), respectively. As population differences in eQTL architecture may reduce gene expression imputation accuracy for the GWAS samples, it is preferable for the eQTL and GWAS data to be collected from the same population (Keys et al., 2020). However, the availability of highly-powered eQTL studies may be limited for the population of interest. Moreover, populations that are closely related still provide good power to detect associations between gene expression and trait (Keys et al., 2020), and the relatively high concordance rate (68.8%) of *cis*-regulation in European and Japanese eQTL studies (Narahara et al., 2014) suggest that European eQTL studies could serve as a reasonable proxy.

TWAS was performed to find genetic loci that may be associated with the traits of interest. For traits where TWAS

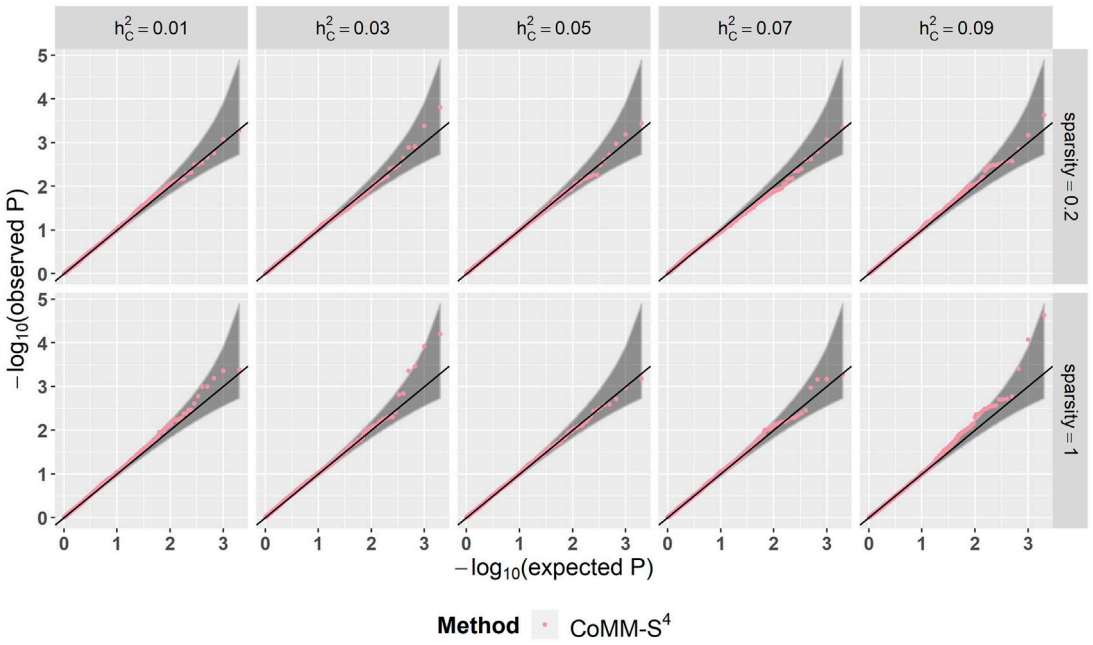


FIGURE 2 | The QQ plot of CoMM-S^4, the model setting is n_1 = 5,000, n_2 = 5,000, n_3 = 400, n_4 = 400, ρ = 0.5, the number of replication is 2,000.

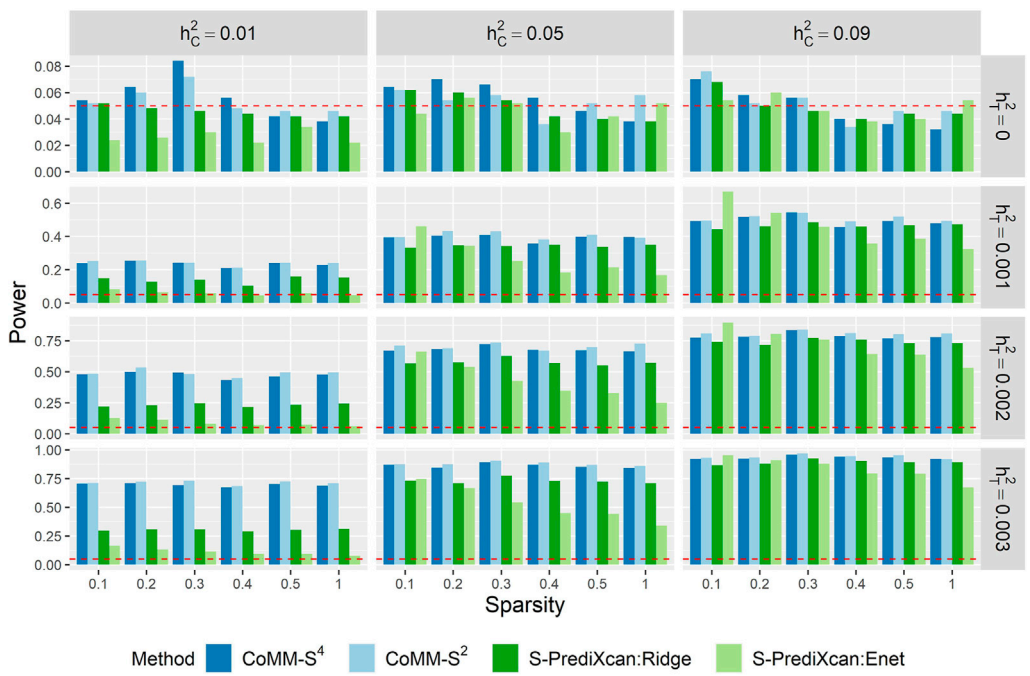


FIGURE 3 | The empirical type I error (h_T^2 = 0) and power ($h_T^2 > 0$) of CoMM-S^4, CoMM-S^2, S-PrediXcan (ridge) and S-PrediXcan (elastic net) across 500 replications. The model setting is n_1 = 500, n_2 = 10,000, n_3 = 400, n_4 = 400, ρ = 0.5.

identified more than 100 statistically significant genes, we further carried out an enrichment analysis based on gene ontology (GO) terms using Enrichr (Chen et al., 2013). The genomic inflation factor is between 1.06 and 1.30 when eQTL summary statistics were obtained from eQTLGen, and between 0.87 and 1.26 when eQTL summary statistics were obtained from GTEx (v8) whole blood (**Table 2**). The number of identified genes (p-value < 5×10^{-6}) ranged from 2 to 450, and there is a high degree of overlap

TABLE 1 | The genomic inflation factor (GIF) and the number of associated genes (p-value $<5 \times 10^{-6}$) found by CoMM-S^4 for the ten NFBC traits. The number within the parentheses is the number of associated genes reported in the NHGRI-EBI GWAS Catalog (Buniello et al., 2018).

	GIF	No. of associated genes (reported in GWAS Catalog)
CRP	0.94	25 (5)
Glucose	0.99	4 (1)
Insulin	0.86	1 (0)
TC	1.06	26 (4)
HDL	1.09	61 (20)
LDL	1.09	64 (13)
TG	1.06	2 (0)
BMI	0.98	3 (1)
SysBP	1.05	0 (0)
DiaBP	0.91	0 (0)

between the genes identified in the two analyses (**Table 2**), indicating robustness to eQTL dataset choice. Moreover, around half or more of the genes identified by CoMM-S^4 have not been previously reported as significant in the GWAS Catalog or the Biobank GWAS analysis (**Table 2**).

The TWAS results recapitulate known or proposed biological mechanisms that give rise to the studied traits. GWAS and animal model studies have implicated MHC molecules, interferon-gamma signalling and apoptosis in the development of Graves' disease (Morshed and Davies, 2015; Okada et al., 2015; Smith and Hegedüs, 2016), rheumatoid arthritis (Castañeda-Delgado et al., 2017; Okada et al., 2019), and chronic hepatitis B infection (Ebert et al., 2015; Zhu et al., 2016). Pathway enrichment analyses recapitulate these findings. For Graves' disease, 143 of the 245 associated genes (TWAS p-value $< 5 \times 10^{-8}$) are involved in GO biological processes, and the 23 significantly enriched processes (FDR <0.05, **Supplementary Table S3**) include interferon-gamma-mediated signaling pathway ($p = 6.86 \times 10^{-10}$), as well as antigen processing and presentation of peptide antigen via MHC class I ($p = 3.14 \times 10^{-8}$) and via MHC class II ($p = 2.76 \times 10^{-6}$). For rheumatoid arthritis, 137 of the 220 associated genes are involved in GO biological processes, and the 25 significantly enriched processes (**Supplementary Table S4**) include interferon-gamma-mediated signaling pathway ($p = 1.20 \times 10^{-11}$), and antigen processing and presentation of exogenous peptide antigen via MHC class II ($p = 4.21 \times 10^{-11}$). For chronic hepatitis B, 91 of the 132 associated genes are involved in GO biological processes, and the 32 significantly enriched processes (**Supplementary Table S5**) include antigen processing and presentation of exogenous peptide antigen via MHC class II ($p = 2.28 \times 10^{-7}$) and positive regulation of apoptotic cell clearance ($p = 9.21 \times 10^{-6}$).

Moreover, CoMM-S^4 is able to identify novel susceptibility loci by aggregating the contributions of SNPs with smaller effect sizes. A comparison of the GWAS results with CoMM-S^4 results based on the highly-powered eQTLGen study shows that the TWAS signal is larger than the GWAS signal at chr17q12 for congestive heart failure (CHF), chr17p13.1 for peripheral artery disease (PAD), chr17q21.31 for ischemic stroke, and chr6q22.33 for osteoporosis (**Supplementary Figures S17–S25**). Plausible mechanisms can be identified for genes at these loci, which may serve as a stepping stone for further investigation. For CHF, the second largest signal at chr17q12 corresponds to *FBXL20* ($p = 1.33 \times 10^{-5}$), which negatively regulates autophagy (Mathiassen and Cecconi, 2017). Reduced autophagy contributes to accelerated cardiac ageing and heart failure (Nishida et al., 2009; Abdellatif et al., 2018; Dong et al., 2019), and may serve as a link between *FBXL20* and CHF. For PAD, the second largest signal at chr17p13.1 corresponds to *GABARAP* ($p = 8.63 \times 10^{-8}$), which is involved in autophagy initiation and autophagosome-lysosome fusion (Schaaf et al., 2016). Impaired autophagy aggravates atherosclerosis (De Meyer et al., 2015), and may serve as a link between *GABARAP* and PAD.

For ischemic stroke, the TWAS signal is larger than the GWAS signal at chr17q21.31. The top association corresponds to *HEXIM1* ($p = 1.07 \times 10^{-6}$), which modulates hypoxia-inducible factor-1 alpha and vascular endothelial growth factor (Ogba et al., 2010; Ketchart et al., 2013), angiogenic factors which may influence stroke risk by mediating neovascularization in atherosclerotic lesions, potentially precipitating thrombi that obstruct blood flow to the brain (Bentzon et al., 2014; Chistiakov et al., 2015; Camaré et al., 2017). For osteoporosis,

TABLE 2 | The genomic inflation factor and number of associated genes (p-value $<5 \times 10^{-6}$) for 9 traits in the Biobank Japan dataset. Two eQTL datasets were used: eQTLGen and GTEx. In parentheses are the number of associated genes that are also present in the other eQTL dataset's gene set. The last column shows the number of associated genes that are common to both the eQTLGen and GTEx analyses; in parentheses are the number of associated genes that are statistically significant in the GWAS analysis (p-value $<5 \times 10^{-8}$), and the number of associated genes reported in the GWAS Catalog.

	eQTLGen		GTEx		eQTLGen and GTEx
	GIF	No. associated genes (No. in GTEx)	GIF	No. associated genes (No. in eQTLGen)	No. common associated genes (sig. in BBJ GWAS; reported in GWAS Catalog)
Graves' disease	1.17	283 (247)	1.09	454 (364)	245 (125; 7)
Rheumatoid arthritis	1.30	266 (230)	1.26	402 (323)	220 (134; 22)
Chronic hepatitis B	1.06	148 (133)	0.87	211 (172)	132 (70; 6)
Chronic hepatitis C	1.09	73 (66)	1.00	163 (145)	64 (4; 1)
Ischemic stroke	1.25	23 (21)	1.24	60 (56)	19 (3; 3)
Congestive heart failure	1.18	4 (2)	1.13	10 (9)	1 (0; 0)
Peripheral artery disease	1.13	13 (10)	0.99	45 (37)	7 (0; 0)
Cerebral aneurysm	1.11	4 (4)	0.99	6 (6)	2 (0; 0)
Osteoporosis	1.07	2 (2)	0.93	7 (6)	1 (0; 0)

the TWAS signal is larger than the GWAS signal at chr6q22.33. The top association corresponds to *RNF146* ($p = 1.05 \times 10^{-8}$), which was shown to promote osteoblast development while antagonizing osteoclast differentiation in mice (Matsumoto et al., 2017). Notably, none of the genes described above are reported as significant in the GWAS Catalog or the Biobank Japan GWAS analysis, thus highlighting the potential utility of applying CoMM-S^4 to identify relevant genes.

On the other hand, the TWAS results are limited by the data availability in the eQTL dataset. Although the TWAS results recapitulate most GWAS results, the Manhattan plots also show some GWAS signals without corresponding TWAS signals (**Supplementary Figures S17–S25**), in part due to the relative sparsity of genes in the eQTL dataset. A further limitation is that TWAS provide information about association, rather than causality. In the present analysis, *TMEM184C* and *PRMT10* showed significant association with cerebral aneurysm. However, a previous report has indicated that these are not the causal genes. Instead, the likely causal gene is *EDNRA*, which is in the same locus as *TMEM184C* and *PRMT10* and regulates response to hemodynamic stress (Low et al., 2012). As *EDNRA* is not present in any of the eQTL datasets, it could not be evaluated in this analysis.

In addition, we compare the CoMM-S^4 results with S-PrediXcan (elastic net) results for the 9 Biobank Japan traits. For S-PrediXcan, gene expression prediction weights for GTEx (v8) whole blood were obtained from the elastic net model in PredictDB (http://predictdb.org/), and the covariance matrix used to calculate the test statistics is based on Japanese samples from the 1,000 Genomes Project. To allow for fair comparison, we consider only genes that are common to both the CoMM-S^4 and S-PrediXcan analyses. Compared with S-PrediXcan (elastic net), CoMM-S^4 identifies a similar number of statistically significant genes for 5 Biobank Japan traits (cerebral aneurysm, congestive heart failure, ischemic stroke, peripheral artery disease, and osteoporosis), and more statistically significant genes for 4 Biobank Japan traits (Graves' disease, rheumatoid arthritis, chronic hepatitis C, and chronic hepatitis B) (**Supplementary Table S2**). The tail behaviour in the QQ plots indicate that the *p*-values tend to be smaller for statistically significant genes (**Supplementary Figure S16**). The higher number of identified genes in the Biobank Japan traits is consistent with the higher power demonstrated in simulations.

4 DISCUSSION

In this article, we have developed a collaborative mixed model using both summary statistics from eQTL and GWAS to examine the expression-trait associations in transcriptome-wide association studies. We compared the performance between CoMM-S^4 and CoMM-S^2, and simulation results demonstrate that CoMM-S^4 performs as well as CoMM-S^2 even though the former uses only summary-level data. Moreover, our analysis of the NFBC1966 cohort has suggested novel susceptibility loci for glucose levels, insulin levels, C-reactive protein, BMI and lipid traits. Our analysis of Biobank Japan traits has similarly suggested novel susceptibility loci for congestive heart failure, ischemic stroke, peripheral artery disease and osteoporosis, and has also recapitulated known and putative mechanisms for Graves' disease, rheumatoid arthritis, chronic hepatitis B and chronic hepatitis C.

CoMM-S^4 has several advantages over CoMM-S^2 and S-PrediXcan. Compared to stage-wise methods like S-PrediXcan, CoMM-S^4 accounts for imputation uncertainty, which makes it statistically more powerful in identifying expression-trait associations. Moreover, CoMM-S^4 requires only summary-level data (z-scores) from eQTL studies, instead of individual-level data. This allows us to make use of eQTL large-scale studies and meta-analyses where individual-level data may be unavailable.

On the other hand, likelihood-ratio tests are less computationally efficient than score-based tests; the relationship between these tests in the context of individual-level data (CoMM and SKAT, respectively) are discussed in detail in (Yang et al., 2018). To reduce the computational time of CoMM-S^4, we have estimated the parameters using variational inference and parameter expansion. Finally, CoMM-S^4, like S-PrediXcan, is not able to distinguish between causal relationship and horizontal pleiotropy. In practice, we can first perform a TWAS to identify regions that show association with the trait of interest, and then apply Mendelian randomization analysis to draw causal conclusions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. Biobank Japan GWAS summary statistics were obtained from http://jenger.riken.jp/en/result; eQTLGen summary statistics were obtained from https://www.eqtlgen.org/; GTEx eQTL summary statistics were obtained from https://www.ebi.ac.uk/eqtl/Studies/.

AUTHOR CONTRIBUTIONS

JL conceived and supervised the study. YY developed the algorithm. KY and YY performed the data analyses and wrote the manuscript with input from JL. All authors have reviewed and approved the final manuscript.

FUNDING

This work was supported by grant R-913-200-098-263 from the Duke-NUS Graduate Medical School, and AcRF Tier 2 (MOE2018-T2-1-046 and MOE2018-T2-2-006) from the Ministry of Education, Singapore. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.704538/full#supplementary-material

REFERENCES

Abdellatif, M., Sedej, S., Carmona-Gutierrez, D., Madeo, F., and Kroemer, G. (2018). Autophagy in Cardiovascular Aging. *Circ. Res.* 123, 803–824. doi:10.1161/circresaha.118.312208

Atkins, I., Kinnersley, B., Ostrom, Q. T., Labreche, K., Il'yasova, D., Armstrong, G. N., et al. (2019). Transcriptome-Wide Association Study Identifies New Candidate Susceptibility Genes for Glioma. *Cancer Res.* 79, 2065–2071. doi:10.1158/0008-5472.can-18-2888

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., et al. (2018). Exploring the Phenotypic Consequences of Tissue Specific Gene Expression Variation Inferred From Gwas Summary Statistics. *Nat. Commun.* 9, 1825. doi:10.1038/s41467-018-03621-1

Bentzon, J. F., Otsuka, F., Virmani, R., and Falk, E. (2014). Mechanisms of Plaque Formation and Rupture. *Circ. Res.* 114, 1852–1866. doi:10.1161/circresaha.114.302721

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2018). The Nhgri-Ebi Gwas Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi:10.1093/nar/gky1120

Camaré, C., Pucelle, M., Nègre-Salvayre, A., and Salvayre, R. (2017). Angiogenesis in the Atherosclerotic Plaque. *Redox Biol.* 12, 18–34. doi:10.1016/j.redox.2017.01.007

Castañeda-Delgado, J. E., Bastián-Hernandez, Y., Macias-Segura, N., Santiago-Algarra, D., Castillo-Ortiz, J. D., Alemán-Navarro, A. L., et al. (2017). Type I Interferon Gene Response Is Increased in Early and Established Rheumatoid Arthritis and Correlates With Autoantibody Production. *Front. Immunol.* 8, 285. doi:10.3389/fimmu.2017.00285

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and Collaborative Html5 Gene List Enrichment Analysis Tool. *BMC bioinformatics.* 14, 1–14. doi:10.1186/1471-2105-14-128

Chistiakov, D. A., Orekhov, A. N., and Bobryshev, Y. V. (2015). Contribution of Neovascularization and Intraplaque Haemorrhage to Atherosclerotic Plaque Progression and Instability. *Acta Physiol.* 213, 539–553. doi:10.1111/apha.12438

De Meyer, G. R., Grootaert, M. O., Michiels, C. F., Kurdi, A., Schrijvers, D. M., and Martinet, W. (2015). Autophagy in Vascular Disease. *Circ. Res.* 116, 468–479. doi:10.1161/circresaha.116.303804

Dong, Y., Chen, H., Gao, J., Liu, Y., Li, J., and Wang, J. (2019). Molecular Machinery and Interplay of Apoptosis and Autophagy in Coronary Heart Disease. *J. Mol. Cell. Cardiol.* 136, 27–41. doi:10.1016/j.yjmcc.2019.09.001

Ebert, G., Preston, S., Allison, C., Cooney, J., Toe, J. G., Stutz, M. D., et al. (2015). Cellular Inhibitor of Apoptosis Proteins Prevent Clearance of Hepatitis B Virus. *Proc. Natl. Acad. Sci.* 112, 5797–5802. doi:10.1073/pnas.1502390112

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A Gene-Based Association Method for Mapping Traits Using Reference Transcriptome Data. *Nat. Genet.* 47, 1091. doi:10.1038/ng.3367

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., et al. (2016). Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies. *Nat. Genet.* 48, 245. doi:10.1038/ng.3506

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., et al. (2018). Transcriptome-Wide Association Study of Schizophrenia and Chromatin Activity Yields Mechanistic Disease Insights. *Nat. Genet.* 50, 538–548. doi:10.1038/s41588-018-0092-1

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying Causal Variants at Loci With Multiple Signals of Association. *Genetics.* 198, 497–508. doi:10.1534/genetics.114.167908

Huang, J., Jiao, Y., Liu, J., and Yang, C. (2021). Remi: Regression With Marginal Information and its Application in Genome-Wide Association Studies. *Stat. Sin.* 31, 1–20. doi:10.5705/ss.202019.018

Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., et al. (2020). Large-Scale Genome-Wide Association Study in a Japanese Population Identifies Novel Susceptibility Loci across Different Diseases. *Nat. Genet.* 52, 669–679. doi:10.1038/s41588-020-0640-3

Ketchart, W., Smith, K. M., Krupka, T., Wittmann, B. M., Hu, Y., Rayman, P. A., et al. (2013). Inhibition of Metastasis by Hexim1 Through Effects on Cell Invasion and Angiogenesis. *Oncogene.* 32, 3829–3839. doi:10.1038/onc.2012.405

Keys, K. L., Mak, A. C., White, M. J., Eckalbar, W. L., Dahl, A. W., Mefford, J., et al. (2020). On the Cross-Population Generalizability of Gene Expression Prediction Models. *PLoS Genet.* 16, e1008927. doi:10.1371/journal.pgen.1008927

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter Expansion to Accelerate Em: the Px-Em Algorithm. *Biometrika.* 85, 755–770. doi:10.1093/biomet/85.4.755

Low, S.-K., Takahashi, A., Cha, P.-C., Zembutsu, H., Kamatani, N., Kubo, M., et al. (2012). Genome-Wide Association Study for Intracranial Aneurysm in the Japanese Population Identifies Three Candidate Susceptible Loci and a Functional Genetic Variant at Ednra. *Hum. Mol. Genet.* 21, 2102–2110. doi:10.1093/hmg/dds020

Mancuso, N., Gayther, S., Gusev, A., Zheng, W., Penney, K. L., Kote-Jarai, Z., et al. (2018). Large-Scale Transcriptome-Wide Association Study Identifies New Prostate Cancer Risk Regions. *Nat. Commun.* 9, 1–11. doi:10.1038/s41467-018-06302-1

Mathiassen, D., and Cecconi, F. (2017). Autophagy and the Cell Cycle: a Complex Landscape. *Front. Oncol.* 7, 51. doi:10.3389/fonc.2017.00051

Matsumoto, Y., La Rose, J., Lim, M., Adissu, H. A., Law, N., Mao, X., et al. (2017). Ubiquitin Ligase Rnf146 Coordinates Bone Dynamics and Energy Metabolism. *J. Clin. Invest.* 127, 2612–2625. doi:10.1172/jci92233

Morshed, S. A., and Davies, T. F. (2015). Graves' Disease Mechanisms: The Role of Stimulating, Blocking, and Cleavage Region Tsh Receptor Antibodies. *Horm. Metab. Res.* 47, 727–734. doi:10.1055/s-0035-1559633

Narahara, M., Higasa, K., Nakamura, S., Tabara, Y., Kawaguchi, T., Ishii, M., et al. (2014). Large-Scale East-Asian Eqtl Mapping Reveals Novel Candidate Genes for Ld Mapping and the Genomic Landscape of Transcriptional Effects of Sequence Variants. *PloS one.* 9, e100924. doi:10.1371/journal.pone.0100924

Nishida, K., Kyoi, S., Yamaguchi, O., Sadoshima, J., and Otsu, K. (2009). The Role of Autophagy in the Heart. *Cel Death Differ.* 16, 31–37. doi:10.1038/cdd.2008.163

Ogba, N., Doughman, Y. Q., Chaplin, L. J., Hu, Y., Gargesha, M., Watanabe, M., et al. (2010). Hexim1 Modulates Vascular Endothelial Growth Factor Expression and Function in Breast Epithelial Cells and Mammary Gland. *Oncogene.* 29, 3639–3649. doi:10.1038/onc.2010.110

Okada, Y., Eyre, S., Suzuki, A., Kochi, Y., and Yamamoto, K. (2019). Genetics of Rheumatoid Arthritis: 2018 Status. *Ann. Rheum. Dis.* 78, 446–453. doi:10.1136/annrheumdis-2018-213678

Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., Matsuda, K., Kamatani, Y., et al. (2015). Construction of a Population-Specific Hla Imputation Reference Panel and its Application to Graves' Disease Risk in Japanese. *Nat. Genet.* 47, 798–802. doi:10.1038/ng.3310

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., et al. (2009). Genome-Wide Association Analysis of Metabolic Traits in a Birth Cohort From a Founder Population. *Nat. Genet.* 41, 35. doi:10.1038/ng.271

Schaaf, M. B., Keulers, T. G., Vooijs, M. A., and Rouschop, K. M. (2016). Lc3/Gabarap Family Proteins: Autophagy-(un)Related Functions. *FASEB J.* 30, 3961–3978. doi:10.1096/fj.201600698r

Smith, T. J., and Hegedüs, L. (2016). Graves' Disease. *New Engl. J. Med.* 375, 1552–1565. doi:10.1056/nejmra1510030

Strunz, T., Lauwen, S., Kiel, C., den Hollander, A., and Weber, B. H. (2020). A Transcriptome-Wide Association Study Based on 27 Tissues Identifies 106 Genes Potentially Relevant for Disease Pathology in Age-Related Macular Degeneration. *Scientific Rep.* 10, 1–16. doi:10.1038/s41598-020-58510-9

The 1000 Genomes Project Consortium (2015). A Global Reference for Human Genetic Variation. *Nature.* 526, 68–74. doi:10.1038/nature15393

The GTEx Consortium (2020). The Gtex Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science.* 369, 1318–1330. doi:10.1126/science.aaz1776

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. New York: Cambridge University Press, Vol. 3.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., et al. (2018). Unraveling the Polygenic Architecture of Complex Traits Using Blood Eqtl Meta-Analysis. *bioRxiv.*, 447367.

Williams, R. B., Chan, E. K., Cowley, M. J., and Little, P. F. (2007). The Influence of Genetic Variation on Gene Expression. *Genome Res.* 17, 1707–1716. doi:10.1101/gr.6981507

Yang, C., Wan, X., Lin, X., Chen, M., Zhou, X., and Liu, J. (2018). CoMM: a Collaborative Mixed Model to Dissecting Genetic Contributions to Complex Traits by Leveraging Regulatory Information. *Bioinformatics.* 35, 1644. doi:10.1093/bioinformatics/bty865

Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., et al. (2020). CoMM-S2: a Collaborative Mixed Model Using Summary Statistics in Transcriptome-Wide Association Studies. *Bioinformatics.* 36, 2009–2016. doi:10.1093/bioinformatics/btz880

Zhu, M., Dai, J., Wang, C., Wang, Y., Qin, N., Ma, H., et al. (2016). Fine Mapping the Mhc Region Identified Four Independent Variants Modifying Susceptibility to Chronic Hepatitis B in Han Chinese. *Hum. Mol. Genet.* 25, 1225–1232. doi:10.1093/hmg/ddw003

Zhu, X., and Stephens, M. (2017). Bayesian Large-Scale Multiple Regression With Summary Statistics from Genome-Wide Association Studies. *Ann. Appl. Stat.* 11, 1561. doi:10.1214/17-aoas1046

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 17 November 2021
doi: 10.3389/fgene.2021.690926

LORSEN: Fast and Efficient eQTL Mapping With Low Rank Penalized Regression

Cheng Gao[1], *Hairong Wei*[2] *and Kui Zhang*[1]*

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, United States, [2]College of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI, United States

Characterization of genetic variations that are associated with gene expression levels is essential to understand cellular mechanisms that underline human complex traits. Expression quantitative trait loci (eQTL) mapping attempts to identify genetic variants, such as single nucleotide polymorphisms (SNPs), that affect the expression of one or more genes. With the availability of a large volume of gene expression data, it is necessary and important to develop fast and efficient statistical and computational methods to perform eQTL mapping for such large scale data. In this paper, we proposed a new method, the low rank penalized regression method (LORSEN), for eQTL mapping. We evaluated and compared the performance of LORSEN with two existing methods for eQTL mapping using extensive simulations as well as real data from the HapMap3 project. Simulation studies showed that our method outperformed two commonly used methods for eQTL mapping, LORS and FastLORS, in many scenarios in terms of area under the curve (AUC). We illustrated the usefulness of our method by applying it to SNP variants data and gene expression levels on four chromosomes from the HapMap3 Project.

Keywords: eQTL mapping, proximal gradient method, cis-eQTL, *trans*-eQTL, penalized regression

OPEN ACCESS

Edited by:
Qi Yan,
Columbia University, United States

Reviewed by:
Rong Zhang,
Amgen , United States
Chi-Yang Chiu,
University of Tennessee Health Science Center (UTHSC), United States

***Correspondence:**
Kui Zhang
kuiz@mtu.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 April 2021*
Accepted: *08 October 2021*
Published: *17 November 2021*

Citation:
Gao C, Wei H and Zhang K (2021) LORSEN: Fast and Efficient eQTL Mapping With Low Rank Penalized Regression. Front. Genet. 12:690926. doi: 10.3389/fgene.2021.690926

1 INTRODUCTION

With rapid advancements in sequencing technologies and high-throughput technologies, a large number of single nucleotide polymorphism (SNP) data and gene expression data have become available. This allows us to investigate the associations between SNP genotypes and gene expression levels. Expression quantitative trait loci (eQTLs) are those genetic variants that can explain variation in gene expression levels and can help to elucidate the underlying genetic mechanisms of human complex traits (Albert and Kruglyak, 2015). eQTL mapping aims to identify eQTLs associated with genes of interest (Hu et al., 2015; Banerjee et al., 2021). In general, eQTLs are classified into two types: *cis*-eQTLs (or local eQTLs) and *trans*-eQTLs (or distant eQTLs) (Cookson et al., 2009). *cis*-eQTLs refer to the genetic variants that functionally act on local genes and are physically located close to the target genes. *trans*-eQTLs are those genetic variants that functionally act on distant genes residing on the same or different chromosome and are physically located far from the target genes. It is worth mentioning that *trans*-eQTLs account for a large proportion of heritability of gene expression levels, though *trans* effects are usually weaker than *cis* effects in humans (Cookson et al., 2009).

In fact, gene expression levels observed are not only regulated by genetic variants but also influenced by non-genetic factors which are known or hidden, for example, batch effects. Therefore, in eQTL mapping, how to account for confounding factors is an important issue and can influence the detection power of eQTL mapping. Up to now, a number of methods have been proposed to

account for confounding factors in eQTL mapping, for example, PANAMA (Fusi et al., 2012), PEER (Stegle et al., 2010), LORS (Yang et al., 2013), HEFT (Gao et al., 2014), LMM-EH-PS (Listgarten et al., 2010) and ECCO (Yue et al., 2020). Another challenge in eQTL mapping is that the number of SNPs involved is usually very large (Yang et al., 2013). This not only results in heavy computational burden for estimating model parameters but also generally results in reduced detection power if all SNPs are included in eQTL mapping. This is because the signal-to-noise ratio (SNR) is very low, meaning only a very small portion of SNPs that are actually associated with gene expression levels. To overcome this problem, a number of SNP screening procedures (Wang et al., 2011; Yang et al., 2013; Jeng et al., 2020) and variable selection techniques (Fan and Lv, 2008) that aim to reduce the number of SNPs and only keep informative SNPs in eQTL mapping have been developed. More importantly, a number of methods based on the penalized regression have been developed to model such sparsity of eQTLs (Lee and Xing, 2012; Yang et al., 2013; Cheng et al., 2014; Jeng et al., 2020).

LORS, a method based on the low rank sparse regression, was proposed for eQTL mapping in (Yang et al., 2013). LORS is based on a linear model with gene expression levels as response variables and SNP genotypes as predictors. To model the sparsity of regression coefficients, LORS poses the L_1 penalty on the regression coefficient matrix. In addition, LORS includes one unknown matrix with the nuclear norm penalty to account for variations caused by non-genetic factors. Yang et al. (2013) applied the coordinate descent algorithm to optimize the objective function and estimate the model parameters. A SNP screening method, called LORS-Screening, was also developed to reduce the number of SNPs involved in the subsequent joint modeling, thus reduce the computational burden greatly. Similar to LORS, FastLORS (Jeng et al., 2020) employs the same low rank sparse regression model that is used in LORS. Different from LORS, FastLORS uses generic proximal gradient algorithm to optimize the objective function and estimate the model parameters. Moreover, Jeng et al. (2020) proposed a SNP screening method based on the Higher Criticism (HC) statistic, called HC-Screening.

To improve the detection power of eQTL mapping, a number of methods have been developed to incorporate the structure information from SNP variants data and gene expression levels, for example, clustering based on gene expression levels (Kendziorski et al., 2006; Chun and Keles, 2009) and gene regulatory networks (Rakitsch and Stegle, 2016), into eQTL mapping. A number of studies have shown that such structure information from SNP variants data and gene expression levels can be effectively used in penalized regression to boost the detection power of eQTL mapping (Chen et al., 2012; Kim and Xing, 2012, 2009). For example, the graph-regularized dual lasso (GDL) proposed by (Cheng et al., 2014) can simultaneously integrate the correlation structures among SNPs and gene expression levels. Through extensive experimental evaluations, Cheng et al. (2014) showed that GDL significantly outperformed the existing method for eQTL mapping. Similar to GDL, the graph-guided fused lasso (GFlasso) proposed by (Lee and Xing, 2012) can also consider the structure of the genetic variants and the structure of the gene expression levels. As a penalized regression method, GFlasso also inherits the benefits from the group lasso. Lee and Xing (2012) showed that GFlasso was able to detect weak association signals between the genetic variants and the gene expression levels.

However, there are some drawbacks for most of the aforementioned methods. First, if two SNPs are highly correlated with each other, and one SNP is associated with some genes, but the other SNP is not associated with them, we should not expect that these two SNPs have similar coefficients for those genes. Similarly, if some SNPs are classified into one group, we should not expect that the SNPs within the same group have similar coefficients for common genes. Second, the group structures of SNP data and gene expression data are usually identified by performing clustering on the data, however, clustering is an unsupervised leaning approach, the number of clusters is usually artificially determined. When we use the resulting clusters of SNPs and gene expressions to design the penalty term, it may lead to loss of detection power and even spurious associations. Third, complicated design of penalty term in penalized regression modeling can result in untractable computational bottleneck, especially when dealing with a large volume of data.

To overcome such limitations of existing methods for eQTL mapping, we proposed a novel method, LOw Rank Sparse regression with Elastic Net penalty, abbreviated as LORSEN. Different from LORS (Yang et al., 2013) and FastLORS (Jeng et al., 2020), we applied the Elastic Net penalty to the association coefficients instead of the L_1 penalty in LORSEN. In addition, we used the low rank approximation to account for non-genetic factors in LORSEN (Yang et al., 2013). There are several advantages to use the Elastic Net penalty instead of the L_1 penalty (Tibshirani, 1996). First, when the number of SNPs p is much larger than the sample size n, theoretically, the methods based on the L_1 penalty can only yield at most n non-zero coefficients. This can lead to the substantial loss of detection power in eQTL mapping since the number of samples is generally much smaller than the number of eQTLs in gene expression studies. Second, when several eQTLs are in linkage disequilibrium (LD), the methods based on the L_1 penalty can only select one of them. In theory, the Elastic Net penalty can overcome these two drawbacks. For the estimation of the model parameters in LORSEN, we developed an efficient optimization algorithm based on the proximal gradient method (Parikh and Boyd, 2014). Our algorithm allows us to perform the eQTL mapping for a large number of SNPs and genes. We evaluated and compared the performance of LORSEN with LORS and FastLORS using extensive simulation studies as well as the HapMap3 data.

2 MATERIAL AND METHODS

2.1 Model

We assume that the genotypes for p SNPs and the gene expression levels for q genes over n samples are collected. Let X denote the $n \times p$ matrix of SNP genotypes coded in an additive manner, and

Y denote the $n \times q$ matrix of gene expression levels. To model the association between SNPs and gene expressions, we can use the following multivariate linear model as proposed in (Yang et al., 2013):

$$Y = XB + L + \mathbf{1}\mu^T + e, \tag{1}$$

where B is a $p \times q$ matrix for the regression coefficients, $\mathbf{1}$ is a n-dimensional all-ones vector, μ is a q-dimensional vector for the intercepts in the regression model, e is a $n \times q$ matrix for the error terms and each element in e has a normal distribution with zero mean and variance σ^2, all e_{ij} are independent, L is a $n \times q$ matrix which is introduced to account for variations caused by non-genetic factors.

For the convenience of description, we first introduce the following notations used in this paper. For a n-dimensional vector v with the elements $v_i (i = 1, \ldots, n)$: the L_1 norm of v is defined as $\|v\|_1 = \Sigma_{i=1}^n |v_i|$ (the sum of absolute values of the elements) and the L_2 norm (also called the Euclidean norm) of v is defined as $\|v\|_2 = \sqrt{\Sigma_{i=1}^n v_i^2}$ (the squared root of the sum of squares of the elements), respectively. For a $m \times n$ matrix M with the elements $M_{ij} (i = 1, \ldots, m; j = 1, \ldots, n)$, the Frobenius norm of M is defined as $\|M\|_F = \sqrt{\Sigma_{i=1}^m \Sigma_{j=1}^n M_{ij}^2}$ (the squared root of the sum of squares of the elements); the nuclear norm $\|M\|_* = \Sigma_{i=1}^r \sigma_i$, where $\sigma_1, \cdots, \sigma_r$ are the singular values of M and r is the rank of M; and the L_1 norm of M is defined as $\|M\|_1 = \Sigma_{i=1}^m \Sigma_{j=1}^n |M_{ij}|$ (the sum of absolute values of the elements).

In this paper, we follow the same sparsity assumptions used in (Yang et al., 2013). First, we assume that there are only a small number of non-genetic factors that influence the gene expression levels globally, not locally. Second, we assume that there are only a small fraction of SNPs that influence the gene expression levels. This assumption implies that the regression coefficient matrix B is sparse. Yang et al. (2013) proposed the following LORS procedure to estimate B, L, μ by solving the optimization problem

$$\min_{B,L,\mu} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda\|B\|_1, \tag{2}$$

where ρ and λ are regularization (tuning) parameters that control the rank of L and the sparsity of B, respectively. When L and μ are fixed, the optimization problem becomes a least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) problem with respect to B. As pointed out in (Zou and Hastie, 2005), the Lasso has some limitations that affect its usefulness. First, when $n < p$ (the number of samples is smaller than the number of SNPs), the Lasso selects at most n SNPs. In the context of eQTL mapping, there are usually a small number of samples available. Even though the proportion of SNPs that are associated with the gene expression levels is small, it is highly likely that the number of SNPs associated with the gene expressions can still be larger than the number of samples. In this case, the L_1 penalty on B will fail to identify some SNPs that are associated with the gene expressions. Second, the Lasso tends to select only one variable among a group of highly correlated variables. This can be problematic in eQTL mapping. For example, if two SNPs are in high linkage disequilibrium and both of them are associated with gene expressions, only one SNP will be selected by the Lasso. Furthermore, if two SNPs are in high linkage disequilibrium and only one of them is associated with gene expressions, the selected SNP by the Lasso may not even be associated with gene expressions.

The use of the Elastic Net penalty (Zou and Hastie, 2005) instead of the L_1 penalty on B can overcome the limitations of the Lasso. Therefore, we propose the following optimization problem to estimate B, L, μ:

$$\min_{B,L,\mu} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2, \tag{3}$$

where ρ, λ_1 and λ_2 are non-negative tuning parameters. For real data sets, it is quite possible that some entries in Y are unobserved (missing). In such scenarios, the missing data will not be used in (**Eq. 3**). As used in (Yang et al., 2013), we use Ω to index the observed entries in Y. Specifically, Ω is a $n \times q$ matrix with the entry

$$\Omega_{ij} = \begin{cases} 0, & Y_{ij} \quad missing \\ 1, & otherwise. \end{cases} \tag{4}$$

Then we define the projection of a matrix A onto Ω as $\tilde{A} = P_\Omega(A) = \Omega \odot A$, where A has the same dimension as Ω and $\odot$ represents Hadamard product, that is, $\tilde{A}_{ij} = A_{ij} \times \Omega_{ij}$. Based on the observed data, the optimization problem becomes

$$\min_{B,L,\mu} \quad \frac{1}{2}\|P_\Omega(Y - XB - L - \mathbf{1}\mu^T)\|_F^2 + \rho\|L\|_* + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2. \tag{5}$$

2.2 Theory and Algorithm

To solve the optimization problem in (**Eq. 5**) efficiently, we developed a fast and efficient algorithm based on proximal gradient method (Parikh and Boyd, 2014).

We first describe the proximal gradient method for a general optimization problem

$$\min_x \quad f(x) = g(x) + h(x), \tag{6}$$

where $g(x)$ is a convex and differentiable function, $h(x)$ is a closed proper convex which means $h(x)$ is a convex function, the epigraph of $h(x)$ is closed and $h(x) < +\infty$ for at least one x and $h(x) > -\infty$ for every x. Furthermore, we assume that $\nabla g(x)$, the gradient of $g(x)$, is Lipschitz continuous with constant ℓ, which implies that $\nabla^2 g(x) \preceq \ell \mathbf{I}$. Two symmetric matrices of the same dimensions A and B have the relationship $A \preceq B$, if $B - A$ is positive semidefinite. Then we have

$$f(x) = g(x) + h(x) \leqslant g(x_0) + \langle \nabla g(x_0), x - x_0 \rangle + \frac{1}{2t}\|x - x_0\|^2 + h(x), \quad t \in (0, \frac{1}{\ell}], \tag{7}$$

where x_0 is an arbitrary point in the domain of $f(x)$ and $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors. Instead of using the optimization problem (**Eq. 6**), we focus on minimizing an upper bound of the objective function, that is,

Algorithm 1 | FISTA with constant step size.

Input: $t_L, t_B, t_\mu, \widetilde{L}_1 = L_0 \in \mathbb{R}^{n\times q}, \widetilde{B}_1 = B_0 \in \mathbb{R}^{p\times q}, \widetilde{\mu}_1 = \mu_0 \in \mathbb{R}^{q\times 1}, t_1 = 1$, the maximum number of iterations N, Ω

Output: optimal feasible solutions L^*, B^*, μ^*

for $k = 1$ to N **do**

$L_k \leftarrow \mathrm{S}_{t_L\rho}(\widetilde{L}_k - t_L\Omega \odot (X\widetilde{B}_k + \mathbf{1}\widetilde{\mu}_k^T + \widetilde{L}_k - Y))$

$B_k^1 \leftarrow \widetilde{B}_k - t_B X^T(\Omega \odot (X\widetilde{B}_k + \mathbf{1}\widetilde{\mu}_k^T + L_k - Y))$

$B_k^2 \leftarrow \mathrm{sign}(\mathrm{B}_k^1) \odot (|B_k^1| - \lambda_1 J)_+$

for $j = 1$ to q **do**

$B_k[,j] \leftarrow \{1 - \frac{\lambda_2}{max\{\|B_k^2[,j]\|_2, \lambda_2\}}\} B_k^2[,j]$

end for

$\mu_k \leftarrow \widetilde{\mu}_k - t_\mu(\Omega \odot (XB_k + \mathbf{1}\widetilde{\mu}_k^T + L_k - Y))^T\mathbf{1}$

$t_{k+1} \leftarrow (1+\sqrt{1+4t_k^2})/2$

$\widetilde{L}_{k+1} \leftarrow \mathrm{L}_k + \frac{t_k-1}{t_{k+1}}(L_k - L_{k-1})$

$\widetilde{B}_{k+1} \leftarrow \mathrm{B}_k + \frac{t_k-1}{t_{k+1}}(B_k - B_{k-1})$

$\widetilde{\mu}_{k+1} \leftarrow \mu_k + \frac{t_k-1}{t_{k+1}}(\mu_k - \mu_{k-1})$

if stopping criteria is satisfied **then**

break;

end if

end for=0

$$\min_x \quad g(x_0) + \langle \nabla g(x_0), x - x_0\rangle + \frac{1}{2t}\|x - x_0\|^2 + h(x), \quad t \in (0, \frac{1}{\ell}], \tag{8}$$

which can be interpreted as an application of majorization-minimization algorithm (Parikh and Boyd, 2014). The optimization problem in (**Eq. 8**) is equivalent to the following optimization problem:

$$\min_x \quad \frac{1}{2t}\|x - (x_0 - t\nabla g(x_0))\|^2 + h(x). \tag{9}$$

Problem (**Eq. 9**) can be solved with an iterative procedure: given the value of x at the k-th iteration, i.e., x_k, the value of x at the $k+1$-th iteration, x_{k+1} can be updated by the following formula

$$\begin{aligned} x_{k+1} &= \underset{x}{argmin}\ \frac{1}{2t}\|x - (x_k - t\nabla g(x_k))\|^2 + h(x) \\ &= Prox_{t,h}(x_k - t\nabla g(x_k)), \end{aligned}$$

where $Prox(\cdot)$ is called proximal operator. The iterative process is repeated until the stopping criterion is satisfied or the maximum number of iterations is reached.

To solve the optimization problem (**Eq. 5**), we adopted an alternating optimization approach that is similar to the method in (Yang et al., 2013). Note that in the following part, t_L, t_B, and t_μ are like t used in problem (**Eq. 9**) and correspond to the variables L, B, and μ, respectively.

First, for fixed B and μ, (**Eq. 5**) becomes

$$\min_L \quad \frac{1}{2}\|Y - XB - \mathbf{1}\mu^T - L\|_F^2 + \rho\|L\|_*. \tag{10}$$

In the setting of optimization problem (**Eq. 10**), $\frac{1}{2}\|Y - XB - \mathbf{1}\mu^T - L\|_F^2$ plays the role of $g(x)$ and $\rho\|L\|_*$ plays the role of $h(x)$ in (**Eq. 6**). By Lemma 1 (**Appendix A**), at the $k+1$-th iteration, we have

$$\begin{aligned} L_{k+1} &= Prox_{t_L,\rho\|\cdot\|_*}(L_k - t_L(XB_k + \mathbf{1}\mu_k^T + L_k - Y)) \\ &= S_{t_L\rho}(L_k - t_L(XB_k + \mathbf{1}\mu_k^T + L_k - Y)), \end{aligned}$$

where $S_{t_L\rho}(\cdot)$ is the singular value shrinkage operator (please refer to the **Appendix A**), t_L is the step size which can be constant or be determined by backtracking line search.

Second, for fixed L and μ, then (**Eq. 5**) becomes

$$\min_B \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2, \tag{11}$$

where t_B is the step size which can be constant or be determined by backtracking line search. By Lemmas 2 and 3 and Theorem 1 (**Appendix A**), we can update B_{k+1} accordingly:

$$\begin{aligned} B_{k+1}^a &= B_k - t_B X^T(XB_k + \mathbf{1}\mu_k^T + L_{k+1} - Y) \\ B_{k+1}^b &= Prox_{t_B,\lambda_1\|\cdot\|_1}(B_{k+1}^a) \\ &= sign(B_{k+1}^a) \circledcirc (|B_{k+1}^a| - \lambda_1 J)_+ \\ B_{k+1}[,j] &= Prox_{t_B,\lambda_2\|\cdot\|_2}(B_{k+1}^b[,j]) \\ &= \{1 - \frac{\lambda_2}{max\{\|B_{k+1}^b[,j]\|_2, \lambda_2\}}\} B_{k+1}^b[,j], \quad j = 1, 2, \ldots, q, \end{aligned}$$

where J is a all-ones $p \times q$ matrix, $B[, j]$ is the j-th column of matrix B and is a p-dimensional vector, $\gamma_+ = \max\{\gamma, 0\}$, the maximum of γ and 0, $|B^a_{k+1}|$, $\text{sign}(B^a_{k+1})$, and $(|B^a_{k+1}| - \lambda_1 J)_+$ are all elementwise operations.

Third, for fixed L and B, the proximal gradient method reduces to the gradient descent method with respect to μ because there is no penalty on μ. At the $k + 1$-th iteration, we have

$$\mu_{k+1} = \mu_k - t_\mu (XB_{k+1} + \mathbf{1}\mu_k^T + L_{k+1} - Y)^T \mathbf{1}.$$

To accelerate the computational speed, we used the accelerated proximal gradient method. Specifically, we applied the fast iterative shrinkage-thresholding algorithm (FISTA) (Beck and Teboulle, 2009) which keeps the simplicity of the iterative shrinkage-thresholding algorithms (ISTA) but has an improved rate $O(1/k^2)$, where k indexes the iteration. In FISTA, the step size can be constant or be determined by backtracking line search. The algorithm to solve LORSEN with FISTA is described in Algorithm 1. For simplicity, here, only the detailed algorithm with the constant step size is described, but the algorithm using the step size determined by backtracking line search is also provided in our R program (https://github.com/gaochengPRC/LORSEN).

2.3 Parameter Tuning

For parameter tuning, we mainly followed the idea described in (Yang et al., 2013). Specifically, we divided the entries of Ω into training entries and testing entries such that training entries and testing entries include roughly the same number of 1's. We define two matrices Ω_1 and Ω_2 such that they have the same dimensions as Ω, Ω_1 contains all training entries and Ω_2 contains all testing entries. Furthermore, we have $\Omega = \Omega_1 + \Omega_2$ and $\Omega_1 \circledcirc \Omega_2 = 0$. For the consistency, we re-parameterized λ_1 and λ_2 as $\lambda \cdot \alpha$ and $\lambda \cdot (1 - \alpha)$, respectively. So the optimization problem (**Eq. 5**) becomes

$$\min_{B,L,\mu} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda\left(\alpha\|B\|_1 + \frac{1-\alpha}{2}\|B\|_F^2\right). \tag{12}$$

This form is the same as that in glmnet (Friedman et al., 2010).

Given the values of parameters (ρ, α, λ), we solve the following optimization problem

$$\min_{B,L,\mu} \quad \frac{1}{2}\|P_{\Omega_1}(Y - XB - L - \mathbf{1}\mu^T)\|_F^2 + \rho\|L\|_* + \lambda\left(\alpha\|B\|_1 + \frac{1-\alpha}{2}\|B\|_F^2\right). \tag{13}$$

The solutions are $B(\rho, \alpha, \lambda)$, $L(\rho, \alpha, \lambda)$ and $\mu(\rho, \alpha, \lambda)$, then we evaluate the parameters by calculating the prediction error

$$Err(\rho, \alpha, \lambda) = \frac{1}{2}\|P_{\Omega_2}(Y - XB(\rho, \alpha, \lambda) - L(\rho, \alpha, \lambda) - \mathbf{1}\mu(\rho, \alpha, \lambda)^T)\|_F^2. \tag{14}$$

The grid search over three parameters may be too computationally intensive. Therefore, we first found an optimal value for ρ, $\hat{\rho}$, which minimizes the prediction error as shown in (Yang et al., 2013) by means of Lemmas 1 and 4 (**Appendix A**). Please refer to (Yang et al., 2013) to find the details about how to find the optimal value of ρ, $\hat{\rho}$. Once the optimal value of ρ, $\hat{\rho}$ is obtained, we selected a value of α from a sequence sequentially, thereafter, we performed one-dimensional grid search for λ for each α. Specifically, we generated a sequence of λ values with length n_λ decreasing from $\lambda_{max}(\hat{\rho}, \alpha)$ to $\epsilon\lambda_{max}(\hat{\rho}, \alpha)$ on the log scale with equal space, where $\lambda_{max}(\hat{\rho}, \alpha)$ is defined as the smallest λ such that $B(\hat{\rho}, \alpha, \lambda(\hat{\rho}, \alpha))$ is a zero matrix. $\lambda_{max}(\hat{\rho}, \alpha)$ is derived as $\frac{1}{\alpha} \max_{i=1,2,\ldots,p} \max_{j=1,2,\ldots,q} |\langle X_i, Y_j \rangle|$ from coordinate-descent algorithm (Friedman et al., 2007), where X_i is the i-th column of X, and Y_j the j-th column of Y. In our R program, we set $\epsilon = 0.02$, $n_\lambda = 50$ and $S_\alpha := (0.2, 0.4, 0.6, 0.8, 0.9)$. The optimal parameters were $(\hat{\rho}, \alpha, \hat{\lambda}(\hat{\rho}, \alpha))$ that minimize the prediction error. The optimal feasible solutions of B, L, and μ were then obtained based on the set of optimal tuning parameters.

TABLE 1 | Simulation scenarios.

Chromosome	#Causal SNPs	Scenario	Method	Screening
Chr 1	60	weak-dense	FastLORS	LORS
	200	strong-sparse	LORSEN	HC
	400		LORS	
Chr 1 + Chr 21	45 + 15	weak-dense	FastLORS	LORS
	150 + 50	strong-sparse	LORSEN	HC
	300 + 100		LORS	

2.4 Single Nucleotide Polymorphism Ranking and Joint Modeling

The procedure to select the set of optimal tuning parameters is computationally intensive. Therefore, as it is discussed in (Yang et al., 2013), it may not be computationally tractable to directly apply such method to the large-scale data sets that contain a large number of gene expression levels and SNPs. A commonly used strategy to reduce such computational burden is to choose a subset of SNPs and then only use them in the subsequent eQTL analysis. In this paper, we used and evaluated two existing methods for the pre-selection of informative SNPs: LORS-Screening (Yang et al., 2013) and Higher Criticism Screening (HC-Screening) (Jeng et al., 2020). For LORS-Screening, we first obtained the initial estimate of β_i's by solving

$$\min_{\beta_i, L, \mu} \quad \frac{1}{2}\|Y - X_i\beta_i^T - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_*, \tag{15}$$

where X_i is the i-th column of X, β_i is a q-dimensional vector for the coefficient of the i-th SNP on q genes, $i = 1, 2, \ldots, p$. For each gene, we selected the top n SNPs in terms of the absolute values of association coefficients, then we obtained the union of selected SNPs for each gene as the final set of SNPs to be involved in the joint modeling. For HC-Screening, we first obtained association coefficients as above, then calculated the standardized estimates of coefficients. For each SNP, the Higher Criticism (HC) statistic (Donoho and Jin, 2004) is calculated based on the standardized

TABLE 2 | Results of the HC-Screening and the LORS-Screening with ten replicates for each simulation scenario.

Chromosome	#Causal SNPs	Scenario	Screening	Average #Slected SNPs	Average #Selected Causal SNPs
Chr 1	60	weak-dense	LORS	1,017	43
			HC	165	7
		strong-sparse	LORS	1,023	60
			HC	165	9
	200	weak-dense	LORS	1,036	130
			HC	165	20
		strong-sparse	LORS	1,095	199
			HC	165	28
	400	weak-dense	LORS	1,045	237
			HC	165	39
		strong-sparse	LORS	1,142	346
			HC	165	44
Chr 1 + Chr 21	45 + 15	weak-dense	LORS	1,044	46
			HC	165	7
		strong-sparse	LORS	1,065	60
			HC	165	10
	150 + 50	weak-dense	LORS	1,064	136
			HC	165	20
		strong-sparse	LORS	1,123	199
			HC	165	28
	300 + 100	weak-dense	LORS	1,064	244
			HC	165	37
		strong-sparse	LORS	1,188	361
			HC	165	44

estimates of coefficients. Then we selected the top n SNPs in terms of the p-values of HC statistics.

2.5 Simulation Design

Our simulation is similar to that described in (Jeng et al., 2020). We first downloaded the genotype data of Chromosome 1 and Chromosome 21 for CEU samples from HapMap3, the third phase of the International HapMap Project (https://www.genome.gov/10001688/international-hapmap-project). CEU samples refer to Utah residents with Northern and Western European ancestry from the CEPH collection. After the quality-control (please refer to Real Data Analysis section), the genotype data of 13,815 SNPs of Chromosome 1 and 2,607 SNPs of Chromosome 21 for $n = 165$ samples were retained in analysis. To simulate gene expression levels for $q = 200$ genes over $n = 165$ samples, we first simulated non-genetic effects of $k = 15$ hidden factors. We randomly generated nk random numbers from $N(0, 1)$ to form a $n \times k$ matrix H, then let $\Sigma = HH^T$. U_j's were simulated from $N(0, 0.1^*\Sigma)$, $j = 1, 2, \ldots, q$ and stacked by column to form a $n \times q$ matrix U. e_j's were simulated from $N(0, I)$ as random noise for j-th gene expression and combined by column to form a $n \times q$ random noise matrix e. Then the expression data of q genes over n samples were simulated by $Y = XB + U + e$, where X is the $n \times p$ genotype data matrix. We set the total number of SNPs $p = 2000$, the number of causal SNPs as 60, 200, or 400. Each causal SNP randomly influences $m = 10$ (or 50) genes. We simulated nonzero genetic effects from a uniform distribution. For the "weak-dense" scenario, each causal SNP affects $m = 50$ randomly selected genes and the corresponding values in B were simulated from a uniform distribution between 0.25 and 0.75. For the "strong-sparse" scenario, each causal SNP affects $m = 10$ randomly selected genes and the corresponding values in B were simulated from a uniform distribution between 1.5 and 2. The different simulation scenarios are summarized in **Table 1**.

3 RESULTS

3.1 Simulation Results

The number of selected SNPs and the number of selected causal SNPs from two screening methods under different simulation scenarios are summarized in **Table 2**. Several conclusions emerge from **Table 2**. First, when the number of samples is much smaller than the number of SNPs and the number of causal SNPs is larger than the number of samples, HC-Screening is seemingly not an appropriate screening tool. This is because the number of causal SNPs retained after the HC-Screening is much smaller than the actual number of causal SNPs, resulting in possible power loss in subsequent analysis. Second, even when the number of causal SNPs is smaller than the number of samples, from **Table 2**, we still observed that the LORS-Screening retains more causal SNPs than the HC-Screening. Of course, the HC-Screening reduces much computational burden especially when the number of samples is much smaller than the number of SNPs.

The area under the curve (AUC) was used to compare the performance between LORSEN and two existing methods, LORS (Yang et al., 2013) and FastLORS (Jeng et al., 2020). For each

TABLE 3 | The average AUC and 95% confidence interval without the SNP screening with ten replicates for each simulation scenario. SNPs are only from chromosome 1. For each simulation scenario, the highest AUC is in bold.

Scenario	Method	#Causal SNPs		
		60	200	400
weak-dense	FastLORS	0.514 (0.511, 0.517)	0.582 (0.580, 0.584)	0.581 (0.580, 0.582)
	LORSEN	**0.651** (0.648, 0.654)	**0.649** (0.647, 0.651)	**0.630** (0.629, 0.631)
	LORS	0.502 (0.499, 0.505)	0.514 (0.512, 0.516)	0.515 (0.514, 0.516)
strong-sparse	FastLORS	0.762 (0.755, 0.769)	**0.840** (0.837, 0.843)	**0.810** (0.807, 0.813)
	LORSEN	0.823 (0.817, 0.829)	0.834 (0.831, 0.837)	0.774 (0.771, 0.777)
	LORS	**0.824** (0.818, 0.830)	0.819 (0.815, 0.823)	0.754 (0.751, 0.757)

TABLE 4 | The average AUC and 95% confidence interval without the SNP screening with ten replicates for each simulation scenario. SNPs are from chromosome 1 and chromosome 21. For each simulation scenario, the highest AUC is in bold.

Scenario	Method	#Causal SNPs		
		60	200	400
weak-dense	FastLORS	0.530 (0.527, 0.533)	0.567 (0.565, 0.569)	0.575 (0.574, 0.576)
	LORSEN	**0.658** (0.655, 0.661)	**0.679** (0.677, 0.681)	**0.625** (0.624, 0.626)
	LORS	0.503 (0.500, 0.506)	0.510 (0.508, 0.512)	0.514 (0.513, 0.515)
strong-sparse	FastLORS	0.774 (0.767, 0.781)	**0.826** (0.822, 0.830)	**0.813** (0.810, 0.816)
	LORSEN	**0.814** (0.807, 0.821)	0.810 (0.806, 0.814)	0.788 (0.785, 0.791)
	LORS	0.813 (0.806, 0.820)	0.801 (0.797, 0.805)	0.756 (0.753, 0.759)

TABLE 5 | The average AUC and 95% confidence interval with the SNP screening with ten replicates for each simulation scenario. SNPs are only from chromosome 1. For each simulation scenario, the highest AUC is in bold.

Scenario	#Causal SNPs	Method	Screening	
			HC	LORS
weak-dense	60	FastLORS	0.514 (0.511, 0.517)	0.596 (0.593, 0.599)
		LORSEN	**0.515** (0.512, 0.518)	**0.618** (0.615, 0.621)
		LORS	0.503 (0.500, 0.506)	0.541 (0.538, 0.544)
	200	FastLORS	**0.512** (0.510, 0.514)	0.583 (0.581, 0.585)
		LORSEN	0.511 (0.509, 0.513)	**0.592** (0.590, 0.594)
		LORS	0.502 (0.500, 0.504)	0.519 (0.517, 0.521)
	400	FastLORS	**0.510** (0.509, 0.511)	**0.557** (0.556, 0.558)
		LORSEN	0.509 (0.508, 0.510)	0.547 (0.546, 0.548)
		LORS	0.502 (0.501, 0.503)	0.511 (0.510, 0.512)
strong-sparse	60	FastLORS	**0.565** (0.557, 0.573)	0.900 (0.895, 0.905)
		LORSEN	0.558 (0.550, 0.566)	**0.903** (0.898, 0.908)
		LORS	0.560 (0.552, 0.568)	0.897 (0.892, 0.902)
	200	FastLORS	**0.552** (0.548, 0.556)	**0.894** (0.891, 0.897)
		LORSEN	0.544 (0.540, 0.548)	**0.894** (0.891, 0.897)
		LORS	0.543 (0.539, 0.547)	0.874 (0.871, 0.877)
	400	FastLORS	**0.536** (0.533, 0.539)	**0.797** (0.794, 0.800)
		LORSEN	0.523 (0.520, 0.526)	0.782 (0.779, 0.785)
		LORS	0.528 (0.525, 0.531)	0.738 (0.735, 0.741)

scenario, we repeated the simulation ten times. We considered the joint modeling of multiple SNPs and multiple gene expression levels with the SNP screening and without the SNP screening. The results without the SNP screening before the eQTL mapping under different simulation scenarios are presented in **Tables 3**, **4**.

From **Tables 3**, **4**, we can see that the average AUC of LORSEN is uniformly larger than those of LORS and FastLORS in the weak-dense scenarios across different number of causal SNPs no matter the SNPs are from single chromosome (Chr 1) or two chromosomes (Chr 1 + Chr 21). For the strong-sparse

TABLE 6 | The average AUC and 95% confidence interval with the SNP screening with ten replicates for each simulation scenario. SNPs are from chromosome 1 and chromosome 21. For each simulation scenario, the highest AUC is in bold.

Scenario	#Causal SNPs	Method	Screening	
			HC	LORS
weak-dense	60	FastLORS	0.518 (0.515, 0.521)	0.606 (0.603, 0.609)
		LORSEN	**0.518** (0.515, 0.521)	**0.629** (0.626, 0.632)
		LORS	0.505 (0.502, 0.508)	0.544 (0.541, 0.547)
	200	FastLORS	**0.512** (0.510, 0.514)	0.591 (0.589, 0.593)
		LORSEN	**0.512** (0.510, 0.514)	**0.615** (0.613, 0.617)
		LORS	0.503 (0.501, 0.505)	0.524 (0.522, 0.526)
	400	FastLORS	**0.510** (0.509, 0.511)	**0.563** (0.562, 0.564)
		LORSEN	0.507 (0.506, 0.508)	0.556 (0.555, 0.557)
		LORS	0.501 (0.500, 0.502)	0.511 (0.510, 0.512)
strong-sparse	60	FastLORS	**0.570** (0.562, 0.578)	0.891 (0.886, 0.896)
		LORSEN	0.563 (0.555, 0.571)	**0.906** (0.901, 0.911)
		LORS	0.564 (0.556, 0.572)	0.891 (0.886, 0.896)
	200	FastLORS	**0.553** (0.549, 0.557)	**0.904** (0.901, 0.907)
		LORSEN	0.547 (0.543, 0.551)	**0.904** (0.901, 0.907)
		LORS	0.544 (0.540, 0.548)	0.883 (0.880, 0.886)
	400	FastLORS	**0.534** (0.531, 0.537)	**0.821** (0.818, 0.824)
		LORSEN	0.524 (0.521, 0.527)	0.813 (0.810, 0.816)
		LORS	0.525 (0.522, 0.528)	0.765 (0.762, 0.768)

scenarios, FastLORS achieves the relatively larger AUC than LORS and LORSEN. For a fixed number of causal SNPs, each method achieves the larger AUC value in the stong-sparse scenario than in the weak-dense scenario. For each method under each simulation scenario, the AUCs in **Tables 3**, **4** are similar, implying that each of three methods has the similar power to detect *cis*-eQTLs and *trans*-eQTLs.

The results with the SNP screening before eQTL mapping under different simulation scenarios are presented in **Tables 5**, **6**. As we have mentioned, the LORS-Screening keeps more SNPs in the analysis, thus retains more causal SNPs than the HC-Screening does. Each method with the LORS-Screening has the larger AUC values than it with the HC-Screening. From **Tables 5**, **6**, we can see that the AUC values of methods with the HC-Screening are quite close to 0.5, which indicates that the HC-Screening can essentially lead to the loss of power of methods. With the LORS-Screening, similar to the non-screening cases, LORSEN has better performance than LORS and FastLORS in the weak-dense scenarios and LORSEN and FastLORS perform similarly and slightly better than LORS in the strong-sparse scenarios. Finally, we find that for the weak-dense scenarios, each method without the SNP screening before joint modeling achieves the larger AUC values than it with the SNP screening. However, for the strong-sparse scenarios, each method with the LORS-Screening before joint modeling achieves the larger AUC values than it without the SNP screening. This may be due to that there are a large number of SNP-gene pairs with the weak association effects in the weak-dense scenarios and many causal SNPs may not be selected by the pre-screening methods. So, in the weak-dense scenarios with the use of pre-screening methods, the computational cost and the detection power can be reduced at the same time. In the strong-sparse scenarios, there are a smaller number of SNP-gene pairs with the stronger association effects than in the weak-dense scenarios, and it is expected that most of the causal SNPs will be selected by the pre-screening methods. Therefore, for the strong-sparse scenarios, the use of pre-screening methods reduce the computational cost while still retain the high detection power.

Our simulation results showed that LORSEN is more powerful to identify weak signals, while it does not have obvious advantage in identifying strong signals compared to LORS. Therefore, we performed additional simulation studies in which the causal variants have mixed weak and strong effects. Specifically, the half of the causal variants had the weak effects and their effects were generated from a uniform distribution between 0.25 and 0.75, while the other half of the causal variants had the strong effects and their effects were generated from a uniform distribution between 1.5 and 2. The number of causal SNPs was set as 60, 200, or 400. The number of genes affected by one causal SNP was set as 50. The AUCs and corresponding 95% confidence intervals are presented in **Supplementary Table S1**. From the results in **Supplementary Table S1**, we can see that LORSEN has the overall highest detection power when the number of causal SNPs is large. It is well known that the rare variants play an important role in the etiology of human complex diseases. Therefore, it is necessary to assess the performance of eQTL mapping methods when most of causal variants are rare. We conducted simulations in which the proportion of rare causal variants was set to be 50 and 75%. Here, the variants with minor allel frequency (MAF) less than 0.03 were considered as the rare variants. The number of causal variants was set as 200. The results from different simulation scenarios (weak-dense and strong-sparse) are presented in **Supplementary Table S2**. From **Supplementary Table S2**, we can see that when the proportion

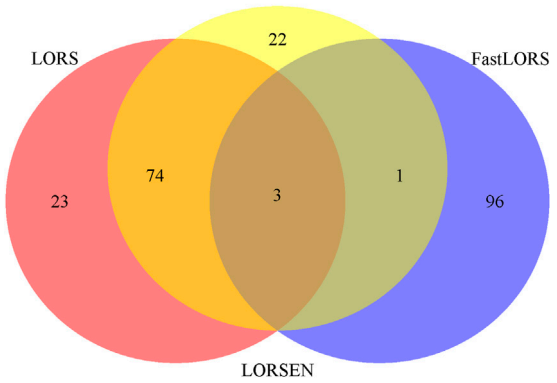


FIGURE 1 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 3.

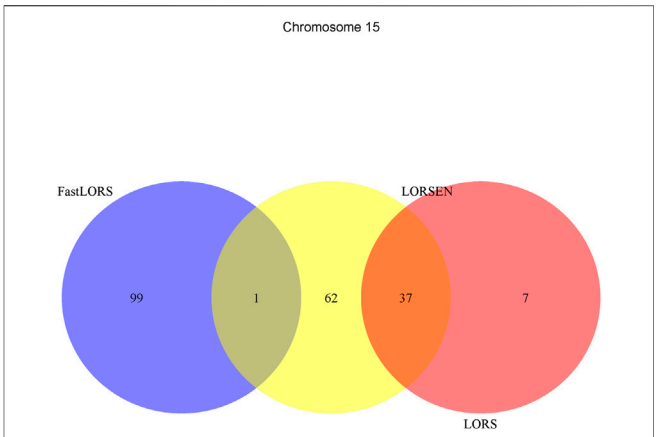


FIGURE 2 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 15.

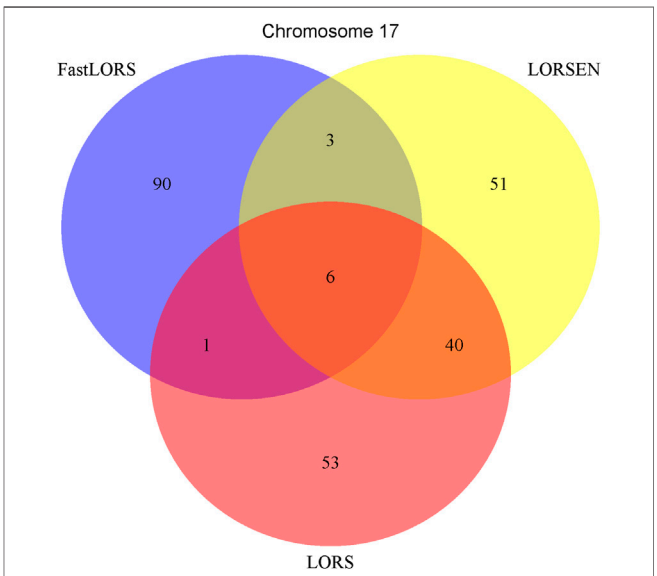


FIGURE 3 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 17.

of causal rare variants is 50%, the AUCs of FastLORS are slightly higher than the AUCs of LORSEN. However, when the proportion of causal rare variants is 75%, the AUCs of LORSEN are at least 10% higher AUCs than the AUCs of FastLORS and about 20% higher than the AUCs of LORS. Our results show that LORSEN has the higher power in detecting rare causal variants. To see how the detection power of LORSEN is affected by the positive and negative effects, we conducted simulations in which the half of the causal variants had the positive effects on genes and the other half of the causal variants had the negative effects on genes. The results from different simulation scenarios (weak-dense and strong-sparse with 60, 200, and 400 causal variants) are presented in **Supplementary Table S3**. From **Supplementary Table S3**, we can see that LORSEN achieves the highest AUCs in almost all simulation scenarios, which implies that the detection power of LORSEN is not affected by the effect directions of causal variants.

In addition to AUC, a commonly used measure to assess the performance of methods for eQTL mapping, we also reported the false positive rates (FPRs) based on four thresholds for the regression coefficients: 0, 10^{-12}, 10^{-6}, 10^{-4}. From the **Supplementary Figures S1–S3**, we can see that FastLORS has the highest FPRs in almost all scenarios, and the FPRs of FastLORS are quite sensitive to the thresholds: the FPRs of FastLORS decrease dramatically for large thresholds. LORS has the smallest FPRs in all simulation scenarios. For LORSEN, it has the small and comparable FPRs with LORS when the effects of the causal variants are all weak or are a mixture of weak and strong effects. LORSEN has the large FPRs when the effects of the causal variants are all strong.

A number of conclusions emerge from the results based on our extensive simulation studies. First, the HC-Screening method retains much smaller number of SNPs than the LORS-Screening method. Second, when all the SNPs are not filtered with the SNP screening method and are used in the analysis, LORSEN is the most powerful method to identify weak signals, while it does not have obvious advantage in identifying strong signals compared to LORS and FastLORS. LORSEN still performs the best with the mixture of the strong and weak effects when the number of causal variants is large. Third, when the SNPs are first filtered with the HC-Screening method, FastLORS performs the best in all simulation scenarios. With the LORS-Screening method, LORSEN has the highest detection power in most of simulation scenarios. Fourth, LORSEN outperforms FastLORS and LORS when a large portion of the causal SNPs are rare and when the causal variants have a mixture of positive and negative effects.

3.2 Real Data Analysis Results

To illustrate our method in real data analysis, we also applied LORS-LORSEN (LORSEN with the LORS-Screening), LORS-LORS (LORS with the LORS-Screening) and HC-FastLORS

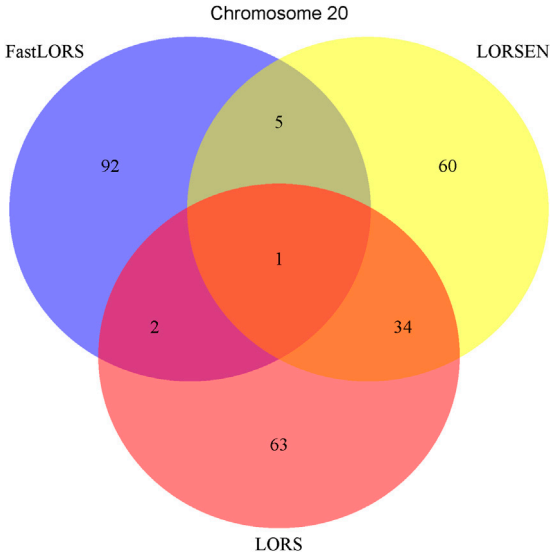


FIGURE 4 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 20.

(FastLORS with the HC-Screening) to the HapMap3 data. Here, we focused on Asian samples (CHB and JPT) in the HapMap3 data and selected four chromosomes for the analysis. SNP genotype data and gene expression data are publicly available, and can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3_r3/plink_format/and http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264/, respectively. Because the set of samples with the SNP genotype data and the set of samples with the gene expression data are slightly different, we only kept the samples that have both the SNP genotype data and the gene expression data in the analysis. We removed SNPs with missing values, and performed the LD pruning using PLINK with its default parameters (window size: 50; moving window increment: five SNPs; cutoff value of R^2: 0.5). After the data pre-processing, a total of 160 samples (CHB: 79; JPT: 81) were included in analysis. The number of SNPs and the number of genes with the expression used in the analysis on chromosome 3 are 4,086 and 1,075, on chromosome 15 are 2,235 and 612, on chromosome 17 are 2,226 and 1,098, on chromosome 20 are 1,863 and 606, respectively. Since the significance tests generally cannot be performed for the penalization based regression models, we focused on the top 100 SNP-probe pairs with the largest absolute regression coefficients. From the Venn diagrams (**Figures 1**–**4**), we

TABLE 7 | Top ten detected SNP-probe pairs for chromosome 3. The SNP-probe pairs that are confirmed in seeQTL database are in bold.

Method	SNP	Probe (gene)	Association coefficient	Distance	Class
HC-FastLORS	rs13084976	ILMN_1657373 (LEPREL1)	0.0430	188.72 mb	distant
	rs17029694	ILMN_1657373 (LEPREL1)	0.0424	188.49 mb	distant
	rs12494696	ILMN_1812093 (UTS2D)	0.0322	189.72 mb	distant
	rs2322212	ILMN_1756501 (ST6GAL1)	0.0310	184.74 mb	distant
	rs17029694	ILMN_1708743 (NT5DC2)	0.0303	49.86 mb	distant
	rs2322212	ILMN_1686920 (CCDC58)	0.0300	120.03 mb	distant
	rs7647780	ILMN_1762084 (DNASE1L3)	0.0292	57.51 mb	distant
	rs1516347	ILMN_1726020 (LOC652670)	0.0278	75.49 mb	distant
	rs13061928	ILMN_1692261 (EPHB1)	0.0273	133.55 mb	distant
	rs1377213	ILMN_1698934 (CMTM7)	0.0270	26.76 mb	distant
LORS-LORSEN	rs1505587	ILMN_1657373 (LEPREL1)	0.3336	127.69 mb	distant
	rs6807033	ILMN_1787750 (CD200)	0.2796	4.163 kb	local
	rs11914577	ILMN_1700967 (C3orf59)	0.2245	113.51 kb	local
	rs1403719	ILMN_1771599 (PLOD2)	0.1963	25.06 mb	distant
	rs628267	ILMN_1760509 (EOMES)	0.1941	302.30 kb	
	rs4016435	**ILMN_1757350 (CTNNB1)**	**0.1908**	**27.772 kb**	**local**
	rs16839507	**ILMN_1761058 (ACAD11)**	**0.1856**	**117.942 kb**	**local**
	rs693430	**ILMN_1657708 (MGLL)**	**0.1796**	**86.074 kb**	**local**
	rs693430	**ILMN_1707310 (MGLL)**	**0.1710**	**47.617 kb**	**local**
	rs1498090	ILMN_1793724 (C3orf31)	0.1662	58.605 kb	local
LORS-LORS	rs1505587	ILMN_1657373 (LEPREL1)	1.2549	127.69 mb	distant
	rs6807033	ILMN_1787750 (CD200)	0.5621	4.163 kb	local
	rs4857653	ILMN_1700967 (C3orf59)	0.3640	16.16 mb	distant
	rs11914577	ILMN_1700967 (C3orf59)	0.2984	113.514 kb	local
	rs1403719	ILMN_1771599 (PLOD2)	0.2824	25.06 mb	distant
	rs628267	ILMN_1760509 (EOMES)	0.2439	302.302 kb	
	rs4016435	**ILMN_1757350 (CTNNB1)**	**0.2404**	**27.772 kb**	**local**
	rs16839507	**ILMN_1761058 (ACAD11)**	**0.2338**	**117.942 kb**	**local**
	rs3773014	ILMN_1762084 (DNASE1L3)	0.2268	29.187 kb	local
	rs1799977	ILMN_1688392 (ZBED2)	0.2234	75.77 mb	distant

TABLE 8 | The SNP-probe pairs found in seeQTL database out of the top ten SNP-probe pairs for chromosomes 3, 15, 17, and 20, respectively.

Chromosome	SNP	Probe (gene)	Method	Information from GTEx
3	rs4016435	ILMN_1757350 (CTNNB1)	LORS-LORSEN, LORS-LORS	Not found in GTEx
3	rs16839507	ILMN_1761058 (ACAD11)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
3	rs693430	ILMN_1657708 (MGLL)	LORS-LORSEN	Not found in GTEx
3	rs693430	ILMN_1657708 (MGLL)	LORS-LORSEN	Not found in GTEx
15	rs7162538	ILMN_1784364 (STARD5)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
15	rs1347069	ILMN_1795822 (DIS3L)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
15	rs2292114	ILMN_1795524 (C15orf44)	LORS-LORS	Not found in GTEx
17	rs4968140	ILMN_1706959 (TIMM22)	HC-FastLORS	Not found in GTEx
17	rs4251704	ILMN_1773352 (CCL5)	LORS-LORSEN	A single hit for sQTLs
17	rs17657522	ILMN_1697227 (USP36)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
17	rs4968140	ILMN_1706959 (TIMM22)	LORS-LORSEN, LORS-LORS	Not found in GTEx
17	rs9905601	ILMN_1750511 (NT5C3L)	LORS-LORS	Not found in GTEx
20	rs16989514	ILMN_1721128 (TOMM34)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
20	rs6041750	ILMN_1702237 (FKBP1A)	HC-FastLORS, LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs

notice that there is a large overlap between the eQTLs identified by LORS-LORS and LORS-LORSEN. However, there is a small overlap between the eQTLs identified by HC-FastLORS and LORS-LORS (or LORS-LORSEN). For example, among the top 100 SNP-probe pairs identified on Chromosome 3 (**Figure 1**), LORS-LORS and LORS-LORSEN share 77 SNP-probe pairs in common, while LORS-LORSEN and HC-FastLORS only share four SNP-probe pairs in common and LORS-LORS and HC-FastLORS share three SNP-probe pairs in common. This observation is consistent with the observation from (Jeng et al., 2020) which also noticed that there is a small overlap between the SNP-probe pairs identified by LORS-LORS and HC-FastLORS. Additionally, as adopted in (Jeng et al., 2020), we classified the detected eQTL as local if the physical distance between the SNP and the probe midpoint is less than 250 kb or as distant if the distance is greater than 5 mb following the criterion described in (Westra et al., 2013). For each chromosome, we report our findings on the top ten identified SNP-probe pairs in **Table 7** and **Supplementary Tables S4–S6** (see **Supplementary Material**). From **Table 7**, we can see that the SNPs in the top ten SNP-probe pairs identified by HC-FastLORS are all *trans*-eQTLs. As a comparison, seven SNPs in the top ten SNP-probe pairs identified by LORS-LORSEN are *cis*-eQTLs and two SNPs are *trans*-eQTLs. Five SNPs in the top ten SNP-probe pairs identified by LORS-LORS are *cis*-eQTLs and four SNPs are *trans*-eQTLs. LORS-LORSEN and LORS-LORS share seven SNP-probe pairs while LORS-LORSEN and LORS-LORS do not share any SNP-probe pair with HC-FastLORS. In addition, the coefficients obtained from HC-FastLORS are ten-fold smaller than those obtained from LORS-LORSEN and LORS-LORS. This indicates that the findings of LORS-LORSEN and LORS-LORS may be more convincing.

To further validate our findings, we searched an existing database called seeQTL (Xia et al., 2011). seeQTL (https://seeqtl.org/) records the eQTLs identified from a meta-analysis (consensus eQTLs) from the HapMap human lymphoblastoid cell lines. A total of fourteen SNP-probe pairs were found in seeQTL and were listed in **Table 8**. Among them, two SNP-probe pairs were identified by HC-FastLORS only, three were identified by LORS-LORSEN only, two were identified by LORS-LORS only, seven were identified by both LORS-LORSEN and LORS-LORS, and one was identified by all three methods. To further validate these fourteen SNP-probe pairs, we searched the eQTL web-browser (http://www.gtexportal.org/home/) built by the Genotype-Tissue Expression Project (GTEx) (https://www.genome.gov/Funded-Programs-Projects/Genotype-Tissue-Expression-Project) to see if those SNP-probe (gene) pairs are listed as the eQTLs and/or sQTLs (splicing quantitative trait locus). A total of seven SNP-probe pairs were also found in GTEx and were presented in **Table 8**. Among seven SNP-probe pairs found both in seeQTL and GTEx, one SNP-probe pair was identified by all three methods, five SNP-probe pairs were identified by both LORS-LORSEN and LORS-LORS, and one SNP-probe pair was identified by LORS-LORSEN only.

A number of conclusions emerge from the results based on HapMap3 data. First, there is a large overlap between the SNP-probe pairs identified by LORS-LORS and LORS-LORSEN but there is a small overlap between the SNP-probe pairs identified by HC-FastLORS and LORS-LORS (or LORS-LORSEN). Second, LORS-LORS and LORS-LORSEN perform similarly and have higher detection power than HC-FastLORS since LORS-LORS and LORS-LORSEN have identified more SNP-probe pairs that are also found in seeQTL and GTEx. Third, five out of seven SNP-probe pairs identified by both LORS-LORS and LORS-LORSEN and found in seeQTL are also found in GTEx, thus it may be beneficial to combine the results from multiple methods to generate a list of SNP-probe pairs for further investigation.

4 DISCUSSION

As more human gene expression data become available, fast and efficient statistical and computational methods are needed to fully take advantage of such data to investigate the relationship between genetic variants and gene expression levels to further reveal the genetic mechanisms that underlie human complex diseases. However, most existing methods are built on small-scale samples and not applicable to human-size datasets. In this paper, we proposed a new low rank penalized regression method (LORSEN) to detect eQTLs. We developed a fast and efficient algorithm to solve optimization problems arising from our methods based on proximal gradient methods. Comprehensive simulation studies showed that LORSEN outperformed two

commonly used methods, LORS and FastLORS, in many simulation scenarios. From our simulation results, we can briefly conclude that, first, LORSEN is more powerful in detecting eQTLs which are rare and/or have weak effects. This is especially an appealing advantage since it is expected that a portion of causal variants are rare and/or have the weak effects in the real world. Second, LORSEN is more powerful when some causal variants have the positive effects and the other causal variants have the negative effects.

Since there are a large number of SNPs and genes to be included in the eQTL mapping and it is expected that only a small portion of SNPs will affect the gene expression levels, a number of pre-screening methods have been developed. In this paper, we used the LORS-Screening (Yang et al., 2013) and the HC-Screening (Jeng et al., 2020). We found that the HC-Screening retained much smaller number of SNPs than the LORS-Screening. Both the LORS-Screening and the HC-Screening can reduce the computational cost, but they may also reduce the detection power in the eQTL mapping, depending on the association patterns between SNPs and gene expression levels. Since we do not know such association patterns in real studies, we should be cautious to apply such pre-screening methods.

There are several limitations for LORSEN. First, as a method based on the penalized regression model, we can rank the SNP-gene pairs in terms of the regression coefficients obtained from LORSEN, but cannot perform the significance test. Second, the computational time of LORSEN depends on many factors such as the number of candidate values of hyperparameters, the initial values of hyperparameters, and the number of samples. The computation was performed parallelly using software R (verson 4.1.1) and 16 cores on a server with 64 Intel(R) Xeon(R) Gold 6130 CPUs @ 2.10 GHz. From **Supplementary Table S7**, we can see that, as expected, LORSEN costs much more time in parameter tuning than other two methods due to the exhaustive grid search. The grid search is easy to be implemented but is computationally intensive. It may not be feasible for large scale data. A more efficient strategy is desirable.

It has shown that the incorporation of the SNP correlation and the gene interaction network can potentially increase the power of detecting eQTLs (Kim and Xing, 2009; Chen et al., 2012; Kim and Xing, 2012; Cheng et al., 2014). We expect that our method can be improved if we use the prior knowledge of correlation structures of SNPs and genes to refine the penalty terms in optimization problems.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CG and KZ conceived the study, CG developed the method and algorithms, CG developed the R program LORSEN, CG performed simulation studies and real data analysis, CG drafted the article, CG, HW, and KZ revised the article. All the authors read and approved the final article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.690926/full#supplementary-material

REFERENCES

Albert, F. W., and Kruglyak, L. (2015). The Role of Regulatory Variation in Complex Traits and Disease. *Nat. Rev. Genet.* 16, 197–212. doi:10.1038/nrg3891

Banerjee, S., Simonetti, F. L., Detrois, K. E., Kaphle, A., Mitra, R., Nagial, R., et al. (2021). Tejaas: Reverse Regression Increases Power for Detecting Trans-eqtls. *Genome Biol.* 22, 142. doi:10.1186/s13059-021-02361-8

Beck, A., and Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.* 2, 183–202. doi:10.1137/080716542

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM J. Optim.* 20, 1956–1982. doi:10.1137/080738970

Chen, X., Shi, X., Xu, X., Wang, Z., Mills, R., Lee, C., et al. (2012). "A Two-Graph Guided Multi-Task Lasso Approach for Eqtl Mapping," in Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. Editors N. D. Lawrence and M. Girolami (La Palma, Canary Islands: PMLR), 208–217. vol. 22 of Proceedings of Machine Learning Research, April 21-23, 2012, La Palma, Canary Islands.

Cheng, W., Zhang, X., Guo, Z., Shi, Y., and Wang, W. (2014). Graph-regularized Dual Lasso for Robust Eqtl Mapping. *Bioinformatics* 30, i139–i148. doi:10.1093/bioinformatics/btu293

Chun, H., and Keleş, S. (2009). Expression Quantitative Trait Loci Mapping with Multivariate Sparse Partial Least Squares Regression. *Genetics* 182, 79–90. doi:10.1534/genetics.109.100362

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping Complex Disease Traits with Global Gene Expression. *Nat. Rev. Genet.* 10, 184–194. doi:10.1038/nrg2537

Donoho, D., and Jin, J. (2004). Higher Criticism for Detecting Sparse Heterogeneous Mixtures. *Ann. Stat.* 32, 962–994. doi:10.1214/009053604000000265

Fan, J., and Lv, J. (2008). Sure independence Screening for Ultrahigh Dimensional Feature Space. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 70, 849–911. doi:10.1111/j.1467-9868.2008.00674.x

Fan, Y., Zhu, H., Song, Y., Peng, Q., and Zhou, X. (2020). Efficient and Effective Control of Confounding in Eqtl Mapping Studies through Joint Differential Expression and Mendelian Randomization Analyses. *Bioinformatics* 37, 296–302. doi:10.1093/bioinformatics/btaa715

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise Coordinate Optimization. *Ann. Appl. Stat.* 1, 302–332. doi:10.1214/07-aoas131

Fusi, N., Stegle, O., and Lawrence, N. D. (2012). Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies. *Plos Comput. Biol.* 8, e1002330. doi:10.1371/journal.pcbi.1002330

Gao, C., Tignor, N. L., Salit, J., Strulovici-Barel, Y., Hackett, N. R., Crystal, R. G., et al. (2014). Heft: Eqtl Analysis of many Thousands of Expressed Genes while Simultaneously Controlling for Hidden Factors. *Bioinformatics* 30, 369–376. doi:10.1093/bioinformatics/btt690

Hanley, J. A., and McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (Roc) Curve. *Radiology* 143, 29–36. doi:10.1148/radiology.143.1.7063747

Hu, Y.-J., Sun, W., Tzeng, J.-Y., and Perou, C. M. (2015). Proper Use of Allele-specific Expression Improves Statistical Power Forcis-eQTL Mapping with RNA-Seq Data. *J. Am. Stat. Assoc.* 110, 962–974. doi:10.1080/01621459.2015.1038449

Jeng, X. J., Rhyne, J., Zhang, T., and Tzeng, J.-Y. (2020). Effective SNP Ranking Improves the Performance of eQtl Mapping. *Genet. Epidemiol.* 44, 611–619. doi:10.1002/gepi.22293

Kendziorski, C. M., Chen, M., Yuan, M., Lan, H., and Attie, A. D. (2006). Statistical Methods for Expression Quantitative Trait Loci (Eqtl) Mapping. *Biometrics* 62, 19–27. doi:10.1111/j.1541-0420.2005.00437.x

Kim, S., and Xing, E. P. (2009). Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *Plos Genet.* 5, e1000587. doi:10.1371/journal.pgen.1000587

Kim, S., and Xing, E. P. (2012). Tree-guided Group Lasso for Multi-Response Regression with Structured Sparsity, with an Application to Eqtl Mapping. *Ann. Appl. Stat.* 6, 1095–1117. doi:10.1214/12-AOAS549

Lee, S., and Xing, E. P. (2012). Leveraging Input and Output Structures for Joint Mapping of Epistatic and Marginal Eqtls. *Bioinformatics* 28, i137–i146. doi:10.1093/bioinformatics/bts227

Listgarten, J., Kadie, C., Schadt, E. E., and Heckerman, D. (2010). Correction for Hidden Confounders in the Genetic Analysis of Gene Expression. *Proc. Natl. Acad. Sci.* 107, 16465–16470. doi:10.1073/pnas.1002425107

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J. Mach. Learn. Res.* 11, 2287–2322.

Parikh, N., and Boyd, S. (2014). Proximal Algorithms. *FNT in Optimization* 1, 127–239. doi:10.1561/2400000003

Rakitsch, B., and Stegle, O. (2016). Modelling Local Gene Networks Increases Power to Detect Trans-acting Genetic Effects on Gene Expression. *Genome Biol.* 17, 33. doi:10.1186/s13059-016-0895-2

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian Framework to Account for Complex Non-genetic Factors in Gene Expression Levels Greatly Increases Power in Eqtl Studies. *Plos Comput. Biol.* 6, e1000770. doi:10.1371/journal.pcbi.1000770

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Wang, P., Dawson, J. A., Keller, M. P., Yandell, B. S., Thornberry, N. A., Zhang, B. B., et al. (2011). A Model Selection Approach for Expression Quantitative Trait Loci (Eqtl) Mapping. *Genetics* 187, 611–621. doi:10.1534/genetics.110.122796

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic Identification of Trans Eqtls as Putative Drivers of Known Disease Associations. *Nat. Genet.* 45, 1238–1243. doi:10.1038/ng.2756

Xia, K., Shabalin, A. A., Huang, S., Madar, V., Zhou, Y.-H., Wang, W., et al. (2011). Seeqtl: a Searchable Database for Human Eqtls. *Bioinformatics* 28, 451–452. doi:10.1093/bioinformatics/btr678

Yang, C., Wang, L., Zhang, S., and Zhao, H. (2013). Accounting for Non-genetic Factors by Low-Rank Representation and Sparse Regression for Eqtl Mapping. *Bioinformatics* 29, 1026–1034. doi:10.1093/bioinformatics/btt075

Yu, Y.-L. (2013). On Decomposing the Proximal Map. *Adv. Neural Inf. Process. Syst.* 26, 91–99.

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. B* 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX A

Lemma 1: For each $\tau \geqslant 0$ and $Y \in \mathbb{R}^{n_1 \times n_2}$, the solution of

$$\min_X \quad \frac{1}{2}\|X - Y\|_F^2 + \tau\|X\|_* \tag{16}$$

is $S_\tau(Y) := US_\tau(\Sigma)V^T (= Prox_{\tau\|\cdot\|_*}(Y))$, where $S_\tau(\Sigma) = diag(\{(\sigma_i - \tau)_+\})$, $Y = U\Sigma V^T$, the singular value decomposition of matrix Y, $\Sigma = diag(\{\sigma_i\}_{1 \leqslant i \leqslant r})$, r is the rank of Y. $S_\tau(\cdot)$ is called singular value shrinkage operator.

Proof: see (Cai et al., 2010) or (Mazumder et al., 2010).

Lemma 2: For each fixed non-negative λ and $v \in \mathbb{R}^n$, the solution of

$$\min_x \quad \frac{1}{2}\|x - v\|_2^2 + \frac{\lambda}{2}\|x\|_2^2 \tag{17}$$

is $(Prox_{\frac{\lambda}{2}\|\cdot\|_2^2}(v))_i = sign(v_i)(|v_i| - \lambda)_+$, $i = 1, 2, \ldots, n$, known as the (elementwise) soft thresholding operator.

Proof: see (Parikh and Boyd, 2014).

Lemma 3: For each fixed non-negative ρ and $v \in \mathbb{R}^n$, the solution of

$$\min_x \quad \frac{1}{2}\|x - v\|_2^2 + \rho\|x\|_1 \tag{18}$$

is $Prox_{\rho\|\cdot\|_1}(v) = (1 - \frac{\rho}{max\{\|v\|_2, \rho\}})v$.

Proof: see (Parikh and Boyd, 2014).

Lemma 4: (soft-impute algorithm)

For the optimization problem

$$\min_X \quad \frac{1}{2}\|P_\Omega(Y - X)\|_F^2 + \tau\|X\|_*$$
$$= \min_X \quad \frac{1}{2}\|[P_\Omega(Y) + P_{\Omega^\perp}(X)] - X\|_F^2 + \tau\|X\|_*,$$

the optimization solution can be obtained via updating X using $X \leftarrow S_\tau(P_\Omega(Y) + P_{\Omega^\perp}(X))$ with an arbitrary initialization.

Proof: see (Mazumder et al., 2010).

Theorem 1: A sufficient condition for $Prox_{f+g} = Prox_f \circ Prox_g$ is $\forall\, x \in \mathcal{H}, \partial g(Prox_f(x)) \supseteq \partial g(x)$, where $\mathcal{H}$ represents Hilbert space and $\circ$ represents composition of two operators.

Proof: see (Yu, 2013).

Details of Confidence Interval of AUC

We followed the method used in (Hanley and McNeil, 1982) to calculate the 95% confidence interval (CI) of AUC. Let $\widehat{AUC}$ and $Var(\widehat{AUC})$ denote the sample mean and the estimated variance of AUCs from ten replicates, respectively, the 95% CI of average AUC was calculated using the following formula:

$$\overline{\widehat{AUC}} \pm 1.96\sqrt{Var(\widehat{AUC})/10}. \tag{19}$$

We used the following formula (Hanley and McNeil, 1982) to calculate $Var(\widehat{AUC})$:

$$Var(\widehat{AUC}) = \frac{q_0 + (n_1 - 1)q_1 + (n_2 - 1)q_2}{n_1 n_2}, \tag{20}$$

where $q_0 = \widehat{AUC}(1 - \widehat{AUC})$, $q_1 = \frac{\widehat{AUC}}{2 - \widehat{AUC}} - \widehat{AUC}^2$, $q_2 = \frac{2\widehat{AUC}^2}{1 + \widehat{AUC}} - \widehat{AUC}^2$, n_1 is the number of true positives, and n_2 is the number of true negatives.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fgene.2021.667074

Identifying Gene–Environment Interactions With Robust Marginal Bayesian Variable Selection

Xi Lu[1], *Kun Fan*[1], *Jie Ren*[2] *and Cen Wu*[1]*

[1] Department of Statistics, Kansas State University, Manhattan, KS, United States, [2] Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, United States

In high-throughput genetics studies, an important aim is to identify gene–environment interactions associated with the clinical outcomes. Recently, multiple marginal penalization methods have been developed and shown to be effective in G×E studies. However, within the Bayesian framework, marginal variable selection has not received much attention. In this study, we propose a novel marginal Bayesian variable selection method for G×E studies. In particular, our marginal Bayesian method is robust to data contamination and outliers in the outcome variables. With the incorporation of spike-and-slab priors, we have implemented the Gibbs sampler based on Markov Chain Monte Carlo (MCMC). The proposed method outperforms a number of alternatives in extensive simulation studies. The utility of the marginal robust Bayesian variable selection method has been further demonstrated in the case studies using data from the Nurse Health Study (NHS). Some of the identified main and interaction effects from the real data analysis have important biological implications.

Keywords: gene-environment interaction, marginal analysis, robust Bayesian variable selection, spike-and-slab priors, markov chain monte carlo method

OPEN ACCESS

Edited by:
Qi Yan,
Columbia University, United States

Reviewed by:
Rong Zhang,
Amgen, United States
Zilin Li,
Harvard University, United States

***Correspondence:**
Cen Wu
wucen@ksu.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *11 February 2021*
Accepted: *13 July 2021*
Published: *08 December 2021*

Citation:
Lu X, Fan K, Ren J and Wu C (2021) Identifying Gene–Environment Interactions With Robust Marginal Bayesian Variable Selection. Front. Genet. 12:667074. doi: 10.3389/fgene.2021.667074

1. INTRODUCTION

The risk and progression of complex diseases including cancer, asthma and type 2 diabetes are associated with the coordinated functioning of genetic factors, the environmental (and clinical) factors, as well as their interactions (Hunter, 2005; Von Mutius, 2009; Cornelis and Hu, 2012; Simonds et al., 2016). The identification of important gene–environment(G×E) interactions leads to novel insight in dissecting the genetic basis of complex diseases in addition to the main effects of genetic and environmental factors. In the last two decades, searching for the important G×E interactions has been extensively conducted based on genetic association studies (Cordell and Clayton, 2005; Wu et al., 2012). One representative example is the genome-wide association study (GWAS), where the statistical significance of interaction between the environmental exposure and the genetic variant has been marginally assessed one at a time across the whole genome. Important findings are evidenced by genome-wide significant p-values after adjusting for multiple comparisons.

Recently, substantial efforts have been devoted to novel penalized variable selection methods for G×E studies (Zhou et al., 2021). In particular, marginal penalization has achieved very competitive performances with the aforementioned significance-based G×E analysis (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020). For example, within the framework of maximum rank correlation, Shi et al. (2014) has developed a penalization method robust to outliers and model

misspecification in determining important G×E interactions one at a time. Zhang et al. (2020) has imposed hierarchical structure between the main effects and interactions in marginal identification of G×E interactions using regularization. Despite success, these studies have limitations. First, as a common tuning parameter is demanded for all the marginal models, its selection requires pooling all genes together to conduct a joint model-based cross-validation. While such a strategy is not rare, it seems not in favor of the marginal nature of the proposed G×E studies. Second, a rigorous measure to quantify uncertainty is not available. Zhang et al. (2020) has constructed 95% confidence intervals based on the observed occurrence index (OOI) values (Huang and Ma, 2010); nevertheless, this measure has been used to demonstrate stability of identified effects rather than quantifying uncertainty of penalized estimates.

These limitations have motivated us to consider Bayesian analyses. In literature, Bayesian variable selection methods have been developed for G×E analysis in multiple studies (Zhou et al., 2021). For example, with indicator model selection, Liu et al. (2015) has imposed hierarchical Bayesian variable selection for linear G×E interactions. Li et al. (2015) has proposed a Bayesian group LASSO to identify non-linear interactions in nonparametric varying coefficient models. Ren et al. (2020) has further incorporated selection of linear and nonlinear G×E interactions simultaneously while accounting for structured identification in the Bayesian adaptive shrinkage framework. All these fully Bayesian methods can efficiently provide uncertainty quantification based on the posterior samples from MCMC. Nevertheless, our limited literature mining shows that none of the marginal Bayesian variable selection methods have been proposed for interaction studies so far.

Historically, marginal analysis has prevailed in G×E interaction studies within the framework of genetic association studies. Although recent studies have confirmed the utility of regularized variable selection in joint G×E analysis, more efforts are needed for marginal penalizations, especially through the Bayesian point of view. The step toward marginal Bayesian variable selection is of particular significance in developing a coherent framework of analyzing G×E interactions.

Here, we propose a novel marginal Bayesian variable selection method for the robust identification of G×E interactions. As heavy-tailed distributions and outliers in the response variable have been widely observed, robust modeling is essential for yielding reliable results. Specifically, the robustness of the proposed method is facilitated by the Bayesian formulation of the least absolute deviation (LAD) regression, which has been a popular choice in frequentist G×E studies but seldom investigated in a similar context from the Bayesian perspective. We consider the Bayesian LAD LASSO for regularized identification of interaction effects. As Bayesian LAD LASSO does not lead to zero coefficients, the spike-and-slab priors (George and McCulloch, 1993; Ishwaran and Rao, 2005) has been incorporated to impose exact sparsity in the adaptive shrinkage framework. The corresponding MCMC algorithm has been developed to accommodate fast computations. We have demonstrated the advantage of the proposed robust Bayesian marginal analysis in simulation. The findings from the case study of the Nurses' Health Study (NHS) with SNP measurements have important biological implications.

2. METHOD

We use Y to denote a continuous response variable representing the cancer outcome or disease phenotype. Let $X = (X_1, \ldots, X_p)$ be the p genetic variants, $E = (E_1, \ldots, E_q)$ be the q environmental factors and $C = (C_1, \ldots, C_m)$ be the m clinical factors. We denote the ith subject with i. Let (Y_i, E_i, C_i, X_i) $(i = 1, \ldots, n)$ be independent and identically distributed random vectors. For X_{ij} $(j = 1, \ldots, p)$, the measurement of the jth genetic factor on the ith subject considers the following marginal model:

$$\begin{aligned} Y_i &= \sum_{k=1}^{q} \alpha_k E_{ik} + \sum_{t=1}^{m} \gamma_t C_{it} + \beta_j X_{ij} + \sum_{k=1}^{q} \eta_{jk} X_{ij} E_{ik} + \epsilon_i \\ &= \sum_{k=1}^{q} \alpha_k E_{ik} + \sum_{t=1}^{m} \gamma_t C_{it} + \beta_j X_{ij} + \eta_j \tilde{W}_i + \epsilon_i, \end{aligned} \tag{1}$$

where α_k's and γ_t's are the regression coefficients corresponding to effects of environmental and clinical factors, respectively. For the jth gene X_j $(j = 1, \ldots, p)$, the G×E interactions effects are defined with $W_j = (X_j E_1, \ldots, X_j E_q)$, $\eta_j = (\eta_{j1}, \ldots, \eta_{jq})^T$. With a slight abuse of notation, denote $\tilde{W} = W_j$. The β_j's and η_{jk}'s are the regression coefficients of the genetic variants and G×E interactions effects, correspondingly. Let us denote $\alpha = (\alpha_1, \ldots, \alpha_q)^T$ and $\gamma = (\gamma_1, \ldots, \gamma_m)^T$. Then model (1) can be written as:

$$Y_i = E_i \alpha + C_i \gamma + X_{ij} \beta_j + \tilde{W}_i \eta_j + \epsilon_i. \tag{2}$$

2.1. Bayesian Formulation of the LAD Regression

The necessity of accounting for robustness in interaction studies has been increasingly recognized (Zhou et al., 2021). Within the frequentist framework, it is essentially dependent on adopting a robust loss function to quantify lack of fit (Wu and Ma, 2015). Among a variety of popular robust losses, the least absolute deviation (LAD) loss function is well known for its advantages in dealing with heavy-tailed error distributions or outliers in response. The estimation of regression coefficients amounts to the following minimization problem:

$$\min_{\alpha, \gamma, \beta_j, \eta_j} \sum_{i=1}^{n} |Y_i - E_i \alpha - C_i \gamma - X_{ij} \beta_j - \tilde{W}_i \eta_j|.$$

Here, we propose the robust marginal Bayesian variable selection based on LAD. As the Laplace distribution is equivalent to the mixture of an exponential distribution and a scaled normal distribution (Kozumi and Kobayashi, 2011), for a Bayesian formulation of LAD regression, we assume that $\epsilon_i (i = 1, \ldots, n)$ are i.i.d. random variables following the Laplace distribution with density:

$$f(\epsilon_i | \tau) = \frac{\tau}{2} \exp(-\tau |\epsilon_i|),$$

where τ is the inverse of the scale parameters from the Laplace density. Then the likelihood function of our marginal G×E model can be expressed as:

$$f(Y|\alpha,\gamma,\beta_j,\eta_j)=\prod_{i=1}^{n}\frac{\tau}{2}\exp(-\tau|Y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j|).$$

The above formulation using Laplace distribution is a special case of the asymmetric Laplace distribution, which has been widely adopted in Baysian quantile regression (Yu and Moyeed, 2001; Yu and Zhang, 2005). In Baysian quantile regression, ϵ_i's are assumed to follow the skewed Laplace distribution with density

$$f(\epsilon|\tau)=\theta(1-\theta)\tau\exp(-\tau\rho_\theta(\epsilon)),$$

where $\rho_\theta(\epsilon)=\epsilon\{\theta-I(\epsilon<0)\}$ is the check loss function. The random errors can be written as

$$\epsilon_i=\xi_1\nu_i+\tau^{-1/2}\xi_2\sqrt{\nu_i}z_i,$$

where

$$\xi_1=\frac{1-2\theta}{\theta(1-\theta)}\quad and\quad \xi_2=\sqrt{\frac{2}{\theta(1-\theta)}}$$

with quantile level $\theta\in(0,1)$, $\nu_i\sim\exp(\tau^{-1})$, and $z_i\sim\mathrm{N}(0,1)$.

The Bayesian LAD regression is a special case of Bayesian quantile regression (Li et al., 2010) with θ=0.5, resulting in that $\xi_1=0$ and $\xi_2=\sqrt{8}$. Therefore, the response Y_i can be written as:

$$\begin{aligned}Y_i&=\mu_i+\tau^{-1/2}\xi_2\sqrt{\nu_i}z_i,\\ \nu_i|\tau&\overset{iid}{\sim}\tau\exp(-\tau\nu_i),\\ z_i&\overset{iid}{\sim}\mathrm{N}(0,1),\end{aligned}\tag{3}$$

where $\mu_i=E_i\alpha+C_i\gamma+X_{ij}\beta_j+\tilde{W}_i\eta_j$.

2.2. Bayesian LAD LASSO With Spike-and-Slab Priors

In model (1), the coefficients β_j and η_j correspond to the main and interaction effects with respect to the jth genetic variant, respectively. When $\beta_j=0$ and $\eta_j=0$, the genetic variant has no effect on the phenotype. A non-zero β_j suggests the presence of main genetic effect. For η_j, if at least one of its component is not zero, then the G×E interaction effect exists. In literature, Bayesian quantile LASSO, with Bayesian LAD LASSO as its special case, has been proposed to conduct variable selection (Li et al., 2010). However, a major limitation is that Bayesian quantile LASSO cannot shrink regression coefficients to 0 exactly, resulting in inaccurate identification and biased estimation. To overcome such a limitation, we incorporate spike-and-slab priors to impose sparsity within Bayesian LAD LASSO framework as follows.

For the jth gene ($j=1,\ldots,p$), the marginal LAD LASSO model is given by

$$\sum_{i=1}^{n}|Y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j|+\lambda_1|\beta_j|+\lambda_2\sum_{k=1}^{q}|\eta_{jk}|.$$

Let $\varphi_1=\tau\lambda_1$ and $\varphi_2=\tau\lambda_2$. Then the conditional Laplace prior on the coefficient of main effect β_j can be expressed as scale mixtures of normals:

$$\begin{aligned}\pi(\beta_j|\tau,\lambda_1)&=\frac{\varphi_1}{2}\exp\{-\varphi_1|\beta_j|\}\\&=\int_0^\infty\frac{1}{\sqrt{2\pi s_1}}\exp(-\frac{\beta_j^2}{2s_1})\frac{\varphi_1^2}{2}\exp(\frac{-\varphi_1^2}{2}s_1)ds_1.\end{aligned}$$

The conditional Laplace prior on the coefficients of interaction effect η_j can be written as:

$$\begin{aligned}\pi(\eta_j|\tau,\lambda_2)&=\prod_{k=1}^{q}\frac{\varphi_2}{2}\exp\{-\varphi_2|\eta_{jk}|\}\\&=\prod_{k=1}^{q}\int_0^\infty\frac{1}{\sqrt{2\pi s_2}}\exp(-\frac{\eta_{jk}^2}{2s_2})\frac{\varphi_2^2}{2}\exp(\frac{-\varphi_2^2}{2}s_2)ds_2.\end{aligned}$$

Therefore, we consider the following hierarchical formulation for the marginal G×E model:

$$\begin{aligned}\beta_j|s_1,\pi_1&\sim(1-\pi_1)\mathrm{N}(0,s_1)+\pi_1\delta_0(\beta_j),\\ s_1|\varphi_1^2&\sim\frac{\varphi_1^2}{2}\exp(-\frac{\varphi_1^2}{2}s_1),\\ \eta_{jk}|s_{2k},\pi_2&\overset{iid}{\sim}(1-\pi_2)\mathrm{N}(0,s_{2k})+\pi_2\delta_0(\eta_{jk})(k=1,\ldots,q),\\ s_{2k}|\varphi_2^2&\overset{iid}{\sim}\frac{\varphi_2^2}{2}\exp(-\frac{\varphi_2^2}{2}s_{2k})(k=1,\ldots,q),\end{aligned}\tag{4}$$

where $\delta_0(\beta_j)$ and $\delta_0(\eta_{jk})$ denote the spike at 0, respectively, and the slab distributions are represented by two normal distributions, $\mathrm{N}(0,s_1)$ and $\mathrm{N}(0,s_2k)$. Here, $\pi_1\in[0,1]$ and $\pi_2\in[0,1]$. The mixture of the spike and slab components facilitate the selection of main and interaction effects. Instead of setting π_1 and π_2 to a fixed value such as 0.5, we assign conjugate beta priors on them as $\pi_1\sim\mathrm{Beta}(r_1,u_1)$ and $\pi_2\sim\mathrm{Beta}(r_2,u_2)$, which account for the uncertainty in π_1 and π_2. In this paper, we choose $r_1=u_1=r_2=u_2=1$ as it gives a prior mean with 0.5 and it also allows a prior to spread out.

In addition, the normal prior has been placed on the coefficients of environmental factor $\alpha_k(k=1,\ldots,q)$ and clinical factor $\gamma_t(t=1,\ldots,m)$ as:

$$\begin{aligned}\alpha_k&\overset{iid}{\sim}\frac{1}{\sqrt{(2\pi\alpha_0)}}\exp(-\frac{\alpha_k^2}{2\alpha_0})(k=1,\ldots,q)\\ \gamma_t&\overset{iid}{\sim}\frac{1}{\sqrt{(2\pi\gamma_0)}}\exp(-\frac{\gamma_t^2}{2\gamma_0})(t=1,\ldots,m),\end{aligned}$$

We also assume conjugate Gamma priors on τ, φ_1^2 and φ_2^2 with

$$\begin{aligned}\tau&\sim\mathrm{Gamma}(a,b),\\ \varphi_1^2&\sim\mathrm{Gamma}(c_1,d_1),\\ \varphi_2^2&\sim\mathrm{Gamma}(c_2,d_2).\end{aligned}$$

In typical G×E studies, the environmental and clinical factors are of low dimensionality and the selection of them is not of interest.

Therefore, the sparsity-inducing priors have not been adopted for these factors. We consider the Bayesian LAD LASSO type of regularization in the proposed study as published studies have demonstrated that baseline penalty such as MCP and LASSO work well for marginal variable selection (Shi et al., 2014; Chai et al., 2017).

It is noted that Zhang et al. (2020) has proposed a marginal sparse group MCP to respect the strong hierarchy between main and interaction effects. Their results are promising when long tailed distributions and outliers are not present in the response variable. Although sparse group (or, bi-level) variable selection has been demonstrated as being very effective in multiple G×E studies based on joint models (Zhou et al., 2021), in our study, there is only one group per each marginal model. The sparse group no longer has significant advantages over individual level selection. Therefore, it has not been considered here.

Our model respects the weak hierarchy of "main effects, interactions." If imposing the strong hierarchy is needed, the genetic factor, once it is not selected given the presence of corresponding interaction effects, can be added back to the identified marginal model for a refit to impose strong hierarchy (Chai et al., 2017). While such a practice is not uncommon in marginal interaction studies, Shi et al. (2014) has also revealed satisfactory performance when strong hierarchy has not been pursued.

2.3. The Gibbs Sampler for Robust Marginal G×E Analysis

For the jth genetic factor, the joint posterior distribution of all the unknown parameters conditional on data can be expressed as

$$
\begin{aligned}
&\pi(\alpha,\gamma,\beta_j,\eta_j,\nu,s_1,s_2,\tau,\varphi_1,\varphi_2,\pi_1,\pi_2,z_i|Y)\\
&\propto\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2\nu_i}}\exp\Big\{-\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2\nu_i}\Big\}\\
&\times\prod_{i=1}^{n}\tau\exp(-\tau\nu_i)\tau^{a-1}\exp(-b\tau)\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z_i^2)\\
&\times\prod_{k=1}^{q}\frac{1}{\sqrt{(2\pi\alpha_0)}}\exp(-\frac{\alpha_k^2}{2\alpha_0})\\
&\times\prod_{t=1}^{m}\frac{1}{\sqrt{(2\pi\gamma_0)}}\exp(-\frac{\gamma_t^2}{2\gamma_0})\\
&\times\Big((1-\pi_1)(2\pi s_1)^{-1/2}\exp(-\frac{\beta_j^2}{2s_1})\mathbf{I}_{\{\beta_j\neq0\}}+\pi_1\delta_0(\beta_j)\Big)\\
&\times\prod_{k=1}^{q}\Big((1-\pi_2)(2\pi s_{2k})^{-1/2}\exp(-\frac{\eta_{jk}^2}{2s_{2k}})\mathbf{I}_{\{\eta_{jk}\neq0\}}+\pi_2\delta_0(\eta_{jk})\Big)\\
&\times\frac{\varphi_1^2}{2}\exp(-\frac{\varphi_1^2}{2}s_1)\\
&\times\prod_{k=1}^{q}\frac{\varphi_2^2}{2}\exp(-\frac{\varphi_2^2}{2}s_{2k})\\
&\times(\varphi_1^2)^{c_1-1}\exp(-d_1\varphi_1^2)\\
&\times(\varphi_2^2)^{c_2-1}\exp(-d_2\varphi_2^2)\\
&\times\pi_1^{r_1-1}(1-\pi_1)^{u_1-1}\\
&\times\pi_2^{r_2-1}(1-\pi_2)^{u_2-1}
\end{aligned}
$$

Let $\mu_{(-\alpha_k)}=E(y_i)-E_{ik}\alpha_k,(i=1,\ldots,n),(k=1,\ldots,q)$, representing the mean effect without the contribution of $E_{ik}\alpha_k$. The posterior distribution of the coefficient of environmental factor α_k conditional on all other parameters can be expressed as:

$$
\begin{aligned}
&\pi(\alpha_k|\text{rest})\\
&\quad\propto\pi(\alpha_k)\pi(Y|\cdot)\\
&\quad\propto\exp\Big\{-\sum_{i=1}^{n}\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2\nu_i}\Big\}\\
&\quad\times\exp(-\frac{\alpha_k^2}{2\alpha_0})\\
&\quad\propto\exp\Big\{-\frac{1}{2}\big[(\sum_{i=1}^{n}\frac{\tau E_{ik}^2}{\xi_2^2\nu_i}+\frac{1}{\alpha_0})\alpha_k^2\\
&\quad-2\sum_{i=1}^{n}\frac{\tau(y_i-\mu_{(-\alpha_k)})E_{ik}}{\xi_2^2\nu_i}\alpha_k\big]\Big\}.
\end{aligned}
$$

Hence, the full conditional distribution of α_k is normal distribution $N(\mu_{\alpha_k},\sigma_{\alpha_k}^2)$with mean

$$\mu_{\alpha_k}=(\sum_{i=1}^{n}\frac{\tau(y_i-\mu_{(-\alpha_k)})E_{ik}}{\xi_2^2\nu_i})\sigma_{\alpha_k}^2,$$

and variance

$$\sigma_{\alpha_k}^2=(\sum_{i=1}^{n}\frac{\tau E_{ik}^2}{\xi_2^2\nu_i}+\frac{1}{\alpha_0})^{-1}.$$

The posterior distribution of the coefficient of clinical factor $\gamma_t(t=1,\ldots,m)$ conditional on all other parameters can be obtained in similar way. Let $\mu_{(-\gamma_t)}=E(y_i)-C_{it}\gamma_t$, $i=1,\ldots,n$, then

$$\gamma_t|\text{rest}\sim N(\mu_{\gamma_k},\sigma_{\gamma_t}^2),$$

where

$$
\begin{aligned}
\mu_{\gamma_t}&=(\sum_{i=1}^{n}\frac{\tau(y_i-\mu_{(-\gamma_t)})C_{it}}{\xi_2^2\nu_i})\sigma_{\gamma_t}^2,\\
\sigma_{\gamma_t}^2&=(\sum_{i=1}^{n}\frac{\tau C_{it}^2}{\xi_2^2\nu_i}+\frac{1}{\gamma_0})^{-1}.
\end{aligned}
$$

Let $\mu_{(-\beta_j)}=E(y_i)-X_{ij}\beta_j$ and $l_1=\pi(\beta_j=0|\text{rest})$, the conditional posterior distribution of the coefficient of genetic factor β_j is a spike-and-slab distribution:

$$\beta_j|\text{rest}\sim(1-l_1)N(\mu_{\beta_j},\sigma_{\beta_j}^2)+l_1\delta_0(\beta_j),\tag{5}$$

where

$$\mu_{\beta_j} = \Big(\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\beta_j)})X_{ij}}{\xi_2^2 \nu_i}\Big)\sigma^2_{\beta_j},$$
$$\sigma^2_{\beta_j} = \Big(\sum_{i=1}^{n} \frac{\tau X_{ij}^2}{\xi_2^2 \nu_i} + \frac{1}{s_1}\Big)^{-1}.$$

We can show that

$$l_1 = \frac{\pi_1}{\pi_1 + (1-\pi_1)s_1^{-1/2}(\sigma^2_{\beta_j})^{1/2}\exp\{\frac{1}{2}(\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\beta_j)})X_{ij}}{\xi_2^2 \nu_i})^2 \sigma^2_{\beta_j}\}}.$$

The posterior distribution of β_j is a mixture of a normal distribution and a point mass at 0. That is, at each iteration of MCMC, β_j is drawn from $N(\mu_{\beta_j}, \sigma^2_{\beta_j})$ with probability $(1 - l_1)$ and is set to 0 with probability l_1.

Similarly, the posterior distribution of the interaction of the jth gene and environmental factors $\eta_{jk}(k = 1, \ldots, q)$ is also a spike-and-slab distribution. Denote $\mu_{(-\eta_{jk})} = E(y_i) - W_{ik}\eta_{jk}$ and $l_{2k} = \pi(\eta_{jk} = 0|\text{rest})$, η_{jk} follows this distribution:

$$\eta_{jk}|\text{rest} \sim (1 - l_{2k})N(\mu_{\eta_{jk}}, \sigma^2_{\eta_{jk}}) + l_{2k}\delta_0(\eta_{jk}), \tag{6}$$

where

$$\mu_{\eta_{jk}} = \Big(\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\eta_{jk})})\tilde{W}_{ik}}{\xi_2^2 \nu_i}\Big)\sigma^2_{\eta_{jk}},$$
$$\sigma^2_{\beta_j} = \Big(\sum_{i=1}^{n} \frac{\tau \tilde{W}_{ik}^2}{\xi_2^2 \nu_i} + \frac{1}{s_{2k}}\Big)^{-1}.$$

And

$$l_{2k} = \frac{\pi_2}{\pi_2 + (1-\pi_2)s_{2k}^{-1/2}(\sigma^2_{\eta_{jk}})^{1/2}\exp\{\frac{1}{2}(\sum_{i=1}^{n} \frac{\tau(y_i - \mu_{(-\eta_{jk})})\tilde{W}_{ik}}{\xi_2^2 \nu_i})^2 \sigma^2_{\eta_{jk}}\}}. \tag{7}$$

The full conditional posterior distribution of s_1 is:

$$\begin{aligned} &s_1|\text{rest} \\ &\propto \pi(\beta_j|s_1, \pi_1)\pi(s_1|\varphi_1^2) \\ &\propto \Big((1-\pi_1)(2\pi s_1)^{-1/2}\exp(-\frac{\beta_j^2}{2s_1})\mathbf{I}_{\{\beta_j \neq 0\}} \\ &+ \pi_1\delta_0(\beta_j)\Big)\exp(-\frac{\varphi_1^2}{2}s_1). \end{aligned} \tag{8}$$

When $\beta_j = 0$, equation (8) is proportional to $\exp(-\frac{\varphi_1^2}{2}s_1)$. Therefore, the posterior distribution of s_1 is $\exp(\frac{\varphi_1^2}{2})$.

When $\beta_j \neq 0$, equation (8) is proportional to

$$\begin{aligned} &\frac{1}{\sqrt{s_1}}\exp(-\frac{\varphi_1^2}{2}s_1)\exp(-\frac{\beta_j^2}{2s_1}) \\ &\propto \frac{1}{\sqrt{s_1}}\exp\Big\{-\frac{1}{2}[\varphi_1^2 s_1 + \frac{\beta_j^2}{s_1}]\Big\}. \end{aligned}$$

Therefore, when $\beta_j \neq 0$, the posterior distribution for s_1^{-1} is Inverse-Gaussian$(\sqrt{\frac{\varphi_1^2}{\beta_j^2}}, \varphi_1^2)$.

Similarly, for $s_{2k}(k = 1, \ldots, q)$, when $\eta_{jk} = 0$, the posterior distribution of s_{2k} is $\exp(\frac{\varphi_2^2}{2})$. When $\eta_{jk} \neq 0$, the posterior distribution for s_{2k}^{-1} is Inverse-Gaussian$(\sqrt{\frac{\varphi_2^2}{\eta_{jk}^2}}, \varphi_2^2)$.

The full conditional posterior distribution of φ_1^2:

$$\begin{aligned} &\varphi_1^2|\text{rest} \\ &\propto \pi(s_1|\varphi_1^2)\pi(\varphi_1^2) \\ &\propto \frac{\varphi_1^2}{2}\exp(-\frac{\varphi_1^2 s_1}{2})(\varphi_1^2)^{c_1 - 1}\exp(-d_1\varphi_1^2) \\ &\propto (\varphi_1^2)^{c_1}\exp\Big(-\varphi_1^2(s_1/2 + d_1)\Big). \end{aligned}$$

Therefore, the posterior distribution for φ_1^2 is Gamma $(c_1 + 1, s_1/2 + d_1)$. Similarly, the posterior distribution for φ_2^2 is Gamma $(c_2 + q, \sum_{k=1}^{q} s_{2k}/2 + d_2)$.

The full conditional posterior distribution of π_1 is given as:

$$\begin{aligned} &\pi_1|\text{rest} \\ &\propto \pi(s_1|\varphi_1^2)\pi(\varphi_1^2) \\ &\propto \pi_1^{r_1 - 1}(1-\pi_1)^{u_1 - 1} \\ &\times \Big((1-\pi_1)(2\pi s_1)^{-1/2}\exp(-\frac{\beta_j^2}{2s_1})\mathbf{I}_{\{\beta_j \neq 0\}} + \pi_1\delta_0(\beta_j)\Big). \end{aligned}$$

Then, the posterior distribution for π_1 is Beta $(1 + r_1 - \mathbf{I}(\beta_j \neq 0), u_1 + \mathbf{I}(\beta_j \neq 0))$.

The full conditional posterior distribution of π_2 is given as:

$$\begin{aligned} &\pi_2|\text{rest} \\ &\propto \pi(s_2|\varphi_2^2)\pi(\varphi_2^2) \\ &\propto \pi_2^{r_2 - 1}(1-\pi_2)^{u_2 - 1} \\ &\times \prod_{k=1}^{q}\Big((1-\pi_2)(2\pi s_{2k})^{-1/2}\exp(-\frac{\eta_{jk}^2}{2s_{2k}})\mathbf{I}_{\{\eta_{jk} \neq 0\}} + \pi_2\delta_0(\eta_{jk})\Big). \end{aligned}$$

So, the posterior distribution for π_2 is Beta $(1 + r_1 - \sum_{k=1}^{q}\mathbf{I}(\eta_{jk} \neq 0), u_1 + \sum_{k=1}^{q}\mathbf{I}(\eta_{jk} \neq 0))$.

The full conditional posterior distribution of τ is given as:

$$\begin{aligned} &\tau|\text{rest} \\ &\propto \pi(\nu|\tau)\pi(\tau)\pi(Y|\cdot) \\ &\propto \tau^{n/2}\exp\Big\{-\sum_{i=1}^{n}\frac{(y_i - E_i\alpha - C_i\gamma - X_{ij}\beta_j - \tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2\nu_i}\Big\} \\ &\times \tau^n\exp(-\tau\sum_{i=1}^{n}\nu_i)\tau^{a-1}\exp(-b\tau) \\ &\propto \tau^{a + \frac{3}{2}n - 1}\exp\Big\{-\tau\Big[\sum_{i=1}^{n}(\frac{(y_i - E_i\alpha - C_i\gamma - X_{ij}\beta_j - \tilde{W}_i\eta_j)^2}{2\xi_2^2\nu_i} \\ &+ \nu_i) + b\Big]\Big\}. \end{aligned}$$

Therefore, the posterior distribution for τ is Gamma$(a + \frac{3}{2}n, \left[\sum_{i=1}^{n}(\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\xi_2^2\nu_i}+\nu_i)+b\right])$.

Last, we have the full conditional posterior distribution of ν_i:

$$\begin{aligned}
&\nu_i|\text{rest}\\
&\propto \pi(\nu|\tau)\pi(Y|\cdot)\\
&\propto \frac{1}{\sqrt{\nu_i}}\exp\Big\{-\frac{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{2\tau^{-1}\xi_2^2\nu_i}\Big\}\\
&\times\exp(-\tau\nu_i)\\
&\propto \frac{1}{\sqrt{\nu_i}}\exp\Big\{-\frac{1}{2}\Big[(2\tau)\nu_i\\
&+\frac{\tau(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}{\xi_2^2\nu_i}\Big]\Big\}.
\end{aligned}$$

It is easy to show that

$$\frac{1}{\nu_i}|\text{rest}\sim \text{Inverse-Gaussian}\left(\sqrt{\frac{2\xi_2^2}{(y_i-E_i\alpha-C_i\gamma-X_{ij}\beta_j-\tilde{W}_i\eta_j)^2}},\ 2\tau\right).$$

The spirit of marginal penalization for G×E interactions lies in the usage of a common sparsity cutoff to determine a list of important main and interaction effects. Instead of focusing on a fixed cutoff, varying the cutoff can generate different lists, resulting in a comprehensive view of important findings. The tuning parameter in penalized estimation serves as the cutoff. Therefore, the same tuning parameter has to be adopted for all the sub-models (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020). To further justify such a common tuning parameter, Zhang et al. (2020) has attempted using the joint model to select the common tuning through cross-validation. However, this seems not coherent with the nature of marginal analysis.

Ideally, the tuning parameter should be determined by each model itself to allow for flexibility in controlling sparsity individually, and a common cutoff is still available to examine different lists of important effects. With the Bayesian formulation, we can avoid such a limitation of frequentist marginal penalization methods. In particular, the priors have been placed on regularization parameters to determine the sparsity in a data-driven manner for each sub-model. With the spike-and-slab priors, the posterior distributions on the coefficients of main and interaction effects naturally lead to the usage of inclusion probability as a common cutoff to pin down the list of important effects, which is described in detail in the next section.

3. SIMULATION

To demonstrate the utility of the proposed approach, we evaluate the performance through simulation study. In particular, we compare the performance of the proposed method, LAD Bayesian Lasso with spike-and-slab priors (denoted as LADBLSS) with three alternatives, LAD Bayesian Lasso (denoted as LADBL), Bayesian Lasso with spike-and-slab priors (denoted as BLSS) and Bayesian Lasso (denoted as BL). LADBL is similar to the proposed method, except that it does not adopt the spike-and-slab prior. The details of posterior inference are given in the **Appendix**.

Under all settings, the sample size is set as $n = 200$, and the number of G factors is $p = 500$ with $q = 4$, $m = 3$. For environmental factors, we simulate four continuous variables from multivariate normal distributions with marginal mean 0, marginal variance 1 and AR1 correlation structure with $\rho = 0.5$. In addition, three clinical factors are generated from a multivariate normal distribution with marginal mean 0 and marginal variance 1 and AR1 structure with $\rho = 0.5$. Among the p main G effects and pq G×E interactions, 8 and 12 effects are set as being associated with the response, respectively. All the environmental and clinical factors are important with nonzero coefficients, which are randomly generated from a uniform distribution Unif[0.1, 0.5]. The random error are generated from: (1) N(0, 1)(Error 1), (2) t-distribution with 2 degrees of freedom ($t(2)$) (Error2), (3) LogNormal(0,2)(Error3), (4) 90%N(0,1)+10%Cauchy(0,1)(Error4), (5) 80%N(0,1)+20%Cauchy(0,1)(Error5). All of them are heavy-tailed distribution except the first one.

In addition, the genetic factors are simulated in the following four settings.

Setting 1: In simulating continuous genetic variants, we generate multivariate normal distributions with marginal mean 0 and variance 1. The AR structure is considered in computing the correlation of G factors, under which gene j and k have correlation $\rho^{|j-k|}$ with $\rho = 0.5$.

Setting 2: We assess the performance under single-nucleotide polymorphism (SNP) data. The SNPs are obtained by dichotomizing the gene expression values at the 1st and 3rd quartiles, with the 3 levels (0,1,2) for genotypes (aa, Aa, and AA). Here, the gene expressions are generated from the first setting.

Setting 3: Consider simulating the SNP data under a pairwise linkage disequilibrium (LD) structure. For the two minor alleles A and B of two adjacent SNPs, let q_1 and q_2 be the minor allele frequencies (MAFs). The frequencies of four haplotypes are as $p_{AB} = q_1q_2 + \delta$, $p_{ab} = (1-q_1)(1-q_2) + \delta$, $p_{Ab} = q_1(1-q_2) - \delta$, and $p_{aB} = (1-q_1)q_2 - \delta$, where δ denotes the LD. Assuming Hardy–Weinberg equilibrium and given the allele frequency for A at locus 1, we can generate the SNP genotype (AA, Aa, aa) from a multinomial distribution with frequencies $(q_1^2, 2q_1(1-q_1), (1-q_1)^2)$. Based on the conditional genotype probability matrix, we can simulate the genotypes for locus 2. With MAFs 0.3 and pairwise correlation $r = 0.6$, we have $\delta = r\sqrt{q_1(1-q_1)q_2(1-q_2)}$.

We collect the posterior samples from the Gibbs Sampler with 10,000 iterations and discard the first 5,000 samples as burn-ins. The posterior medians are used to estimate the coefficients. For approaches incorporating spike-and-slab priors, we consider computing the inclusion probability to indicate the importance of predictors. Here, we use a binary indicator ϕ to denote that the membership of the non-spike distribution. Take the main effect of the jth genetic factor, X_j, as an example. Suppose we

have collected H posterior samples from MCMC after burn-ins. The jth G factor is included in the marginal G×E model at the hth MCMC iteration if the corresponding indicator is 1, i.e., $\phi_j^{(h)} = 1$. Subsequently, the posterior probability of retaining the jth genetic main effect in the final marginal model is defined as the average of all the indicators for the jth G factor among the H posterior samples. That is,

$$p_j = \hat{\pi}(\phi_j = 1|y) = \frac{1}{H}\sum_{h=1}^{H}\phi_j^{(h)},\ j = 1, \ldots, p.$$

A larger posterior inclusion probability p_j indicates a stronger empirical evidence that the jth genetic main effect has a non-zero coefficient, i.e., a stronger association with the phenotypic trait.

To comprehensively assess the performance of the proposed and alternative methods, we consider a sequence of probabilities as cutting-offs in inclusion probability for methods with spike-and-slab priors. Given a cutoff probability, the main or interaction is included in the final marginal model if its posterior inclusion probability is larger than the cutoff, and is excluded otherwise. Provided with a sequence of cutting-off probabilities from small to large, we can investigate the set of identified effects and calculate the true/false positive rates (T/FPR) as the ground truth is known in simulation. For the sequence of cut-offs, we are able to compute the area under curve (AUC) as a comprehensive measure. Besides, for methods without spike-and-slab priors, the confidence level of the credible intervals can be adopted as the cut-off to compute TPR and FPRs. Therefore, all the methods under comparison can be evaluated on the same ground.

In addition, we also consider Top100, which is defined as the number of true signals when 100 important main effects (or interactions) are identified. For methods with spike-and-slab priors, 100 main effects or interactions are chosen with the highest inclusion probabilities. For methods without spike-and-slab priors, the indicators of all effects are computed for a sequence of credible levels. The top 100 main effects or interactions are chosen in terms of the highest average identification values.

Simulation results for the gene expression data in the first setting are tabulated in **Tables 1**, **2**. We can observe that the proposed method has the best performance among all approaches, especially when the response variable has heavy-tailed distributions. First, the performance of methods with spike-and-slab priors is consistently better than methods without spike-and-slab priors. For example, in **Table 1**, under error 3, the AUC of LADBLSS is 0.9558 (sd 0.0161), which is much larger than that of the robust method without spike-and-slab priors, i.e., 0.8432(sd 0.0115) from LADBL. Also, the AUC of robust methods is much larger than that of non-robust methods, especially in the presence of heavy-tailed errors. For instance, in the first setting under error 3, the AUC of LADBLSS is 0.9558 and the AUC of LADBL is 0.8432 while that of BLSS and BL is around 0.5. Similar advantageous performance can also be observed from the identification results with Top100. In **Table 2** under error 5, LADBLSS identifies 7.80 (sd 0.55) out of the 8 main effects and 10.53 (sd 1.36) out of the 12 interaction effects.

TABLE 1 | Simulation results of the first setting for BL (Bayesian LASSO), BLSS (Bayesian LASSO with spike-and-slab priors), LADBL (LAD Bayesian LASSO), and LADBLSS (LAD Bayesian LASSO with spike-and-slab priors).

		BL	BLSS	LADBL	LADBLSS
Error 1	AUC	0.9182	0.9901	0.9258	0.9887
N(0,1)	SD	0.0052	0.0021	0.0076	0.0026
Error 2	AUC	0.8332	0.9420	0.9004	0.9841
$t(2)$	SD	0.0107	0.0235	0.0078	0.0031
Error 3	AUC	0.5343	0.5473	0.8432	0.9558
Lognormal(0,2)	SD	0.0144	0.0576	0.0115	0.0161
Error 4	AUC	0.8221	0.9124	0.9222	0.9895
90%N(0,1) + 10%Cauchy(0,1)	SD	0.0212	0.0410	0.0071	0.0024
Error 5	AUC	0.7507	0.8431	0.9192	0.9904
80%N(0,1) + 20%Cauchy(0,1)	SD	0.0217	0.0633	0.0059	0.0018

AUC (mean of AUC) and SD (sd of AUC) based on 100 replicates. n = 200, p = 500, q = 4, and m = 3.

This is higher than the results of LADBL with 7.57 (sd 0.57) of main effects and 6.83 (sd 1.07) of interaction effects. Second, among all the methods with spike-and-slab priors, Bayesian LAD method with spike-and-slab priors has the best performance in all identification results. Under error 3, in **Table 1**, the AUC of LADBLSS is 0.9558(sd 0.0161) while the AUC of BLSS is 0.5473(sd 0.0576). Under error 4 in **Table 2**, LADBLSS identifies 7.77(sd 0.57) main effects and 10.67(sd 1.50) interaction effects while BLSS identifies 6.2(sd 2.62) main effects and 8.3(sd 3.98) interaction effects, respectively.

Similar patterns can be observed in Tables 4, 5 in **Appendix** for the second setting, and Tables 6, 7 in **Appendix** for the third setting in **Appendix**. We have also investigated the performance of when n = 2,000 under setting 1. While the difference among the 4 methods significantly diminishes with such a large sample size, we can still observe the superior performance of LADBLSS by using a shorter list of top ranked effects. The results are provided in the table from **Supplementary Material**. Overall, the advantages of conducting robust Bayesian G×E analysis using the proposed approach can be justified based on the results of comprehensive simulation studies. The convergence of the MCMC chains with the potential scale reduction factor (PSRF) (Brooks and Gelman, 1998) has been conducted. In this study, we use PSRF $\leq$ 1.1 (Gelman et al., 2004) as the cut-off point, which indicates that chains converge to a stationary distribution. The convergence of chains after burn-ins has been checked for all parameters with the value of PSRF <1.1. **Figure 1** shows the convergence pattern of PSRF for the main and interaction coefficients of the first genetic factors in Example 1 under error 3.

In simulation, the hyperparameters for the Gamma priors and Beta priors specified in section Bayesian LAD LASSO With Spike-and-slab Priors are set to 1. In addition, the initial values of the regression parameters are also set to 1. Based on our experiments, the results and convergence of the MCMC algorithm are not sensitive to the choice of these parameters. We have observed satisfactory convergence for all of our simulations. For one simulated dataset under the first setting with n = 200, p = 500

TABLE 2 | Identification results of the first setting with Top100 method for BL (Bayesian LASSO), BLSS (Bayesian LASSO with spike-and-slab priors), LADBL (LAD Bayesian LASSO) and LADBLSS (LAD Bayesian LASSO with spike-and-slab priors).

		Main	Interaction	Total
Error 1	BL	7.60(0.49)	6.80(1.6)	14.40(1.73)
N(0,1)	BLSS	7.80(0.41)	10.80(0.92)	18.60(1.13)
	LADBL	7.67(0.55)	6.53(1.85)	14.20(1.81)
	LADBLSS	7.76(0.5)	10.53(1.36)	18.30(1.49)
Error 2	BL	6.37(1.90)	3.90(2.07)	10.27(3.19)
t(2)	BLSS	6.33(1.63)	8.53(2.46)	14.87(3.71)
	LADBL	7.43(0.94)	5.80(1.71)	13.23(2.01)
	LADBLSS	7.53(0.51)	9.90(1.56)	17.43(1.76)
Error 3	BL	0.90(1.21)	0.50(0.97)	1.40(1.45)
Lognormal(0,2)	BLSS	0.73(0.94)	0.47(0.68)	1.20(1.35)
	LADBL	6.27(1.55)	3.67(1.94)	9.93(2.75)
	LADBLSS	6.10(1.37)	8.93(2.02)	15.03(3.09)
Error 4	BL	5.57(2.99)	3.63(2.53)	9.20(5.05)
90%N(0,1)	BLSS	6.20(2.62)	8.30(3.98)	14.50(6.39)
+10%Cauchy(0,1)	LADBL	7.77(0.43)	7.00(1.93)	14.77(1.81)
	LADBLSS	7.77(0.57)	10.67(1.50)	18.23(1.67)
Error 5	BL	5.07(2.89)	3.00(2.49)	8.07(5.01)
80%N(0,1)	BLSS	4.60(3.25)	5.70(4.23)	10.30(7.27)
+20%Cauchy(0,1)	LADBL	7.57(0.57)	6.83(1.07)	14.40(1.83)
	LADBLSS	7.80(0.55)	10.53(1.36)	18.33(1.69)

Mean(sd) based on 100 replicates. $n = 200$, $p = 500$, $q = 4$, and $m = 3$.

and standard normal error, the CPU time (in minutes) for fitting all the 500 marginal models through 10,000 MCMC iterations on a laptop with standard configurations are 1.27(BL), 1.75(BLSS), 6.16(LADBL), and 5.95 (LADBLSS) minutes, respectively. The source codes of implementing all the methods under comparison are included in the **Supplementary Material**.

4. REAL DATA ANALYSIS

In this study, we analyze the type 2 diabetes (T2D) data from Nurses' Health Study (NHS), which is a well-characterized cohort study of women with high-dimensional SNP data, as well as measurements on lifestyle and dietary factors. We consider SNPs on chromosome 10 to identify main and gene–environment interactions associated with weight, which is an important phenotypic trait related to type 2 diabetes. Here, weight is used as response and five environment factors, age (age), total physical activity (act), trans fat intake (trans), cereal fiber intake (ceraf), and reported high blood cholesterol (chol), are considered. Data are available on 3,391 subjects and 17,016 gene expressions after cleaning the raw data through matching phenotypes and genotypes and removing SNPs with MAF <0.05. A prescreening is done before downstream analysis. We use a marginal linear model with weight as response and age, act, trans, ceraf, and chol as environment factors. Note that 10,000 SNPs that have at least two main or interaction effects with $p < 0.05$ are kept.The scale of working data is generally not a major concern for marginal analysis, as the computation can be done in a highly parallel manner. Here, we focus on chromosome 10 which has been reported to harbor interesting genes in existing studies.

We use Top 100 method to identify 100 most important main and interaction effects. The proposed method LADBLSS identifies 20 main SNP effects and 80 gene–environment interactions, which are listed in Table 8 in **Appendix**. Our study provides crucial implications in identifying the important main and interactions of SNPs and its associations with weight. For example, three SNPs, rs17011106, rs4838643 and rs17011115, located within gene WDFY4 are identified. WDFY4 has been observed as an influential factor related to weight and obesity (Barclay et al., 2015; Martin et al., 2019). In addition, SNPs rs10994364, rs10821773, and rs10994308, located within gene ANK3, are identified with interacting environment factors age and chol. There are findings showing an association between ANK3 and higher systolic blood pressure (Ghanbari et al., 2014). Published studies have also shown that ANK3 is linked to pulmonary and renal hypertension (Ghanbari et al., 2014). Allele risk variants have been identified in ANK3, and these variants explain a proportion of the heritability of BD (bipolar disorder), which is associated with higher body mass index (BMI) and increased metabolic comorbidity and the genetic risk for BD relates to common genetic risk with T2D (Winham et al., 2014). Our proposed method identifies its interaction with chol, the high blood cholesterol. Data from several sources suggest that islet cholesterol metabolism contributes to the pathogenesis of T2D (Brunham et al., 2008). Furthermore, the SNP rs1244416, corresponding to gene ATP5C1, interacts with the reported high blood cholesterol. This gene has been found to be deregulated in T2D skeletal muscle through pathway-based microanalysis (Morrison et al., 2012). The interactions between SNP rs10857590 and trans fat intake has also been identified by using the proposed method. The SNP is within gene ARHGAP22, which has been investigated in Huang et ail. (2018). As a diabetic retinopathy (DR) susceptibility gene, the expression of ARHGAP22 is positively associated with endothelial progenitor cells (EPC) levels in T2D patients with DR.

Analysis with alternatives BL, BLSS, and LADBL has also been conducted. To compare the alternative methods with the proposed method, we provide the numbers of main effects and interactions identified by these methods with pairwise overlaps in **Table 3**. It clearly shows that the proposed one results in a very different set of effects compared to alternatives. We refit the regularized marginal models by LADBL and LADBLSS using robust Bayesian Lasso, and those identified by BL and BLSS using Bayesian Lasso. In addition, the inclusion probabilities of the selected main and interaction effects using LADBLSS are provided in Table 9 in **Appendix**. Results from the alternative methods are available from the **Supplementary Material**. The proposed method selects the 100 most important effects with the inclusion probability larger than 0.9, demonstrating its superiority in quantifying uncertain compared to marginal penalization methods (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020). We noticed

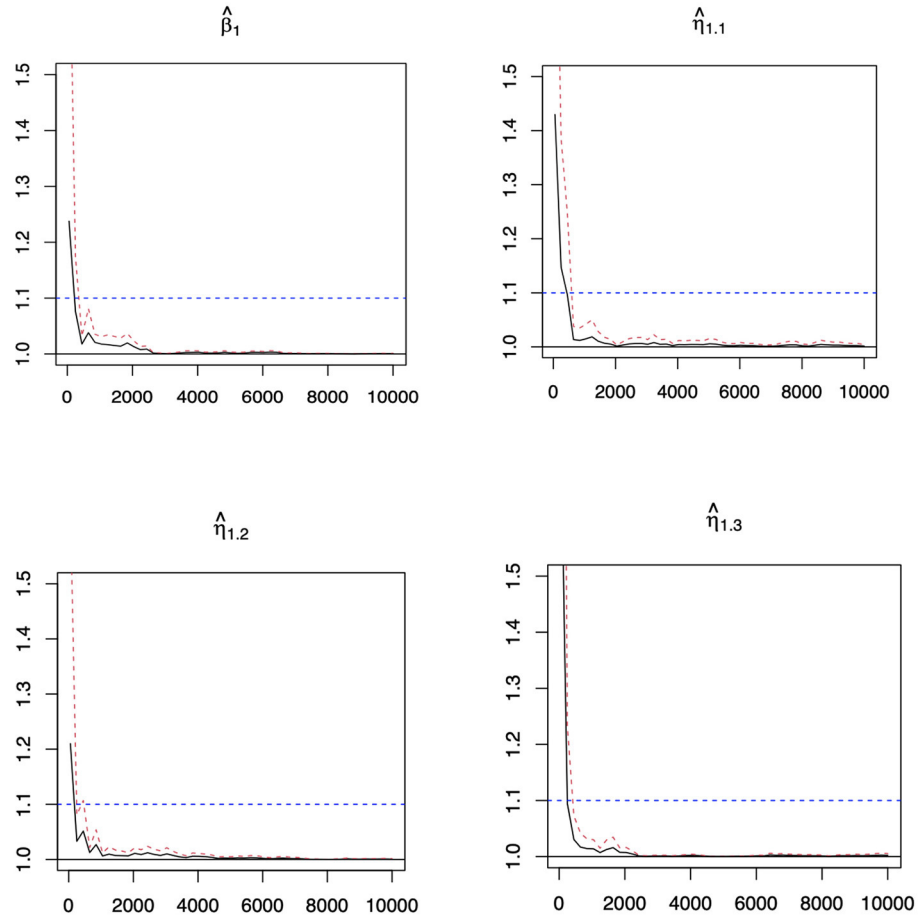


FIGURE 1 | Potential scale reduction factor (PSRF) against iterations for the coefficients of the first genetic factors and its interaction with environmental factors in Example 1 under error 3. Black line: the PSRF. Red dotted line: the upper limits of the 95% confidence interval for the PSRF. Blue dotted line: The threshold of 1.1. The $\hat{\beta}_1$ represents the estimated coefficients of the main effects for the first genetic factor. The $\hat{\eta}_{11}$ to $\hat{\eta}_{13}$ represent the estimated coefficients of the first three interaction effects for the first genetic factor.

the small magnitude of refitted regression coefficients from LAD-based methods compared to those obtained by the non-robust method in the **Supplementary Material**. This is due to the difference between the LAD-based and least square based loss function for robust and non-robust methods, respectively. The advantage of LADBLSS over the non-robust methods can be clearly observed. First, majority of the top 100 important effects identified by BL are main genetic effects. This is less likely to be reasonable as the response variable weight has been well acknowledged to be also dependent on gene–environment interactions. For BLSS, the inclusion probabilities are low compared to those of the LADBLSS, suggesting lower level of certainty and confidence in the regression coefficients obtained from BLSS. The inferior performance of BL and BLSS further justifies the need of developing robust methods in marginal gene–environment interaction studies. Overall, LADBLSS leads to identification results significantly different from all the alternatives, as well as main and interaction effects of important biological implications that are not discovered by the benchmarks.

TABLE 3 | The numbers of main G effects and interactions identified by different approaches and their overlaps for BL (Bayesian LASSO), BLSS (Bayesian LASSO with spike-and-slab priors), LADBL (LAD Bayesian LASSO), and LADBLSS (LAD Bayesian LASSO with spike-and-slab priors).

T2D	Main				Interaction			
	BL	**BLSS**	**LADBL**	**LADBLSS**	**BL**	**BLSS**	**LADBL**	**LADBLSS**
BL	86	5	6	8	14	14	4	8
BLSS		24	3	6		76	20	23
LADBL			20	12			80	50
LADBLSS				20				80

5. DISCUSSION

In the past, G×E interaction studies have been mainly conducted through marginal hypothesis testing, based on a diversity of study designs utilizing parametric, nonparametric, and semiparametric models (Murcray et al., 2009; Thomas, 2010; Mukherjee et al.,

2012), which later have been extended to joint analyses driven primarily by the pathway or gene set based association studies (Wu and Cui, 2013a; Jin et al., 2014; Jiang et al., 2017). In addition, published literature has also reported the success of marginal screening studies, including those based on partial correlations (Niu et al., 2018; Xu et al., 2019). Recently, the effectiveness of regularized variable selection in G×E interaction studies has been increasingly recognized, and a large number of regularization methods have been proposed for joint interaction studies (Zhou et al., 2021). Marginal penalization has also been demonstrated as promising competitors, although they have only been investigated in a limited number of frequentist studies (Shi et al., 2014; Chai et al., 2017; Zhang et al., 2020).

Therefore, the proposed marginal robust Bayesian variable selection is of particular importance, since joint and marginal analysis cannot replace each other and marginal Bayesian penalization has not been examined for G×E studies so far. In particular, with the robustness and incorporation of spike-and-slab priors in the adaptive Bayesian shrinkage, the LADBLSS has an analysis framework more coherent with that of the joint robust analysis[1], which significantly facilitates methodological developments for interaction studies.

Nevertheless, the proposed method has limitations. As a fully Bayesian methods based on MCMC algorithms, the computation cost is generally high due to the tradeoff for quantifying uncertainty using posterior samples. Such a drawback can be addressed through conducting the computation in a parallel manner given the marginal nature of the method. Besides, the variable selection conducted in our study is based on the L1 penalty within the Bayesian framework. As this structure ignores the correlation among genetic features, a possible direction for future improvement is to incorporate network or gene set information in the identification of important gene–environment interactions (Wang et al., 2021). Furthermore, in our study, the genetic factor is represented by one SNP coded as a triadic factor. A closer look at both the additive and dominant penetrance effects of the SNP will lead to elucidation of the genetic basis using marginal interaction studies on a finer scale. For gene–environment interaction studies, marginal and joint analysis are the two major paradigms, and cannot replace each other (Zhou et al., 2021). It is always on a safe side to perform marginal analysis in G×E studies in addition to the joint ones, facilitating a more comprehensive understanding on the genetic architecture of complex diseases.

The marginal Bayesian regularization can be extended to different types of response, for example, under binary, categorical, prognostic and multivariate outcomes. Nevertheless, considering robustness in the generalized models with the Bayesian framework is not trivial, especially under the multivariate responses (Wu et al., 2014; Zhou et al., 2019). We postpone the investigations to the future studies.The interaction between genetic and environmental factors in this study has been modeled as the product of the two corresponding variables, which amounts to "linear" interactions. In practice, the linear interaction assumption has been frequently violated (Ma et al., 2011; Wu and Cui, 2013b; Zhao et al., 2019), which demands accommodation of these nonlinear effects through nonparametric and semiparametric models (Li et al., 2015; Wu et al., 2015, 2018; Ren et al., 2020). It is of great interest and importance to migrate the nonlinear G×E studies to marginal cases in the near future.

[1]Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y., and Wu, C. (under revision). Robust Bayesian variable selection for gene-environment interactions. *Biometrics*.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Authorized access should be granted before accessing the data. Applications to access the data should be sent to dbGap (accession number phs000091.v2.p1). For more information, please refer to NIH dbGap (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by This study is a secondary data analysis. The dataset has been applied through NIH dbGap (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1). In the dataset, the patient information has been de-identified. As indicated from the dbGap website under section Authorized Access/Use Restrictions, IRB is not required for accessing and using the data. According to the original publication, The study was approved by the institutional review board of Brigham and Women's Hospital in Boston; completion of the self-administered questionnaire was considered to imply informed consent. For more information regarding study population, please refer to the original publication: Hu et al. (2001). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XL and CW: conceptualization and writing—original draft preparation. XL, KF, JR, and CW: methodology and writing—review and editing. XL: data analysis. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank the editors and reviewers for their careful review and insightful comments. This work was supported by an innovative research award from K State Johnson Cancer Research Center.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.667074/full#supplementary-material

REFERENCES

Barclay, S. F., Rand, C. M., Borch, L. A., Nguyen, L., Gray, P. A., Gibson, W. T., et al. (2015). Rapid-Onset Obesity with Hypothalamic Dysfunction, Hypoventilation, and Autonomic Dysregulation (ROHHAD): exome sequencing of trios, monozygotic twins and tumours. *Orphanet J. Rare Dis.* 10:103. doi: 10.1186/s13023-015-0314-x

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Brunham, L. R., Kruit, J. K., Verchere, C. B., and Hayden, M. R. (2008). Cholesterol in islet dysfunction and type 2 diabetes. *J. Clin. Invest.* 118, 403–408. doi: 10.1172/JCI33296

Chai, H., Zhang, Q., Jiang, Y., Wang, G., Zhang, S., Ahmed, S. E., et al. (2017). Identifying gene-environment interactions for prognosis using a robust approach. *Econometr. Stat.* 4, 105–120. doi: 10.1016/j.ecosta.2016.10.004

Cordell, H. J., and Clayton, D. G. (2005). Genetic association studies. *Lancet* 366, 1121–1131. doi: 10.1016/S0140-6736(05)67424-7

Cornelis, M. C., and Hu, F. B. (2012). Gene-environment interactions in the development of type 2 diabetes: recent progress and continuing challenges. *Ann. Rev. Nutr.* 32, 245–259. doi: 10.1146/annurev-nutr-071811-150648

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2004). *Bayesian Data Analysis*. London; Boca Raton, FL: Chapman and Hall/CRC.

George, E. I., and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* 88, 881–889. doi: 10.1080/01621459.1993.10476353

Ghanbari, M., de Vries, P. S., de Looper, H., Peters, M. J., Schurmann, C., Yaghootkar, H., et al. (2014). A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. *Hum. Mutat.* 35, 1524–1531. doi: 10.1002/humu.22706

Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., et al. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N. Engl. J. Med.* 345, 790–797. doi: 10.1056/NEJMoa010492

Huang, J., and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* 16, 176–195. doi: 10.1007/s10985-009-9144-2

Huang, Y. C., Liao, W. L., Lin, J. M., Chen, C. C., Liu, S. P., Chen, S. Y., et al. (2018). High levels of circulating endothelial progenitor cells in patients with diabetic retinopathy are positively associated with ARHGAP22 expression. *Oncotarget* 9, 17858. doi: 10.18632/oncotarget.24909

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nat. Rev. Genet.* 6, 287–298. doi: 10.1038/nrg1578

Ishwaran, H., and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.* 33, 730–773. doi: 10.1214/009053604000001147

Jiang, Y., Huang, Y., Du, Y., Zhao, Y., Ren, J., Ma, S., et al. (2017). Identification of prognostic genes and pathways in lung adenocarcinoma using a Bayesian approach. *Cancer Inform.* 1:1176935116684825. doi: 10.1177/1176935116684825

Jin, L., Zuo, X., Su, W., Zhao, X., Yuan, M., Han, L., et al. (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics* 12, 210–220. doi: 10.1016/j.gpb.2014.10.002

Kozumi, H., and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *J. Stat. Comput. Simul.* 81, 1565–1578. doi: 10.1080/00949655.2010.496117

Li, J., Wang, Z., Li, R., and Wu, R. (2015). Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat.* 9, 640. doi: 10.1214/15-AOAS808

Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Anal.* 5, 533–556. doi: 10.1214/10-BA521

Liu, C., Ma, J., and Amos, C. I. (2015). Bayesian variable selection for hierarchical gene–environment and gene–gene interactions. *Hum. Genet.* 134, 23–36. doi: 10.1007/s00439-014-1478-5

Ma, S., Yang, L., Romero, R., and Cui, Y. (2011). Varying coefficient model for gene-environment interaction: a non-linear look. *Bioinformatics* 27, 2119–2126. doi: 10.1093/bioinformatics/btr318

Martin, C. L., Jima, D., Sharp, G. C., McCullough, L. E., Park, S. S., Gowdy, K. M., et al. (2019). Maternal pre-pregnancy obesity, offspring cord blood DNA methylation, and offspring cardiometabolic health in early childhood: an epigenome-wide association study. *Epigenetics* 4, 325–340. doi: 10.1080/15592294.2019.1581594

Morrison, F., Johnstone, K., Murray, A., Locke, J., and Harries, L. W. (2012). Oxidative metabolism genes are not responsive to oxidative stress in rodent beta cell lines. *Exp. Diabetes Res.* 2012:793783. doi: 10.1155/2012/793783

Mukherjee, B., Ahn, J., Gruber, S. B., and Chatterjee, N. (2012). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am. J. Epidemiol.* 175, 177–190. doi: 10.1093/aje/kwr367

Murcray, C. E., Lewinger, J. P., and Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* 169, 219–226. doi: 10.1093/aje/kwn353

Niu, Y. S., Hao, N., and Zhang, H. H. (2018). Interaction screening by partial correlation. *Stat Its Interface* 11, 317–325. doi: 10.4310/SII.2018.v11.n2.a9

Ren, J., Zhou, F., Li, X., Chen, Q., Zhang, H., Ma, S., et al. (2020). Semiparametric Bayesian variable selection for gene-environment interactions. *Stat. Med.* 39, 617–638. doi: 10.1002/sim.8434

Shi, X., Liu, J., Huang, J., Zhou, Y., Xie, Y., and Ma, S. (2014). A penalized robust method for identifying gene–environment interactions. *Genet. Epidemiol.* 38, 220–230. doi: 10.1002/gepi.21795

Simonds, N. I., Ghazarian, A. A., Pimentel, C. B., Schully, S. D., Ellison, G. L., Gillanders, E. M., et al. (2016). Review of the gene-environment interaction literature in cancer: what do we know?. *Genetic Epidemiol.* 40, 356–365. doi: 10.1002/gepi.21967

Thomas, D. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Ann. Rev. Public Health* 31, 21–36. doi: 10.1146/annurev.publhealth.012809.103619

Von Mutius, E. (2009). Gene-environment interactions in asthma. *J. Allergy Clin. Immunol.* 123, 3–11. doi: 10.1016/j.jaci.2008.10.046

Wang, H., Ye, M., Fu, Y., Dong, A., Zhang, M., Feng, L., et al. (2021). Modeling genome-wide by environment interactions through omnigenic interactome networks. *Cell Rep.* 35, 109114. doi: 10.1016/j.celrep.2021.109114

Winham, S. J., Cuellar-Barboza, A. B., Oliveros, A., McElroy, S. L., Crow, S., Colby, C., et al. (2014). Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. *Mol. Psychiatry* 19, 1010–1016. doi: 10.1038/mp.2013.159

Wu, C., and Cui, Y. (2013a). Boosting signals in gene-based association studies via efficient SNP selection. *Brief. Bioinformatics* 15, 279–291. doi: 10.1093/bib/bbs087

Wu, C., and Cui, Y. (2013b). A novel method for identifying nonlinear gene–environment interactions in case–control association studies. *Hum. Genet.* 132, 1413–1425. doi: 10.1007/s00439-013-1350-z

Wu, C., Cui, Y., and Ma, S. (2014). Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. *Stat. Med.* 33, 4988–4998. doi: 10.1002/sim.6287

Wu, C., Li, S., and Cui, Y. (2012). Genetic association studies: an information content perspective. *Curr. Genomics* 13, 566–573. doi: 10.2174/138920212803251382

Wu, C., and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics, *Brief. Bioinformatics* 16, 873–883. doi: 10.1093/bib/bbu046

Wu, C., Shi, X., Cui, Y., and Ma, S. (2015). A penalized robust semiparametric approach for gene–environment interactions. *Stat. Med.* 34, 4016–4030. doi: 10.1002/sim.6609

Wu, C., Zhong, P. S., and Cui, Y. (2018). Additive varying-coefficient model for nonlinear gene-environment interactions. *Stat. Appl. Genet. Mol. Biol.* 17:/j/sagmb.2018.17.issue-2/sagmb-2017-0008/sagmb-2017-0008.xml. doi: 10.1515/sagmb-2017-0008

Xu, Y., Wu, M., Zhang, Q., and Ma, S. (2019). Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. *Genomics* 111, 1115–1123. doi: 10.1016/j.ygeno.2018.07.006

Yu, K., and Moyeed, R. A. (2001). Bayesian quantile regression. *Stat. Probab. Lett.* 54, 437–447. doi: 10.1016/S0167-7152(01)00124-9

Yu, K., and Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Commun. Stat. Theory Methods* 34, 1867–1879. doi: 10.1080/03610920500199018

Zhang, S., Xue, Y., Zhang, Q., Ma, C., Wu, M., and Ma, S. (2020). Identification of gene–environment interactions with marginal penalization. *Genet. Epidemiol.* 44, 159–196. doi: 10.1002/gepi.22270

Zhao, N., Zhang, H., Clark, J. J., Maity, A., and Wu, M. C. (2019). Composite kernel machine regression based on likelihood ratio test for joint testing of genetic and gene–environment interaction effect. *Biometrics* 75, 625–637. doi: 10.1111/biom.13003

Zhou, F., Ren, J., Li, G., Jiang, Y., Li, X., Wang, W., et al. (2019). Penalized variable selection for lipid–environment interactions in a longitudinal lipidomics study. *Genes* 10, 1002. doi: 10.3390/genes10121002

Zhou, F., Ren, J., Lu, X., Ma, S., and Wu, C. (2021). Gene–Environment Interaction: a Variable Selection Perspective. Epistasis. *Methods Mol. Biol.* 2212, 191–223. doi: 10.1007/978-1-0716-0947-7_13

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 24 January 2022
doi: 10.3389/fgene.2021.749573

RFtest: A Robust and Flexible Community-Level Test for Microbiome Data Powerfully Detects Phylogenetically Clustered Signals

Lujun Zhang[1,2], *Yanshan Wang*[3], *Jingwen Chen*[4]* and *Jun Chen*[5]*

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, United States, [2]Institute of Soil and Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, China, [3]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, United States, [4]Department of General Surgery, Zhongshan Hospital, Fudan University, Shanghai, China, [5]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States

Random forest is considered as one of the most successful machine learning algorithms, which has been widely used to construct microbiome-based predictive models. However, its use as a statistical testing method has not been explored. In this study, we propose "Random Forest Test" (RFtest), a global (community-level) test based on random forest for high-dimensional and phylogenetically structured microbiome data. RFtest is a permutation test using the generalization error of random forest as the test statistic. Our simulations demonstrate that RFtest has controlled type I error rates, that its power is superior to competing methods for phylogenetically clustered signals, and that it is robust to outliers and adaptive to interaction effects and non-linear associations. Finally, we apply RFtest to two real microbiome datasets to ascertain whether microbial communities are associated or not with the outcome variables.

Keywords: random forest, hypothesis testing, community-wide test, microbiome, omics association test

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong, Hong Kong SAR, China

Reviewed by:
Chaolong Wang,
Huazhong University of Science and Technology, China
Xihao Li,
Harvard University, United States

***Correspondence:**
Jingwen Chen
Riceawen@163.com
Jun Chen
chen.jun2@mayo.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *29 July 2021*
Accepted: *09 November 2021*
Published: *24 January 2022*

Citation:
Zhang L, Wang Y, Chen J and Chen J (2022) RFtest: A Robust and Flexible Community-Level Test for Microbiome Data Powerfully Detects Phylogenetically Clustered Signals. Front. Genet. 12:749573. doi: 10.3389/fgene.2021.749573

1 INTRODUCTION

The microbiome, the collection of microorganisms and their genetic materials in an environment, has been intricately related to human health (Gao et al., 2018; Gentile and Weir, 2018) and ecosystem functioning (Fierer, 2017). Studying the composition and function of the microbiome has been greatly facilitated by next-generation sequencing *via* marker gene (Weisburg et al., 1991) and/or shotgun metagenomic sequencing techniques (Handelsman, 2004). For the past three decades, the marker gene sequencing has been the dominant approach to investigate the phylogenies and the abundance of microbial groups (Weisburg et al., 1991), while shotgun metagenomics has become increasingly popular to study the functional potential of the microbiome (Quince et al., 2017). Sequences stemming from this marker gene sequencing procedure are usually quality-filtered, merged, and clustered into operational taxonomic units (OTUs) (Schloss et al., 2009; Edgar, 2013) or denoised into amplicon sequence variants (ASVs) (Callahan et al., 2016; Bharti and Grimm, 2021). These OTUs and ASVs are regarded as surrogates of microbial taxa, and downstream statistical analyses are then performed based on the OTU/ASV abundance table, which records the frequencies of the detected OTUs/ASVs in each microbiome sample, together with a phylogenetic tree relating the OTUs/ASVs and the metadata describing the characteristics of the samples.

One central task of microbiome data analyses is to test the association between the microbiome and a variable of interest, while adjusting for potential confounders. Although the ultimate goal is to identify specific microbial taxa associated with the variable of interest, a process also known as differential abundance analysis (Chen et al., 2018), the large abundance variation, weak effects, and the need for multiple testing correction makes differential abundance analysis underpowered for a moderate sample size. It is not uncommon that differential abundance analysis fails to make any discoveries after multiple testing correction when a number of microbial taxa are weakly associated with the variable of interest. In such cases, a community-level test, which jointly analyzes the abundance data at the community level, may be more powerful due to its ability to pool individual weak signals and no need for multiple testing correction. It is also possible to explore the interspecific interactions (Zengler and Zaramela, 2018) and phylogenetic relations (Washburne et al., 2018) in the test to further improve the statistical power. In fact, the community-level tests have been routinely applied, as the first step in statistical analysis of microbiome data, to establish an overall association between the microbiome and the variable of interest. They have been instrumental in disentangling microbial association with, for example, clinical outcomes (Clooney et al., 2021) and environmental gradients (Zhang et al., 2021).

The first community-level test for microbiome data is based on permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2001). PERMANOVA is a distance-based permutation test for assessing the association between a multivariate outcome and a covariate of interest, where the variability of the multivariate outcome is summarized in a distance/dissimilarity matrix. In microbiome applications, ecologically motivated distances/dissimilarities, such as UniFrac (Lozupone and Knight, 2005; Lozupone et al., 2007) distance and Bray–Curtis dissimilarity (Bray and Curtis, 1957), are frequently used. As an alternative to PERMANOVA, the microbiome regression-based kernel association test (MiRKAT) (Zhao et al., 2015) follows a similar logic but treats the abundance data as the covariate and transforms those distance or dissimilarity matrices into kernels; subsequently, community-level associations are evaluated using semi-parametric kernel machine regressions. MiRKAT is computationally efficient, allows a straightforward adjustment for covariates, and accommodates multiple distance kernels through an omnibus test (Zhao et al., 2015). Another community-level test is the adaptive microbiome-based sum of powered scores (aMiSPU), which is an adaptive test based on a series of microbiome-based sum of powered scores (MiSPU) calculated using different powers (Wu et al., 2016). aMiSPU utilizes the variable selection/weighting of the SPU framework (Pan et al., 2014) based on weighted and unweighted generalized taxon proportions and is designed to adapt to the underlying signal structure. Combining the strength of MiRKAT and aMiSPU, the optimal microbiome-based association test (OMiAT) (Koh et al., 2017) substitutes MiSPU with its non-phylogenetic version, sum of powered scores (SPU), and integrates these two criteria *via* an omnibus *p*-value to improve power. These methods all use permutation to assess the statistical significance and hence the type I error rates are well controlled (Anderson, 2001; Zhao et al., 2015; Wu et al., 2016; Koh et al., 2017). However, their power relies on the choice of candidate distances/kernels or specific data transformation (e.g., the power function for MiSPU). Moreover, they have limited ability to exploit the interactions among taxa, which are expected to be prevalent in microbiome data (Zengler and Zaramela, 2018). Additionally, they have not leveraged the strength of machine learning algorithms, which have been shown to be effective in building up microbiome-based predictive models (Marcos-Zambrano et al., 2021).

In the present study, we propose a community-level test based on random forest (RFtest) for testing the associations between the microbiome and an outcome variable. Random forest (Breiman, 2001) is considered as one of the most successful machine learning algorithms, which can be readily applied to diverse tasks, such as variable selection and prediction from high-dimensional omics datasets (Degenhardt et al., 2019). As a non-parametric decision tree-based method, it is robust to outliers and can automatically adapt to the complex relationship between the taxa abundance and the outcome variable without the need for data transformation. Moreover, they can capture high-order interactions in the data without prior knowledge provided (Wright et al., 2016). The proposed method RFtest uses the generalization error estimate of random forest as the test statistic and uses permutation to calculate *p*-values. It incorporates the phylogenetic information *via* creating features that accumulate OTU/ASV abundance along the branches of the phylogenetic tree. RFtest is flexible and can be applied to different types of outcomes. It can also adjust covariates, which facilitates confounder adjustment in microbiome association analysis. By comprehensive simulations, we show that our approach has controlled type I error rates, and is particularly powerful to detect phylogenetically clustered signal, robust to outliers, and capable of detecting complex relationships between microbial taxa, and between the taxa and the outcome.

2 METHODS AND MATERIALS

2.1 Notations

Suppose that we have abundance measurements from n independent microbiome samples and p OTUs/ASVs, denoted by $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_i, \ldots, \mathbf{X}_n)^{\mathrm{T}}$ $(1 \le i \le n)$, where $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip})^{\mathrm{T}}$ $(1 \le j \le p)$ and x_{ij} is the (normalized) abundance of the j^{th} OTU/ASV in the i^{th} sample. Let $\mathbf{Y} = (y_1, y_2, \ldots, y_i, \ldots, y_n)^{\mathrm{T}}$ $(1 \le i \le n)$ denote the vector for the outcome variable, such as clinical outcomes and environmental gradients. Additionally, we may have q covariates, such as age and biological sex, which are denoted by $\mathbf{Z}_{n\times q} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_i, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$, where $\mathbf{Z}_i = (z_{i1}, z_{i2}, \ldots, z_{ik}, \ldots, z_{iq})^{\mathrm{T}}$ $(1 \le k \le q)$ are the measurement of the q covariates in the ith sample. Moreover, we may have a rooted phylogenetic tree G capturing the phylogenetic relatedness of the OTUs/ASVs. G has p leaves (terminal vertices with a degree of 1) and one node (an internal vertex with a degree greater than 1) called root. The p leaves correspond to the p OTUs/ASVs while the root is theoretically assumed to be the last common ancestor

of all vertices in the phylogenetic tree. In a path connecting a leaf and the root, the vertices closer to the root are regarded as "ancestors" of vertices that are farther from; thus, this ancestral relationship describes the relative closeness of vertices to the root of G. The aim for RFtest is to test the association between $\mathbf{Y}_{n\times 1}$ and $\mathbf{X}_{n\times p}$ while adjusting $\mathbf{Z}_{n\times q}$.

2.2 Methods

The tree of life underpins our understanding towards microorganisms (Washburne et al., 2018). Closely related microorganisms share similar biological traits and association signals tend to be clustered with respect to their phylogenetic relationship (Xiao et al., 2017; Xiao et al., 2018a; Xiao et al., 2018b). We therefore aim to utilize the phylogenetic information in the random forest test to improve its power. We incorporate such phylogenetic information by augmenting the OTU/ASV-level abundance data with the abundances of the internal nodes of the phylogenetic tree G. This is achieved by creating an n-by-m feature matrix $\mathbf{W}_{n\times m} = (w_{il})_{n\times m}$ for the m internal nodes in G, where the features accumulate the abundance of OTUs/ASVs belonging to the same ancestor in G. As each leaf corresponds to one OTU/ASV in microbiome and there exists exactly one path between each leaf and the root, the total abundance of all OTU/ASV leaves that shares a specific common ancestor or internal node l is well-defined. Thus, we have

$$w_{il} = \sum_{j\in\mathcal{A}} x_{ij} \tag{1}$$

where w_{il} is the collective abundance of the l^{th} internal node of the i^{th} sample and $\mathcal{A}$ is the set of OTUs/ASVs whose ancestor is the l^{th} internal node.

The RFtest uses the generalization error rate estimate (Breiman, 2001) of random forest as a test statistic, and uses permutation to calculate p-values. Specifically, random forest is firstly grown using the "ranger" package (Wright and Ziegler, 2017) in the R platform (Team, 2020) using $\mathbf{Y}_{n\times 1}$ as outcome variable and $\mathbf{X}_{n\times p}$ and $\mathbf{W}_{n\times m}$ as input features, and the observed out-of-bag (OOB) error rate T_{obs} is used as the test statistic. The OOB error is the average error for each observation calculated using predictions from the trees that do not contain in their respective bootstrap sample. Here, we use the probabilistic prediction for classification and the OOB error is essentially a Brier's score (Malley et al., 2012). Regression and classification trees are used for continuous and binary **Y**s, respectively. When there are no covariates, it permutes the outcome $\mathbf{Y}_{n\times 1}$ B times and calculates the OOB error rate $\tilde{T}^b$ ($b = 1, \ldots, B$) based on the permuted $\mathbf{Y}_{n\times 1}$. The p-value is calculated using:

$$p\text{-value} = \left[\#\left(\tilde{T}^b \le T_{\text{obs}}\right) + 1\right]/(B+1) \tag{2}$$

where $\#(\tilde{T}^b \le T_{\text{obs}})$ is the number of permuted datasets satisfying $\tilde{T}^b \le T_{\text{obs}}$.

When covariates are present, RFtest accommodates covariates using the following steps. Firstly, $\mathbf{Y}_{n\times 1}$ is regressed on covariate $\mathbf{Z}_k$ ($1 \le k \le q$) using linear model if **Y** is continuous:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \ldots + \hat{\beta}_q z_{iq} + e_i = \hat{\beta}_0 + \sum_{k=1}^{q} \hat{\beta}_k z_{ik} + e_i \tag{3}$$

and using logistic regression model if **Y** is binary:

$$\text{logit}\left(P\left(y_i = 1\right)\right) = \hat{\beta}_0 + \sum_{k=1}^{q} \hat{\beta}_k z_{ik} \tag{4}$$

where $\hat{\beta}_0$ and $\hat{\beta}_k$ ($1 \le k \le q$) are the estimated coefficients, and e_i are regression residuals. Next, for a continuous **Y**, we generate $\tilde{\mathbf{Y}}^b$ using residual permutation. The observed error rate T_{obs} is calculated based on the input features $\mathbf{X}_{n\times p}$ and $\mathbf{W}_{n\times \text{m}}$ and the adjusted outcome $\mathbf{Y}_{\text{adj}} = (e_1, e_2, \ldots, e_i, \ldots, e_n)^{\text{T}}$ ($1 \le i \le n$). Thereafter, the permutated $\tilde{\mathbf{Y}}^b = (\tilde{y}_1^b, \tilde{y}_2^b, \ldots, \tilde{y}_i^b, \ldots, \tilde{y}_n^b)^{\text{T}}$ is generated by

$$\tilde{y}_i^b = \tilde{e}_i^b \tag{5}$$

where $\tilde{e}_i^b$ is the permutated regression residuals for the ith sample. For a binary covariate **Y**, $\tilde{\mathbf{Y}}^b$ is generated using a (0, 1) random number generator according to adjusted probabilities of

$$\text{logit}\left(P\left(\tilde{y}_i^b = 1 \,\middle|\, \sum_i \tilde{y}_i^b = \sum_i y_i\right)\right) = \hat{\beta}_0 + \sum_{k=1}^{q} \hat{\beta}_k z_{ik} \tag{6}$$

where we conditioned on the number of observed cases. Finally, we calculate the error rate $\tilde{T}^b$ under permutation based on $\tilde{\mathbf{Y}}^b$ similarly. Consequently, p-value can be obtained using (**Eq. 2**).

We implemented the random forest test in the package "RFtest" on the R platform, which is available on GitHub (https://github.com/Lujun995/Random-forest-test-RFtest).

2.3 Simulation Studies

Simulations were conducted under various scenarios to study whether RFtest would control type I error rates at desired levels and whether it would be a powerful testing approach compared with competing methods. Instead of using a parametrical model such as the Dirichlet-multinomial model (Chen and Li, 2013), the microbiome data were directly resampled from a large gut microbiome study by Hale et al. (2017). Briefly, the study compared the fecal microbiome profiles of patients with adenomas versus healthy controls. 16s rRNA sequences were analyzed using IM-TORNADO pipeline (Jeraldo et al., 2014), OTUs were clustered at 97% identity, and singletons were removed (Hale et al., 2017). After rarefaction to 20,000 counts per sample, the adenoma dataset contained 439 samples and 2,100 OTUs, where we resampled $n = 50$ samples, i.e., $\mathbf{X}_{50\times p}$, without replacement for each simulated dataset. We then constructed the outcome variable $\mathbf{Y}_{50\times 1}$ under six scenarios, following the strategy by Zhao et al. (2015). Let S denote the set that comprises OTUs associated with **Y**. We generated the continuous and binary outcome $\mathbf{Y} = (y_1, y_2, \ldots, y_i, \ldots, y_{50})^{\text{T}}$ ($1 \le i \le 50$) based on

$$y_i = \beta_0 + z_i + \beta\ \text{scale}\left[\sum_{j\in\mathcal{S}}\left(x_{ij}\right)\right] + \varepsilon_i, \tag{7}$$

and

$$\text{logit}\left(P\left(y_i = 1\right)\right) = \beta_0 + z_i + \beta \text{ scale}\left[\sum_{j\in\mathcal{S}}\left(x_{ij}\right)\right], \quad (8)$$

where β_0 is a constant, β is an adjustable effect size, $\varepsilon_i \sim N(0, \sigma^2)$, and the "scale" function standardizes the data to have mean 0 and standard deviation 1. We used $\beta_0 = 10$ for a continuous **Y** and $\beta_0 = 0$ for a binary **Y**, $\varepsilon_i \sim N(0, 1)$.

The first scenario (S0) was used to study the type I error rate of RFtest by setting the effect size $\beta = 0$ under three cases, including no covariates [$z_i = 0$ and $\varepsilon_i \sim N(0, 1)$], one covariate independent of **X** ($z_i \sim N(0, 1)$ and $\varepsilon_i \sim N(0, 9)$), and one covariate associated with **X** {z_i = scale[$\Sigma_{j\in\mathcal{S}}(x_{ij})$] + $N(0, 1)$ and $\varepsilon_i \sim N(0, 9)$}, respectively. In the third case, $\mathcal{S}$ consisted of OTUs from an abundant lineage $\mathcal{S}_A$, which contributed to 15% of the total OTU number and 21% of the total abundance.

The other five scenarios (S1–S5) were used to evaluate the power of RFtest. No covariates were included ($z_i = 0$) in these scenarios. In S1, we investigated different signal types (phylogenetically clustered vs. non-phylogenetically clustered) and different signal densities (5% vs. 15%). For phylogenetically clustered signals, the signal OTUs for 5% and 15% densities were from two abundant lineages $\mathcal{S}_B$ and $\mathcal{S}_A$, respectively, where $\mathcal{S}_B$ was contained in $\mathcal{S}_A$ described above and contributed to 5% of the total OTU number and 11% of the total abundance. For non-clustered signals, the signal OTUs were randomly selected and OTUs for 5% density were also contained in those for 15%. We further substituted the term $\Sigma_{j\in\mathcal{S}}(x_{ij})$ in **Eq. 7** and **Eq. 8** with $\Sigma_{j\in\mathcal{S}}(x_{ij}/\overline{x_{\cdot j}})$, where $\overline{x_{\cdot j}} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij})$, to avoid several OTUs dominating the overall signal strength.

The scenario S2 was designed to further validate the results of clustered signals in S1 using different lineages. We studied seven disjoint major lineages ($\mathcal{S}_I$: $I = A, C, D, E, F, G, H$) in the dataset of Hale et al. (2017), which spanned 80% of the total OTU number and more than 80% of the total abundance. Each lineage possessed 5%–20% of the total OTU number and 1%–40% of the overall abundance. The simulations in this scenario were conducted under $\beta = 2.25$ for a binary **Y** and $\beta = 0.75$ for a continuous **Y**.

The scenario S3 was to evaluate the power of the RFtest when the outcome variable was non-linearly associated with the signal OTUs. We applied a non-linear link function f_{link} to x_{ij}. Specifically,

$$y_i = \beta_0 + \beta \text{ scale}\left[\sum_{j\in\mathcal{S}} f_{link}\left(x_{ij}\right)\right] + \varepsilon_i \quad (9)$$

for a continuous **Y**, and

$$\text{logit}\left(P\left(y_i = 1\right)\right) = \beta_0 + \beta \text{ scale}\left[\sum_{j\in\mathcal{S}} f_{\text{link}}\left(x_{ij}\right)\right] \quad (10)$$

for a binary **Y**, where $f_{\text{link}}(x_{ij}) = \log_2(x_{ij} + 1)$ ($x_{ij} \geq 0$).

The scenario S4 studied a complex association between **Y** and **X** where there was interaction between two sets of signal OTUs. Particularly, for a continuous **Y**, it was generated *via*

$$y_i = \beta_0 + \beta \text{ scale}\left[\sum_{j\in\mathcal{S}}\left(x_{ij}\right)\right] \cdot \text{scale}\left[\sum_{j'\in\mathcal{S}'}\left(x_{ij'}\right)\right] + \varepsilon_i \quad (11)$$

and for a binary **Y**, it was generated using

$$\text{logit}\left(P\left(y_i = 1\right)\right) = \beta_0 + \beta \text{ scale}\left[\sum_{j\in\mathcal{S}}\left(x_{ij}\right)\right] \cdot \text{scale}\left[\sum_{j'\in\mathcal{S}'}\left(x_{ij'}\right)\right], \quad (12)$$

where β was fixed at 1.33 and 5 for a continuous and binary **Y**, respectively, and $\mathcal{S}$ and $\mathcal{S}'$ were two disjoint sets comprising 15% and 13% of total OTUs, respectively. For phylogenetic signals, we let $\mathcal{S} = \mathcal{S}_A$ and $\mathcal{S}' = \mathcal{S}_C$, where $\mathcal{S}_A$ had been characterized in S0 and $\mathcal{S}_C$ was another major lineage accounting for 12% of the total abundance. For non-phylogenetic signal, the terms $\sum_{j\in\mathcal{S}}(x_{ij})$ and $\sum_{j'\in\mathcal{S}'}(x_{ij'})$ in **Eq. 11** and **Eq. 12** were normalized using $\sum_{j\in\mathcal{S}\text{ or }\mathcal{S}}(x_{ij}/\overline{x_{\cdot j}})$, where $\overline{x_{\cdot j}} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij})$. The sample size used in this scenario ranged from 50 to 250 as detection of an interaction generally requires a relatively large sample size.

The last scenario (S5) was used to assess the robustness of RFtest to outliers. Firstly, the outcome variable **Y** was generated according to the procedure in S1, using clustered or non-clustered signals with a density of 15%. Subsequently, the order of OTUs in 0, 1, or 3 samples was randomly shuffled, yielding 0, 1, or 3 outliers; therefore, these outliers would possess distinct microbiome profiles.

The source code of this section is available at GitHub (https://github.com/Lujun995/RFtest-Simulations).

2.4 Competing Methods and Evaluation

The competing methods include the optimal microbiome regression-based kernel association test (optimal MiRKAT) (version 1.1.1, https://cran.r-project.org/package=MiRKAT) (Zhao et al., 2015), the adaptive microbiome-based sum of powered score test (aMiSPU) (version 1.0, https://cran.r-project.org/package=MiSPU) (Wu et al., 2016) and optimal microbiome-based association test (OMiAT) (version 6.0, https://github.com/hk1785/OMiAT) (Koh et al., 2017). While multiple distance or dissimilarity functions could be used in MiRKAT, we followed the example in the "MiRKAT" package (Zhao et al., 2015) and selected weighted and unweighted UniFrac distance (Lozupone and Knight, 2005; Lozupone et al., 2007) and Bray–Curtis dissimilarities (Bray and Curtis, 1957), which have been widely used in microbiome studies. All the results were averaged over 1,000 simulation runs.

3 RESULTS

3.1 Simulation Studies

3.1.1 Factors Influencing the Power of RFtest

We first studied factors that might influence the performance of RFtest including choice of the test statistic, method for *p*-value calculation, sparsity filtering, and the parameters of the random forest ("ranger"). Results of these evaluations were obtained under the scenario S1 (binary outcome).

TABLE 1 | Estimated type I error rate of the random forest test (RFtest).

	Binary outcome variable (Y)	Continuous Y
No covariates (Z)	4.7% (3.6%, 6.2%)[a]	5.3% (4.1%, 6.9%)
Z independent with microbiome data (X)	5.2% (4.0%, 6.8%)	4.7% (3.6%, 6.2%)
Z correlated with X	3.6% (2.6%, 4.9%)	2.9% (2.0%, 4.1%)

[a]*Data are presented as "proportion (L, U)," where the proportion is a point estimate of type I error rate and the L and the U are the lower and upper bounds of Wilson's 95% confidence interval for proportion data. Type I error rates are expected to be ≤ ~5%.*

For the choices of test statistic, we investigated the OOB error rate ("OOB_P"), training error, 0.632 error, and 0.632 + error based on probabilistic predictions. It is well known that the training error underestimates the generalization error while OOB error overestimates it. The 0.632 and 0.632 + rule proposed by Efron and Tibshirani (Efron and Tibshirani, 1997) tried to obtain a more unbiased estimate. In addition to the use of probabilistic predictions, we also compared to the OOB error rate based on binary prediction ("OOB_noP"). **Supplementary Figure S1** shows that error rates based on probability predictions were found to be more powerful than that based on binary predictions, while for different types of error rates based on probabilistic predictions, their performance was similar (**Supplementary Figure S1**). Thus, we selected the OOB error rate with probabilistic predictions as the test statistic. Next, we compared the permutation test to a naïve test, which applied a Wilcoxon rank sum test based on the OOB predicted probabilities. We observed that their *p*-values were highly correlated (**Supplementary Figure S2**); nonetheless, the naïve approach was unable to adjust for covariates and slightly less powerful than the permutation-based RFtest (**Supplementary Figure S3**). We also examined the effect of sparsity filtering on power and computational time of RFtest by filtering features at sparsity thresholds of 98%, 96%, 90%, and 80%. **Supplementary Figure S4** shows that mild filtering (e.g., filter OTUs present in less than 4%–10% of samples) was more beneficial than no filtering or aggressive filtering. Such mild filtering could remarkably reduce computation time while maintaining a similar power. Finally, we studied the impact of the parameters of random forest ("ranger") on the power of RFtest. Concerning the number of split variables, splitting a proportion of 2%–3% of the total OTU number (close to the default) generally performed well under both phylogenetic and non-phylogenetic signals while a greater or smaller number might be preferrable for phylogenetic or non-phylogenetic signals, respectively (**Supplementary Figure S5**). A larger number of decision trees in random forest would stabilize the error rate (**Supplementary Figure S6A**); however, the variance of the sampling distribution of the error rate under permutation was observed 10 times larger than the variance of the OOB error rate across different runs (**Supplementary Figure S6A**). Thus, a larger number would hardly increase the power of the RFtest (**Supplementary Figure S6B**) but significantly increase computational burden. Based on these evaluations, we used an ensemble of 500 decision trees in the RFtest to accelerate the computation and stabilized the estimated error rate by averaging over three runs.

3.1.2 Type I Error Control

We studied the type I error rate control of RFtest by simulating null datasets (S0) with or without covariates. At the nominal level of 5%, we observed that the type I error was controlled at the desired level in situations where a covariate was absent, independent with **X** or correlated with **X** (**Table 1**).

3.1.3 Power Studies

Next, we studied the power of RFtest under different scenarios with association signals (S1–S5). In scenario S1, RFtest was more powerful than competing methods under phylogenetically clustered signals across signal densities for both binary and continuous outcomes (**Figures 1C,DS7c** & **S7d**). While the margin by which the RFtest led might expand or contract for different OTU clusters defined based on the phylogenetic tree in scenario S2 (**Figure 2** & **S8**), RFtest was generally considered as a leading test among all competing methods except in lineage "3590" (**Figure 2** & **S8**). Furthermore, this margin was more notable when the outcome variable is binary (**Figures 1C,D**, **S7c** & **S7d**). For random or non-phylogenetic signals, however, the RFtest appeared to be less powerful than OMiAT and optimal MiRKAT but outperformed aMiSPU (**Figures 1A,B**, **1b**, **S7a** & **S7b**).

Scenarios S3–6 demonstrated the robustness of the RFtest to outliers and its adaptivity to diverse association patterns between **X** and **Y**. In scenario S3, the microbiome profile **X** was related to **Y** on the log scale yielding a non-linear relationship. We found that the results remained similar to those in scenario S1, where a linear relationship was assumed. The RFtest was observed to maintain a leading position under phylogenetical signals but became relatively less powerful under non-phylogenetic signals (**Figure 3** & **S9**). However, compared to scenario S1, the difference diminished among the RFtest, the optimal MiRKAT, and the OMiAT (**Figure 3** & **S9**). These three methods also outperformed aMiSPU (**Figure 3** & **S9**).

In scenario S4, where we simulated interaction effects between OTU clusters, we observed that while the RFtest was a leading method in this scenario, the pattern differed for a binary and continuous outcome. For a binary outcome, RFtest could effectively detect interactions between two phylogenetic clusters or non-phylogenetic groups at a relatively larger sample sizes (**Figure 4**). Meanwhile, the competing methods appeared powerless for both phylogenetic and non-phylogenetic signals (**Figure 4**). For a continuous outcome, RFtest could powerfully detect the association for both types of signals (**Supplementary Figure S10**). Meanwhile, the optimal MiRKAT and the OMiAT became considerably more powerful

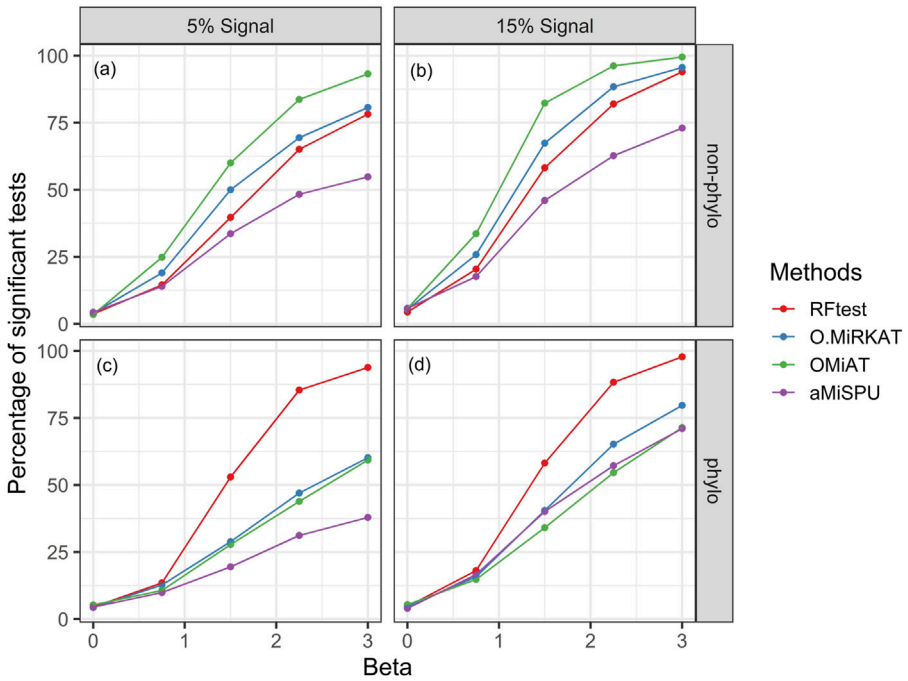


FIGURE 1 | Power comparison among the competing methods for a binary outcome variable under different signal types and densities. Abbreviation: O.MiRKAT, optimal MiRKAT. **(A,B)** Random signals with a density of 5% and 15%, respectively. **(C,D)** Phylogenetically clustered signal with a density of 5% and 15%, respectively.

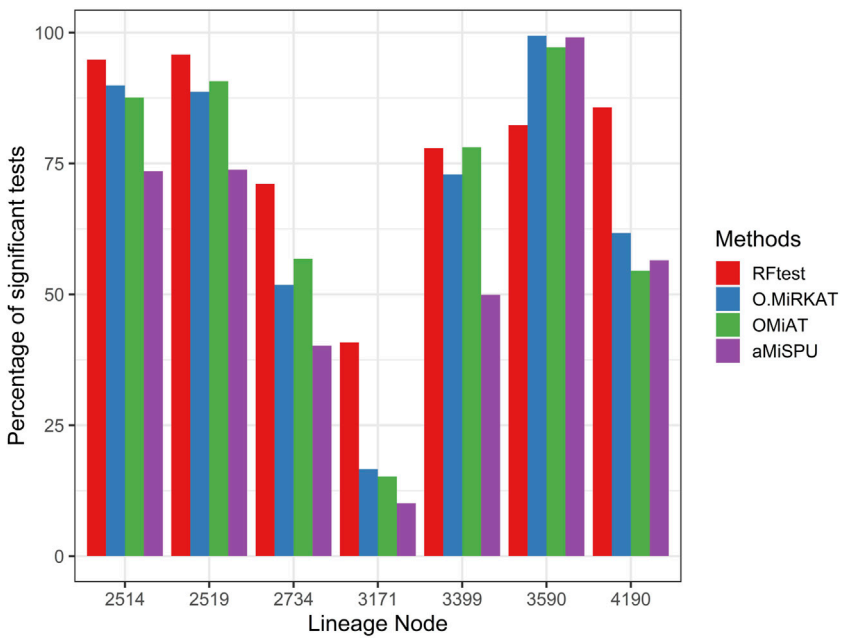


FIGURE 2 | Power comparison among the four competing methods under signals from seven major lineages. The lineage numbers correspond to node numbers in the phylogenetic tree used in simulation in the present study. These lineages span ≥80% of the total OTUs and the total abundance. Abbreviations have the same meaning as in **Figure 1**.

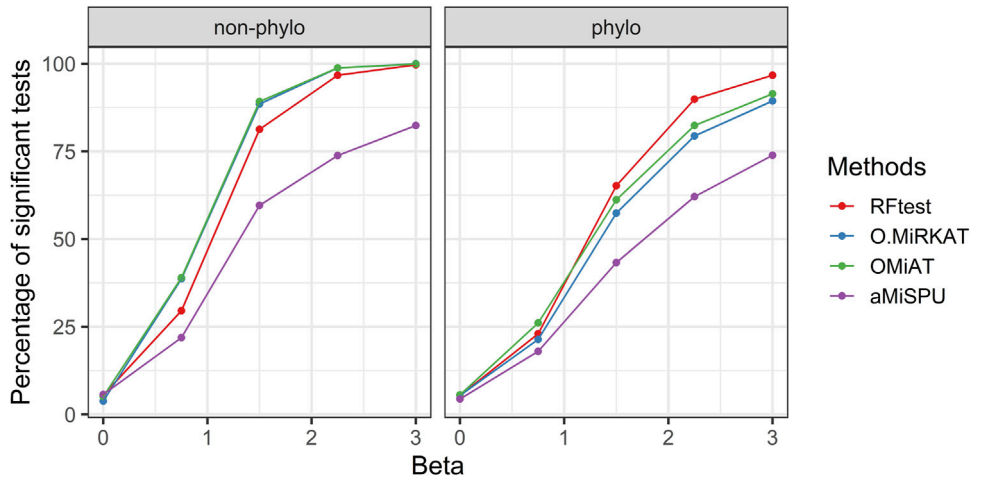


FIGURE 3 | Power comparison among the competing methods for a binary outcome variable when **X** and **Y** are non-linearly associated. The raw OTU abundance data were transformed using a link function of $f_{log2}(x_{ij}) = \log_2(x_{ij} + 1)(x_{ij} \geq 0)$. Two signal types, phylogenetic and non-phylogenetic, with a density of 15% were used.

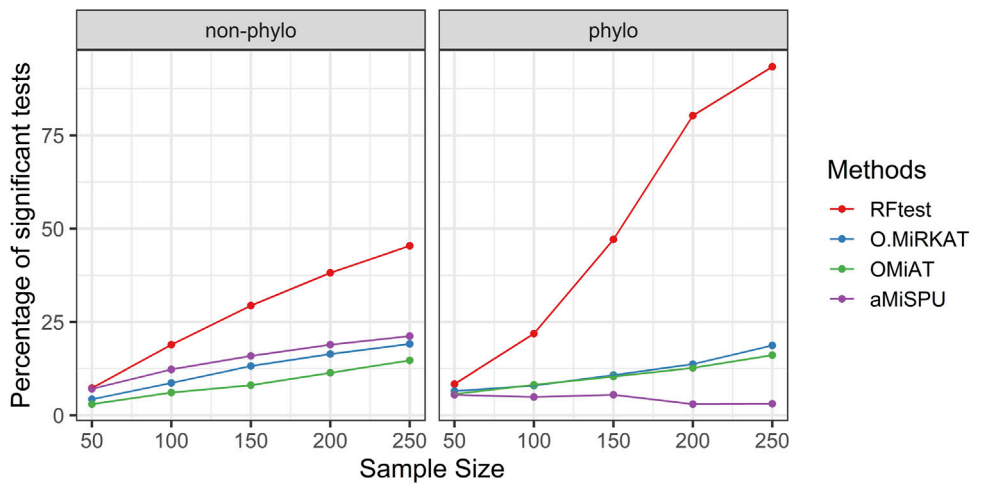


FIGURE 4 | Power comparison among the competing methods when there was interaction between two microbial groups. The outcome variable was binary, and two signal types, phylogenetic and non-phylogenetic, were investigated. The two microbial groups comprised 13% and 15% of the total OTUs.

than the binary case under a non-phylogenetic signal (**Supplementary Figure S10**); however, they remained underpowered under a phylogenetic signal (**Supplementary Figure S10**).

In scenario S5, we simulated one and three outliers to assess the reduction in power when outlier samples were present. The results indicated that RFtest was the most robust among the competing methods, and that the presence of several outliers did not affect the power much for both binary and continuous outcomes with phylogenetic or non-phylogenetic signals, while the power of other methods might be considerably reduced (**Figure 5**; **Supplementary Figure S11**).

3.2 Real Data Analysis

In this section, we aimed to compare the results of RFtest, optimal MiRKAT, aMiSPU, and OMiAT in real-world examples. We re-analyzed the relationship between outcome variables and microbiome profiles in two published datasets. The first example was taken from a study on the throat microbiome (Charlson et al., 2010). That study investigated the effect of smoking on human microbiota in the upper respiratory tract. While detailed information of sample collecting and data processing procedures can be accessed from Charlson et al. (2010), a summary is provided here. Nylon-flocked swabs were taken from the nasopharynx and oropharynx of 62 healthy subjects, including 33 non-smokers and 29 smokers. From each swab, DNA was extracted using the QIAamp DNA Stool Minikit (Qiagen) and the V1–V2 region of the 16S rRNA was amplified. Thereafter, this 16S rRNA was sequenced using a 454 Life Sciences Genome Sequencer FLX instrument (Roche). The sequence reads were denoised (Quince et al., 2009), analyzed

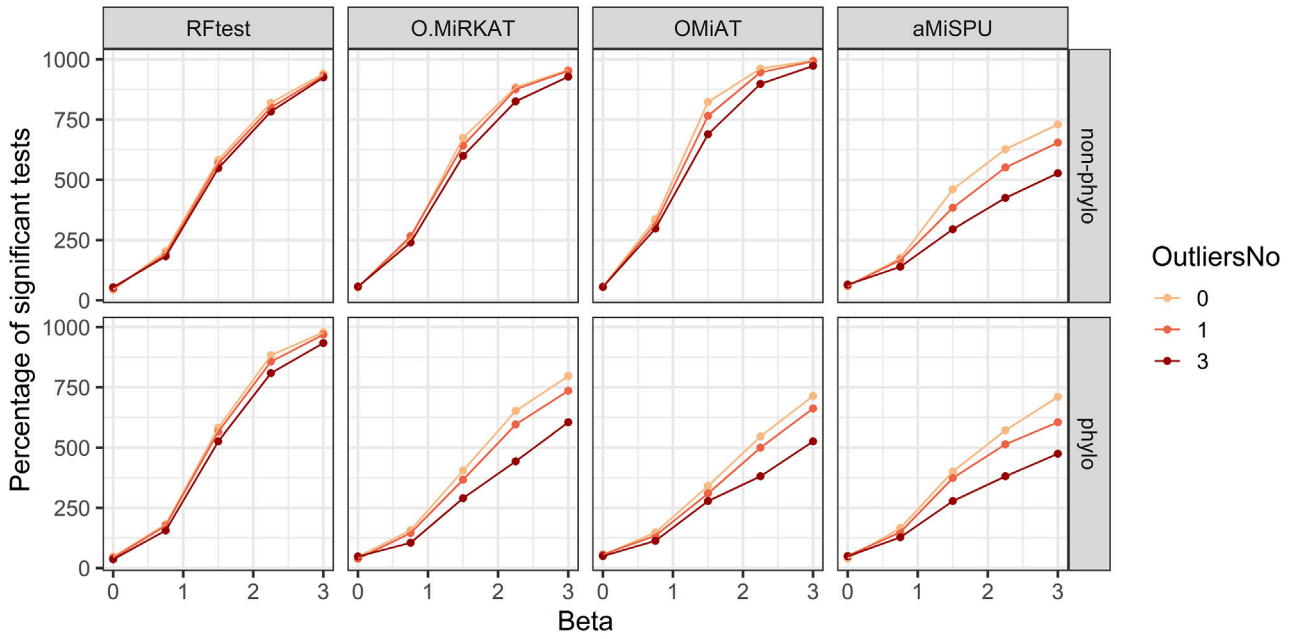


FIGURE 5 | Power curves of the competing methods when outliers were present. The outcome variable was binary, and two signal types, phylogenetic and non-phylogenetic, with a density of 15% were assessed. Zero to three outlier samples were included.

using the QIIME pipeline (Caporaso et al., 2010), and clustered into OTUs at 97% similarity using UCLUST (Edgar, 2010).

In the original study (Charlson et al., 2010), the association between smoking and the respiratory tract microbiome was tested by Permutational Multivariate Analysis of Variance (Anderson, 2001). based on weighted and unweighted UniFrac distances (Lozupone and Knight, 2005; Lozupone et al., 2007). A difference in microbial community structure was reported between smokers and non-smokers ($p < 0.05$). In the present study, we re-analyzed the microbiome data and found consistent results with previous studies (Charlson et al., 2010; Zhao et al., 2015; Wu et al., 2016). When no covariate was considered, the p-value estimated by the RFtest was 0.001 while those of the optimal MiRKAT, the OMiAT, and the aMiSPU were 0.006, 0.008, and 0.009, respectively. When biological sex was included as a confounder, the estimated p-values became 0.002, 0.009, 0.010, and 0.005 for the RFtest, the optimal MiRKAT, the OMiAT, and the aMiSPU, respectively. The RFtest provided more significant p-values in general, while all competing methods rejected the null hypotheses at a significance level of 0.01.

Another relevant example was taken from a study of the distance–decay relationship in microbial ecology (Xue et al., 2021). This relationship can be portrayed as relatedness of microbial communities decreases as their spatial distance increases (Hanson et al., 2012). In brief, surface soil was collected intact from a paddy field in Wenling, Zhejiang Province, China (28°21′ N, 121°15′ E) in November 2017. From the sample, a soil cube (2.0 cm × 2.0 cm × 2.0 cm) was selected and further divided into 4 × 4 × 4 cubes of which each had sides 0.5 cm in length. DNA samples were extracted from these sub-cubes, and the V4–V5 region of the 16S rDNA genes was amplified and subsequently sequenced using an Illumina HiSeq platform. After removal of adaptors and quality control, 16s rDNA sequences were aligned using USEARCH11 (https://www.drive5.com/usearch/) and OTUs were clustered at 97% identity using UPARSE (Edgar, 2013). Finally, the microbial communities were rarefied to 41,752 sequences per sample.

As one of the original findings (Xue et al., 2021), a decreased community similarity, measured by 1 – Bray–Curtis dissimilarity (Bray and Curtis, 1957) between microbial communities, was observed as the spatial distance increased in the 64 sub-cubes (Mantel test, $p = 0.001$). Herein, we re-examined this distance–decay association using the RFtest *via* an assessment of microbial changes along each spatial axis of the *xyz*-coordinate defined in the study of Xue et al. (2021). We found a similar result that the microbiome was associated with the x- and y-axes, and p-values by the RFtest were 0.001, 0.001, and 0.310 for the x-, y-, and z-axes, respectively. Those of the optimal MiRKAT were 0.011, 0.001, and 0.618, respectively; those of the OMiAT were 0.001, 0.001, and 0.265; and those of aMiSPU were 0.006, 0.001, and 0.135. While all methods discovered a statistically significant association between microbial changes and the x-axis, the RFtest reported a more significant p-value than the optimal MiRKAT and the aMiSPU, rejecting the null hypotheses at a significance level of 0.01.

4 DISCUSSION

Random forest has been one of the most successful machine learning methods for microbiome data (Marcos-Zambrano et al., 2021). The superior predictive performance of random forest is

due to its ability to model a complex nonlinear relationship between the microbiome and the outcome, to capture high-order interactions among taxa, and to accommodate a large number of taxa. In this study, we proposed a random forest-based test (RFtest) to assess the association between the microbiome and an outcome variable, borrowing the strengths of random forest in prediction. In RFtest, we incorporated phylogenetic structure by creating features that accumulate OTU abundance along the branches of the phylogenetic tree and used residual permutation to address covariates. Simulation results showed that RFtest could control type I error rate at the desired level with or without confounders (**Table 1**). This approach was closely linked to the naïve approach (**Supplementary Figure S2**); however, the naïve method could not address covariates, which limits its use in real-world applications.

Our benchmarking study further revealed that RFtest had a clear edge over the competing methods to detect phylogenetically clustered signals (**Figure 1**; **Supplementary Figure S11**). This is because our approach incorporates topological information of a phylogenetic tree G into random forest *via* creating features that accumulate leaf OTU abundances. This strategy could also be explored in other machine learning algorithms to capture a clustered signal. Conversely, when the signal OTUs are randomly distributed in the phylogenetic tree, the OMiAT (Koh et al., 2017) and optimal MiRKAT (Zhao et al., 2015) may become a better choice than the RFtest (**Figure 1A**; **Supplementary Figure S7A**). Though non-phylogenetic signal cases were less advantageous to RFtest, we consider that the superior power of RFtest for phylogenetically clustered signals may be practically more important, since phylogenetic signals are extensively observed in microbiome studies, and phylogenetic approaches are of particular interest in microbiome analysis (Washburne et al., 2018).

Our simulation results also demonstrated the robustness of RFtest to outliers and its adaptivity to various types of associations (**Figures 3–5**; **Supplementary Figure S9–11**). Microbiome composition is highly variable, which would largely be ascribed to stochasticity rather than explained (Clooney et al., 2021). Such large biological variation might consequently result in several outliers in a study. Remarkably, outliers affected the power of RFtest minimally, and RFtest was the most robust method to outliers in our benchmarking study (**Figure 5**; **Supplementary Figure S11**). Moreover, microbial communities have been portrayed as a complex ecosystem, in which its components closely interact with each other (Zengler and Zaramela, 2018). These interactions are generally categorized into two groups—beneficial and neutral relationships, such as mutualism and commensalism, and antagonistic relationships, such as competition and predation (Little et al., 2008). For mutualism and commensalism, they can be depicted as a non-linear, positive correlation between bacterial lineages and the outcome variable $\mathbf{Y}$. For antagonistic relationships, a possible signal indicating competitive exclusion, denoted by $\mathbf{Y} = 0$, would occur when one of two lineages overwhelms the other, denoted by $\mathbf{X}_1$ (+), $\mathbf{X}_2$ (−); otherwise, $\mathbf{Y} = 1$ when $\mathbf{X}_1, \mathbf{X}_2$ (+) or $\mathbf{X}_1, \mathbf{X}_2$ (−). Therefore, they would be identified as interaction effects. Notably, our results showed that the RFtest was efficient in discovering a non-linear relationship (**Figure 3** & S9) as well as an interaction effect (**Figure 4** & S10). Given the relatively high performance of the RFtest under these complex conditions (**Figures 3–5**, **S9**, **S10** & **S11**), it may be projected that the RFtest can be flexibly applied to a wide range of data structures to ascertain associations between a microbiome profile $\mathbf{X}$ and an outcome variable $\mathbf{Y}$.

There are several limitations for our proposed method. First, because of the use of bootstrapping in the random forest algorithm, RFtest can be computationally intensive. For example, it took 68 s and 70 MB in memory using a single core on a laptop computer to test the dataset of throat microbiome in our first real data example, compared to 2–4 s and 60–100 MB memory usage of its counterparts. Although computation is usually not a problem for a small dataset, more time would be required for larger datasets. The computation time of random forest increases linearly with the number of variables, i.e., p, and approximately linearly with the sample size n (Wright and Ziegler, 2017). To accelerate the computation of RFtest, we have implemented parallel computing in our software, where each permutation could be run in parallel. Moreover, we could perform sparsity-based filtering to reduce the number of input features to speed up the computation, without affecting the power much (**Supplementary Figure S4**). Another limitation may be that current random forest test could not as effectively identify random, non-phylogenetical signals as OMiAT (**Figures 1A**,**B**; **Supplementary Figures S7A,B**). Increasing the power for non-phylogenetic signal is our future direction of research, for example, by leveraging multiple weighting schemes in RFtest from external data with an omnibus test (Li et al., 2020).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The package "RFtest" was implemented on the R platform, which can be found on GitHub (https://github.com/Lujun995/Random-forest-test-RFtest). The source code of the simulations in the present study is available at GitHub (https://github.com/Lujun995/RFtest-Simulations).

AUTHOR CONTRIBUTIONS

JnC, LZ, and JgC conceived the idea, implemented the method, designed and conducted the simulation studies, and drafted the manuscript. YW contributed to the data analysis and polished the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Center for Individualized Medicine at Mayo Clinic, NIH 1 R21 HG011662 and National Science Foundation NSF-DMS 2113360.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.749573/full#supplementary-material

REFERENCES

Anderson, M. J. (2001). A New Method for Non-parametric Multivariate Analysis of Variance. *Austral Ecol.* 26 (1), 32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Bharti, R., and Grimm, D. G. (2021). Current Challenges and Best-Practice Protocols for Microbiome Analysis. *Brief Bioinform.* 22 (1), 178–193. doi:10.1093/bib/bbz155

Bray, J. R., and Curtis, J. T. (1957). An Ordination of the upland forest Communities of Southern Wisconsin. *Ecol. Monogr.* 27 (4), 325–349. doi:10.2307/1942268

Breiman, L. (2001). Random Forests. *Mach Learn.* 45 (1), 5–32.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-Resolution Sample Inference from Illumina Amplicon Data. *Nat. Methods* 13 (7), 581–583. doi:10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nat. Methods* 7 (5), 335–336. doi:10.1038/nmeth.f.303

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., et al. (2010). Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. *PLoS One* 5 (12), e15216. doi:10.1371/journal.pone.0015216

Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., et al. (2018). An Omnibus Test for Differential Distribution Analysis of Microbiome Sequencing Data. *Bioinformatics* 34 (4), 643–651. doi:10.1093/bioinformatics/btx650

Chen, J., and Li, H. (2013). Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. *Ann. Appl. Stat.* 7 (1), 418–442. doi:10.1214/12-aoas592

Clooney, A. G., Eckenberger, J., Laserna-Mendieta, E., Sexton, K. A., Bernstein, M. T., Vagianos, K., et al. (2021). Ranking Microbiome Variance in Inflammatory Bowel Disease: a Large Longitudinal Intercontinental Study. *Gut* 70 (3), 499–510. doi:10.1136/gutjnl-2020-321106

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets. *Brief Bioinformatics* 20 (2), 492–503. doi:10.1093/bib/bbx124

Edgar, R. C. (2010). Search and Clustering Orders of Magnitude Faster Than BLAST. *Bioinformatics* 26 (19), 2460–2461. doi:10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: Highly Accurate OTU Sequences from Microbial Amplicon Reads. *Nat. Methods* 10 (10), 996–998. doi:10.1038/nmeth.2604

Efron, B., and Tibshirani, R. (1997). Improvements on Cross-Validation: the 632+ Bootstrap Method. *J. Am. Stat. Assoc.* 92 (438), 548–560. doi:10.1080/01621459.1997.10474007

Fierer, N. (2017). Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome. *Nat. Rev. Microbiol.* 15 (10), 579–590. doi:10.1038/nrmicro.2017.87

Gao, L., Xu, T., Huang, G., Jiang, S., Gu, Y., and Chen, F. (2018). Oral Microbiomes: More and More Importance in Oral Cavity and Whole Body. *Protein Cell* 9 (5), 488–500. doi:10.1007/s13238-018-0548-1

Gentile, C. L., and Weir, T. L. (2018). The Gut Microbiota at the Intersection of Diet and Human Health. *Science* 362 (6416), 776–780. doi:10.1126/science.aau5812

Hale, V. L., Chen, J., Johnson, S., Harrington, S. C., Yab, T. C., Smyrk, T. C., et al. (2017). Shifts in the Fecal Microbiota Associated with Adenomatous Polyps. *Cancer Epidemiol. Biomarkers Prev.* 26 (1), 85–94. doi:10.1158/1055-9965.epi-16-0337

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev.* 68 (4), 669–685. doi:10.1128/mmbr.68.4.669-685.2004

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., and Martiny, J. B. H. (2012). Beyond Biogeographic Patterns: Processes Shaping the Microbial Landscape. *Nat. Rev. Microbiol.* 10 (7), 497–506. doi:10.1038/nrmicro2795

Jeraldo, P., Kalari, K., Chen, X., Bhavsar, J., Mangalam, A., White, B., et al. (2014). IM-TORNADO: A Tool for Comparison of 16S Reads from Paired-End Libraries. *PLoS ONE* 9 (12), e114804. doi:10.1371/journal.pone.0114804

Koh, H., Blaser, M. J., and Li, H. (2017). A Powerful Microbiome-Based Association Test and a Microbial Taxa Discovery Framework for Comprehensive Association Mapping. *Microbiome* 5 (1), 45. doi:10.1186/s40168-017-0262-x

Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., et al. (2020). Dynamic Incorporation of Multiple In Silico Functional Annotations Empowers Rare Variant Association Analysis of Large Whole-Genome Sequencing Studies at Scale. *Nat. Genet.* 52 (9), 969–983. doi:10.1038/s41588-020-0676-4

Little, A. E. F., Robinson, C. J., Peterson, S. B., Raffa, K. F., and Handelsman, J. (2008). Rules of Engagement: Interspecies Interactions that Regulate Microbial Communities. *Annu. Rev. Microbiol.* 62, 375–401. doi:10.1146/annurev.micro.030608.101423

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors that Structure Microbial Communities. *Appl. Environ. Microbiol.* 73 (5), 1576–1585. doi:10.1128/aem.01996-06

Lozupone, C., and Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* 71 (12), 8228–8235. doi:10.1128/aem.71.12.8228-8235.2005

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., and Ziegler, A. (2012). Probability Machines. *Methods Inf. Med.* 51 (01), 74–81. doi:10.3414/me00-01-0052

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Front. Microbiol.* 12, 634511. doi:10.3389/fmicb.2021.634511

Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A Powerful and Adaptive Association Test for Rare Variants. *Genetics* 197 (4), 1081–1095. doi:10.1534/genetics.114.165035

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., et al. (2009). Accurate Determination of Microbial Diversity from 454 Pyrosequencing Data. *Nat. Methods* 6 (9), 639–641. doi:10.1038/nmeth.1361

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun Metagenomics, from Sampling to Analysis. *Nat. Biotechnol.* 35 (9), 833–844. doi:10.1038/nbt.3935

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing Mothur: Open-Source, Platform-independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* 75 (23), 7537–7541. doi:10.1128/aem.01541-09

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Washburne, A. D., Morton, J. T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A. M., et al. (2018). Methods for Phylogenetic Analysis of Microbiome Data. *Nat. Microbiol.* 3 (6), 652–661. doi:10.1038/s41564-018-0156-0

Weisburg, W. G., Barns, S. M., Pelletier, D. A., and Lane, D. J. (1991). 16S Ribosomal DNA Amplification for Phylogenetic Study. *J. Bacteriol.* 173 (2), 697–703. doi:10.1128/jb.173.2.697-703.1991

Wright, M. N., Ziegler, A., and König, I. R. (2016). Do little Interactions Get Lost in Dark Random Forests. *BMC Bioinformatics* 17, 145. doi:10.1186/s12859-016-0995-8

Wright, M. N., and Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* 77 (1), 1–17. doi:10.18637/jss.v077.i01

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An Adaptive Association Test for Microbiome Data. *Genome Med.* 8 (1), 56. doi:10.1186/s13073-016-0302-3

Xiao, J., Cao, H., and Chen, J. (2017). False Discovery Rate Control Incorporating Phylogenetic Tree Increases Detection Power in Microbiome-wide Multiple Testing. *Bioinformatics* 33 (18), 2873–2881. doi:10.1093/bioinformatics/btx311

Xiao, J., Chen, L., Johnson, S., Yu, Y., Zhang, X., and Chen, J. (2018). Predictive Modeling of Microbiome Data Using a Phylogeny-Regularized Generalized Linear Mixed Model. *Front. Microbiol.* 9, 1391. doi:10.3389/fmicb.2018.01391

Xiao, J., Chen, L., Yu, Y., Zhang, X., and Chen, J. (2018). A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data. *Front. Microbiol.* 9, 3112. doi:10.3389/fmicb.2018.03112

Xue, R., Zhao, K., Yu, X., Stirling, E., Liu, S., Ye, S., et al. (2021). Deciphering Sample Size Effect on Microbial Biogeographic Patterns and Community Assembly Processes at Centimeter Scale. *Soil Biol. Biochem.* 156, 108218. doi:10.1016/j.soilbio.2021.108218

Zengler, K., and Zaramela, L. S. (2018). The Social Network of Microorganisms - How Auxotrophies Shape Complex Communities. *Nat. Rev. Microbiol.* 16 (6), 383–390. doi:10.1038/s41579-018-0004-5

Zhang, L., Ma, B., Tang, C., Yu, H., Lv, X., Mazza Rodrigues, J. L., et al. (2021). Habitat Heterogeneity Induced by Pyrogenic Organic Matter in Wildfire-Perturbed Soils Mediates Bacterial Community Assembly Processes. *ISME J.* 15 (7), 1943–1955. doi:10.1038/s41396-021-00896-z

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *Am. J. Hum. Genet.* 96 (5), 797–807. doi:10.1016/j.ajhg.2015.04.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fgene.2022.729011

Efficient Approximation of Statistical Significance in Local Trend Analysis of Dependent Time Series

Ang Shan[1,2], *Fang Zhang*[1] *and Yihui Luan*[1]*

[1]*Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China,* [2]*Postdoctoral Programme of Zhongtai Securities Co. Ltd, Jinan, China*

Biological time series data plays an important role in exploring the dynamic changes of biological systems, while the determinate patterns of association between various biological factors can further deepen the understanding of biological system functions and the interactions between them. At present, local trend analysis (LTA) has been commonly conducted in many biological fields, where the biological time series data can be the sequence at either the level of gene expression or OTU abundance, etc., A local trend score can be obtained by taking the similarity degree of the upward, constant or downward trend of time series data as an indicator of the correlation between different biological factors. However, a major limitation facing local trend analysis is that the permutation test conducted to calculate its statistical significance requires a time-consuming process. Therefore, the problem attracting much attention from bioinformatics scientists is to develop a method of evaluating the statistical significance of local trend scores quickly and effectively. In this paper, a new approach is proposed to evaluate the efficient approximation of statistical significance in the local trend analysis of dependent time series, and the effectiveness of the new method is demonstrated through simulation and real data set analysis.

Keywords: local trend analysis, dependent time series, statistical significance, Markov chain model, spectral decomposition theory

OPEN ACCESS

Edited by:
Jun Chen,
Mayo Clinic, United States

Reviewed by:
Yinglei Lai,
George Washington University, United States
Guosheng Han,
Xiangtan University, China

***Correspondence:**
Yihui Luan
yhluan@sdu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *22 June 2021*
Accepted: *01 March 2022*
Published: *26 April 2022*

Citation:
Shan A, Zhang F and Luan Y (2022) Efficient Approximation of Statistical Significance in Local Trend Analysis of Dependent Time Series. Front. Genet. 13:729011. doi: 10.3389/fgene.2022.729011

1 INTRODUCTION

Due to the rapid development of molecular biology technology and the significant reduction to sequencing cost, a large amount of biological time series data has been generated in molecular biological research over the past decade. Among the statistical methods used for time series, local similarity analysis (LSA) has been extensively carried out to identify the correlation between various factors, which can be the genes used in gene expression analysis or operational taxonomic units (OTUs) in metagenomics (Qian et al., 2001; Ruan et al., 2006). Extending the LSA method to the study on the local correlation of repeated time series data, Xia et al. (2011) proposed the extended Local Similarity Analysis method(eLSA), where the confidence interval of LSA was constructed by bootstrap. Due to the ease to use allowed by LSA, it has been widely applied in various fields, for example gene expression profiling (Ji and Tan, 2004; Balasubramaniyan et al., 2005), gene regulatory network construction (Madeira et al. (2010)), symbiotic relationship pattern recognition (Beman et al., 2011; Steele et al.. 2011; Goncalves and Madeira, 2014; Cram et al., 2015) etc. Initially, the permutation test is commonly performed to evaluate the statistical significance of LSA, however, both the approximations of statistical significance and permutation test require the assumption that the time series are independent identically distributed

(i.i.d.), which can be violated in most time series data. In order to analyze the statistical significance of LSA for stationary time series, an approach based on moving block bootstrap was proposed by Zhang et al. (2018), and it is referred to as Moving Block Bootstrap LSA (MBBLSA). To assess statistical significance of LSA for stationary time series data, Zhang et al. (2019) developed a theoretical method, which is known as Data Driven LSA (DDLSA). According to DDLSA, long run variance estimated by a nonparametric kernel method is applied to adjust the asymptotic theory of LSA, on the basis of which the limit distribution of LS score for stationary time series can be obtained.

As suggested by Ji and Tan (2004), the degree of similarity shown by rising, unchanged, or falling trends in time series data can be taken as another indicator of the correlation among various biological factors, which is known as local trend analysis (LTA). In LTA, local similarity analysis is performed on the transformed trend sequence, and the corresponding similarity measure is referred to as the local trend score. Local trend analysis is an extension of local similarity analysis, which can better preserve the changing trend of time series. In addition, the discretization of the original sequence can transform some non-stationary time series into stationary Markov series, which is a big advantage of local trend analysis. He and Zeng (2006) applied dynamic programming algorithm to calculate this value, and then conducted permutation test to evaluate statistical significance. Currently, LTA has been widely adopted in many biological fields, including gene association network (He et al.. 2012; Goncalves et al., 2012; Seno et al., 2006; Skreti et al.. 2014) and transcription factor network (Wu et al., 2010). Nevertheless, it takes long to evaluate the statistical significance of local trend analysis through permutation test. In this case, bioinformatics scientists have shifted attention to exploring how the statistical significance of local trend scores can be evaluated quickly and effectively. By extending the statistical significance evaluation method of local similarity analysis theory to local trend analysis, Xia et al. (2015) developed the statistical significance evaluation method of local trend analysis. However, this method is effective only when the original sequence is independent and identically distributed. On the basis of this and prior studies, this paper improves the approximation method proposed by Xia et al. to develop a general method of statistical significance evaluation for local trend analysis.

This paper is organized as follows. In **Section 2**, an introduction is made of the concept of local trend analysis, and a general method of theoretical evaluation regarding the statistical significance of local trend scores is proposed. In **Section 3**, the effectiveness of the new method is demonstrated by simulation and real data analysis. Finally, the conclusions and future work are indicated in **Section 4**.

2 MATERIAL AND METHODS

2.1 Introduction to Local Trend Analysis

The first step in local trend analysis is to convert time series data into a change trend sequence. In general, if the change trend is indicated by two states, decline and rise, the change trend state set can be set as $\Sigma = (D, U)$ or $\Sigma = (-1, 1)$. If the change trend is indicated by three states decline, unchanged and rise, the change trend state set can be set as $\Sigma = (D, N, U)$ or $\Sigma = (-1, 0, 1)$. Undoubtedly, a collection with more changing trend states can be chosen, but it is rare in practice. For a given time series $X_1, X_2, \ldots, X_n$, they can be converted into $d_i^X\ (i = 1, 2, \ldots, n-1)$ as follows:

when $X_i \neq 0$,

$$d_i^X = \begin{cases} 1 & \text{if} \quad \dfrac{X_{i+1} - X_i}{|X_i|} \geq t \\ 0 & \text{if} \quad -t < \dfrac{X_{i+1} - X_i}{|X_i|} < t \\ -1 & \text{if} \quad \dfrac{X_{i+1} - X_i}{|X_i|} \leq -t \end{cases}, \tag{1}$$

where $t \geq 0$ is a threshold to determine whether there is a trend of change; when $X_i = 0$,

$$d_i^X = \begin{cases} 1 & \text{if} \quad X_i = 0, X_{i+1} > 0 \\ 0 & \text{if} \quad X_i = 0, X_{i+1} = 0 \\ -1 & \text{if} \quad X_i = 0, X_{i+1} < 0 \end{cases}. \tag{2}$$

When $t = 0$, d_i^X involves only two states, and the change trend state set is $\Sigma = (-1, 1)$; when $t \neq 0$, d_i^X involves three states, and the change trend state set is $\Sigma = (-1, 0, 1)$. It is assumed that two time series X_t and Y_t are of the same length, $t = 1, 2, \ldots, n$. First of all, X_t and Y_t are converted into tred series d_i^X and d_i^Y, $i = 1, 2, \ldots, n-1$. Given the maximum time delay $D > 0$, the local similarity analysis is conducted on the transformed trend sequence d_i^X and d_i^Y to obtain the local trend score $LT(D)$, i.e.,

$$LT(D) = \max_{0 \leq i,j,k \leq n; |i-j| \leq D} \left| \sum_{l=0}^{k-1} d_{i+l}^X d_{j+l}^Y \right|. \tag{3}$$

2.2 Statistical Significance Analysis of Local Trend Score

After the local trend score is obtained, it is necessary to evaluate its statistical significance which can be estimated by means of permutation test. In the permutation test, however, only the p value obtained by fully permutating the original data is regarded as an accurate estimate. Since the full permutation is a lengthy process, part permutation is usually selected on a random basis. The p value obtained at this time is limited to an approximate estimate. Besides, the p value obtained may deviate from the actual p value if the number of replacements is too small.

In case that the asymptotic distribution result of the local trend score is obtainable, then the p value of the local trend score can be obtained through the limit distribution. Probability statisticians have obtained the asymptotic distribution theory of the local similarity scores of Markov chains with a mean value of 0, finite second-order moment, and finite subset in $\mathbb{R}$ (Feller, 1951; Daudin et al., 2003; Etienne and Vallois, 2004), as shown in the following theorem.

Theorem 1. Assume that Z_i, $i = 1, 2, \ldots, n$, Markov chains with a mean value of 0, finite second-order moment, and finite subset in $\mathbb{R}$. Assume $\mathbb{E}_\nu(Z_1) = 0, \sigma^2 = \mathbb{E}_\nu(Z_1^2) + 2\sum_{k=1}^{\infty} \mathbb{E}_\nu(Z_1 Z_{k+1})$, where ν is the stationary distribution of Z_i. S_k is the random walk process of Z_i:

$$S_0 = 0, S_k = \sum_{i=1}^{k} Z_i, 1 \le k \le n.$$

Let

$$H_n = \max_{0 \le i \le j \le n} \left(S_j - S_i\right) = \max_{0 \le i \le j \le n} \left(Z_{i+1} + \cdots + Z_j\right).$$

Then $H_n/(\sigma\sqrt{n})$ is the convergence in probability of W^*, where $W^* = \max_{0 \le v \le 1}|W_v|$, W_t is a standard Brownian motion. Xia et al. (2015) used the Theorem 1 to obtain a theoretical evaluation method of statistical significance for local trend analysis. Different from the theoretical evaluation method of statistical significance for local similarity analysis, in local trend analysis, even if the original sequence X_t is independent, the transformed trend sequence d_i^X $(i = 1, 2, \ldots, n-1)$ is not independent, because d_i^X and d_{i+1}^X both depend on X_i. In order to facilitate the use of Theorem 1 to calculate the p value of the local trend score, the following assumptions are proposed.

Assumption 1. d_i^X and d_i^Y are mutually independent first-order Markov chains, and the product of d_i^X and d_i^X is also a first-order Markov chain, namely

$$P\left(d_i^X d_i^Y | d_{i-1}^X d_{i-1}^Y, \ldots, d_1^X d_1^Y\right) = P\left(d_i^X d_i^Y | d_{i-1}^X d_{i-1}^Y\right). \quad (4)$$

Under the Assumption 1, d_i^X and d_i^Y are irreducible non-periodic Markov chains, so the theoretical method in Feller (1951), Daudin et al. (2003) and Etienne and Vallois (2004) can be directly applied. Xia et al. (2015) suggested a method of theoretically evaluating statistical significance for local trend analysis, with the approximate p value of the local trend score $LT(D)$ obtained as:

$$P(LT(D) \ge s_D) = P\left(\frac{LT(D)}{\sigma\sqrt{n}} \ge \frac{s_D}{\sigma\sqrt{n}}\right) \approx \mathcal{L}_D\left(\frac{s_D}{\sigma\sqrt{n}}\right), \quad (5)$$

where s_D represents the local trend score of X_t and Y_t, and the definition of the tail probability distribution function $\mathcal{L}_D(x)$ is expressed as follows:

$$\mathcal{L}_D(x) = 1 - 8^{2D+1}\left[\sum_{k=1}^{\infty}\left\{\frac{1}{x^2} + \frac{1}{(2k-1)^2\pi^2}\right\}\exp\left\{-\frac{(2k-1)^2\pi^2}{2x^2}\right\}\right]^{2D+1}. \quad (6)$$

It can be found out that σ^2 plays a vital role in the p value approximation **Eq. 5** of the local trend score, which is referred to as the variance of Markov chain. From the formula $\sigma^2 = \mathbb{E}_v(Z_1^2) + 2\sum_{k=1}^{\infty}\mathbb{E}_v(Z_1 Z_{k+1})$, it can be seen that when the stationary distribution of Markov chain v and k step transition probability matrix are known, $\mathbb{E}_v(Z_1 Z_k)$ $(k \ge 1)$ can be obtained. Thus, σ^2 can be obtained easily through calculation. Xia et al. presented the display expression of σ^2 when the original sequence is independent and identically distributed. In practice, however, the original sequence contradicts the assumption of independent and identical distribution. Zhang et al. (2019) proposed an asymptotic statistical significance for local similarity analysis, with the approximate p value of the local similarity score $LS(D)$ similar to $LT(D)$:

$$P(LS(D) \ge s_D) = P\left(\frac{LS(D)}{\omega\sqrt{n}} \ge \frac{s_D}{\omega\sqrt{n}}\right) \approx \mathcal{L}_D\left(\frac{s_D}{\omega\sqrt{n}}\right), \quad (7)$$

where $\omega = \lim_{n\to\infty}\sqrt{var\left(\sum_{i=1}^{n} Z_i\right)/n}$ is referred to as the long-run variance, and $\mathcal{L}_D(x)$ is expressed as **Eq. 6**. Because Markov chains can be regarded as time series, they also satisfy **Eq. 7**. It is obvious that ω for Markov chains is σ. Therefore, we can get the statistical significance for local trend analysis of non-independent identically distributed time series if the σ^2 is obtained.

Next, the formula of σ^2 is proposed for the local trend score of the time series in general using the spectral decomposition theory of the matrix.

2.2.1 Spectral Decomposition Theorem of Matrix

First, the definition and properties of simple matrix are given.

Definition 1. Let matrix $\boldsymbol{A} \in \boldsymbol{C}^{n\times n}$, λ_i be the differential eigenvalues of $\boldsymbol{A}$, $i = 1, 2, \ldots, s$, and the characteristic polynomial of $\mathbf{A}$ is

$$det(\lambda\boldsymbol{I} - \boldsymbol{A}) = (\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2}\ldots, (\lambda - \lambda_s)^{m_s},$$

where $\sum_{i=1}^{s} m_i = n$. Call m_i the algebraic multiplicity of the eigenvalues λ_i of the matrix $\boldsymbol{A}$.

Definition 2. The solution space V_{λ_i} of the homogeneous equation set $\boldsymbol{Ax} = \lambda_i x$ (i = 1, 2, ..., s) is called the eigenspace of $\boldsymbol{A}$ corresponding to the eigenvalue λ_i, and the dimension of V_{λ_i} is called the geometric multiplicity of the eigenvalue λ_i of the matrix $\boldsymbol{A}$.

Definition 3. If the algebraic multiplicity of each eigenvalue of the matrix $\boldsymbol{A}$ is equal to its geometric multiplicity, then $\boldsymbol{A}$ is called a simple matrix.

Theorem 2. (Spectral decomposition theorem) Let matrix $\boldsymbol{A} \in \boldsymbol{C}^{n\times n}$, λ_i be the differential eigenvalues of $\boldsymbol{A}$, m_i is the algebraic multiplicity of λ_i, i = 1, 2, ..., s, then the sufficient and necessary condition of $\boldsymbol{A}$ being a simple matrix is that there is a unique $\boldsymbol{E}_i \in \boldsymbol{C}^{n\times n}$, i = 1, 2, ..., s, so

1) $\sum_{i=1}^{s} \boldsymbol{E}_i = \boldsymbol{I}$.

2) $\boldsymbol{E}_i\boldsymbol{E}_j = \begin{cases} \boldsymbol{E}_i, & i = j \\ 0, & i \ne j \end{cases}$.

3) $\boldsymbol{A} = \sum_{i=1}^{s} \lambda_i \boldsymbol{E}_i$.

2.2.2 Two-State Markov Chain Model

Firstly, the two-state Markov chain model is studied. When $t = 0$, d_i^X and d_i^Y, $i = 1, 2, \ldots, n-1$ can be obtained by discretizing the original sequence X_t and Y_t. Assume that the distribution of the original sequence is symmetrical, and the mean is 0. Also assume that d_i^X is a first-order stationary Markov chain. Since the original sequence distribution is symmetrical, the stationary distribution of d_i^X is $P(d_i^X = 1) = P(d_i^X = -1) = 1/2$, $\mathbb{E}((d_1^X)^2) = 1^2 \times \frac{1}{2} + (-1)^2 \times \frac{1}{2} = 1$. It is assumed that the transition probability matrices of d_i^X and d_i^Y are $\boldsymbol{T}_X$ and $\boldsymbol{T}_Y$ respectively, as expressed below.

$$\boldsymbol{T}_X = \begin{array}{c|cc} & -1 & 1 \\ \hline -1 & a_X & 1-a_X \\ 1 & 1-a_X & a_X \end{array} \quad \boldsymbol{T}_Y = \begin{array}{c|cc} & -1 & 1 \\ \hline -1 & a_Y & 1-a_Y \\ 1 & 1-a_Y & a_Y \end{array} \tag{8}$$

It can be obtained by calculation, $\mathbb{E}(d_1^X d_{k+1}^X) = (2a_X - 1)^k$, $\mathbb{E}((d_1^X)^2) = \mathbb{E}((d_1^Y)^2) = 1$, $\mathbb{E}(d_1^Y d_{k+1}^Y) = (2a_Y - 1)^k$ (**Supplementary Material S1**). Under the null hypothesis that X_i and Y_i are uncorrelated,

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left(\left(d_1^X d_1^Y\right)^2\right) + 2\sum_{k=1}^{\infty}\mathbb{E}\left(\left(d_1^X d_{k+1}^X\right)\left(d_1^Y d_{k+1}^Y\right)\right) \\
&= \mathbb{E}\left(\left(d_1^X\right)^2\right)\mathbb{E}\left(\left(d_1^Y\right)^2\right) + 2\sum_{k=1}^{\infty}\mathbb{E}\left(d_1^X d_{k+1}^X\right)\mathbb{E}\left(d_1^Y d_{k+1}^Y\right) \\
&= 1 + 2\sum_{k=1}^{\infty}(2a_X - 1)^k(2a_Y - 1)^k \\
&= 1 + 2 \times \lim_{k\to\infty}\frac{(2a_X - 1)(2a_Y - 1) - (2a_X - 1)^{k+1}(2a_Y - 1)^{k+1}}{1 - (2a_X - 1)(2a_Y - 1)} \\
&= 1 + 2 \times \frac{(2a_X - 1)(2a_Y - 1)}{1 - (2a_X - 1)(2a_Y - 1)} \\
&= \frac{1 + (2a_X - 1)(2a_Y - 1)}{1 - (2a_X - 1)(2a_Y - 1)}.
\end{aligned} \tag{9}$$

thus, when $t = 0$, the p value of the local trend score $LT(D)$ is written as

$$P(LT(D) \geq s_D) = \mathcal{L}_D\left(\frac{s_D}{\sigma\sqrt{n}}\right), \tag{10}$$

where s_D indicates the local trend score of X_i and Y_i, σ is obtained using the **Eq. 9**, and $\mathcal{L}_D(x)$ is defined as **Eq. 6**.

2.2.3 Three-State Markov Chain Model

Secondly, the three-state Markov chain model is studied. When $t \neq 0$, d_i^X and d_i^Y are three-state Markov chains. Similarly, it is assumed that the transition probability matrices of d_i^X and d_i^Y are $\boldsymbol{T}_X$ and $\boldsymbol{T}_Y$ respectively, as expressed below.

$$\boldsymbol{T}_X = \begin{array}{c|ccc} & -1 & 0 & 1 \\ \hline -1 & b_X & 1-b_X-c_X & c_X \\ 0 & d_X & 1-2d_X & d_X \\ 1 & c_X & 1-b_X-c_X & b_X \end{array} \quad \boldsymbol{T}_Y = \begin{array}{c|ccc} & -1 & 0 & 1 \\ \hline -1 & b_Y & 1-b_Y-c_Y & c_Y \\ 0 & d_Y & 1-2d_Y & d_Y \\ 1 & c_Y & 1-b_Y-c_Y & b_Y \end{array} \tag{11}$$

It can be obtained by calculation, $\mathbb{E}(d_1^X d_{k+1}^X) = \varphi_1^X T_{1,1}^{X,k} + \varphi_{-1}^X T_{-1,-1}^{X,k} - \varphi_1^X T_{1,-1}^{X,k} - \varphi_{-1}^X T_{-1,1}^{X,k}$, $\mathbb{E}((d_1^X)^2) = \varphi_{-1}^X + \varphi_1^X$, $\mathbb{E}((d_1^Y)^2) = \varphi_{-1}^Y + \varphi_1^Y$, $\mathbb{E}(d_1^Y d_{k+1}^Y) = \varphi_1^Y T_{1,1}^{Y,k} + \varphi_{-1}^Y T_{-1,-1}^{Y,k} - \varphi_1^Y T_{1,-1}^{Y,k} - \varphi_{-1}^Y T_{-1,1}^{Y,k}$ (**Supplementary Material S2**). Under the null hypothesis that X^i and Y^i are uncorrelated,

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left(\left(d_1^X d_1^Y\right)^2\right) + 2\sum_{k=1}^{\infty}\mathbb{E}\left(\left(d_1^X d_{k+1}^X\right)\left(d_1^Y d_{k+1}^Y\right)\right) \\
&= \mathbb{E}\left(\left(d_1^X\right)^2\right)\mathbb{E}\left(\left(d_1^Y\right)^2\right) + 2\sum_{k=1}^{\infty}\mathbb{E}\left(d_1^X d_{k+1}^X\right)\mathbb{E}\left(d_1^Y d_{k+1}^Y\right) \\
&= \left(\varphi_{-1}^X + \varphi_1^X\right)\left(\varphi_{-1}^Y + \varphi_1^Y\right) \\
&\quad + 2\sum_{k=1}^{\infty}\left(\varphi_1^X T_{1,1}^{X,k} + \varphi_{-1}^X T_{-1,-1}^{X,k} - \varphi_1^X T_{1,-1}^{X,k} - \varphi_{-1}^X T_{-1,1}^{X,k}\right) \\
&\quad \left(\varphi_1^Y T_{1,1}^{Y,k} + \varphi_{-1}^Y T_{-1,-1}^{Y,k} - \varphi_1^Y T_{1,-1}^{Y,k} - \varphi_{-1}^Y T_{-1,1}^{Y,k}\right) \\
&= 4\varphi_1^X\varphi_1^Y \quad + 2\varphi_1^X\varphi_1^Y \\
&\quad \times \sum_{k=1}^{\infty}\left(T_{1,1}^{X,k} + T_{-1,-1}^{X,k} - T_{1,-1}^{X,k} - T_{-1,1}^{X,k}\right) \quad \left(T_{1,1}^{Y,k} + T_{-1,-1}^{Y,k} - T_{1,-1}^{Y,k} - T_{-1,1}^{Y,k}\right) \\
&= 4\varphi_1^X\varphi_1^Y + 2\varphi_1^X\varphi_1^Y\sum_{k=1}^{\infty}2(b_X - c_X)^k \quad \times 2(b_Y - c_Y)^k \\
&= 4\varphi_1^X\varphi_1^Y \quad \left(1 + 2\lim_{k\to\infty}\frac{(b_X - c_X)(b_Y - c_Y) - (b_X - c_X)^{k+1}(b_Y - c_Y)^{k+1}}{1 - (b_X - c_X)(b_Y - c_Y)}\right) \\
&= 4\left(\frac{d_X}{1 - b_X - c_X + 2d_X}\right) \\
&\quad \times \left(\frac{d_Y}{1 - b_Y - c_Y + 2d_Y}\right) \quad \left(\frac{1 + (b_X - c_X)(b_Y - c_Y)}{1 - (b_X - c_X)(b_Y - c_Y)}\right).
\end{aligned} \tag{12}$$

Thus, when $t \neq 0$, the p value of the local trend score $LT(D)$ is expressed as

$$P(LT(D) \geq s_D) = \mathcal{L}_D\left(\frac{s_D}{\sigma\sqrt{n}}\right), \tag{13}$$

where s_D represents the local trend score of X_i and Y_i, σ is obtained using the **Eq. 12**, and $\mathcal{L}_D(x)$ is defined as **Eq. 6**.

2.2.4 Mixed-State Markov Chain Model

Thirdly, the mixed-state Markov chain model is studied. When $t \neq 0$, d_i^X or d_i^Y is potentially a two-state Markov chain as well. At this time, if d_i^X and d_i^Y are both two-state Markov chains, σ^2 can be estimated using the two-state Markov chain model. The circumstance where only d_i^X or d_i^Y is a two-state Markov chain is defined as a mixed-state Markov chain model. Without any compromise on generality, it is supposed that d_i^X is a two-state Markov chain while d_i^Y is a three-state Markov chain.

It can obtained by the previous derivation that

$$\begin{aligned}
&\mathbb{E}\left(\left(d_1^X\right)^2\right) = 1, \\
&\mathbb{E}\left(d_1^X d_{k+1}^X\right) = (2a_X - 1)^k, \\
&\mathbb{E}\left(\left(d_1^Y\right)^2\right) = \varphi_{-1}^Y + \varphi_1^Y = \frac{2d_Y}{1 - b_Y - c_Y + 2d_Y}, \\
&\mathbb{E}\left(d_1^Y d_{k+1}^Y\right) = \varphi_1^Y T_{1,1}^{Y,k} + \varphi_{-1}^Y T_{-1,-1}^{Y,k} - \varphi_1^Y T_{1,-1}^{Y,k} - \varphi_{-1}^Y T_{-1,1}^{Y,k}, \\
&\qquad = 2\left(\frac{d_Y}{1 - b_Y - c_Y + 2d_Y}\right)(b_Y - c_Y)^k.
\end{aligned}$$

So,

$$\begin{aligned}\sigma^2 &= \mathbb{E}\left(\left(d_1^X d_1^Y\right)^2\right) + 2\sum_{k=1}^{\infty}\mathbb{E}\left(\left(d_1^X d_{k+1}^X\right)\left(d_1^Y d_{k+1}^Y\right)\right)\\ &= \mathbb{E}\left(\left(d_1^X\right)^2\right)\mathbb{E}\left(\left(d_1^Y\right)^2\right) + 2\sum_{k=1}^{\infty}\mathbb{E}\left(d_1^X d_{k+1}^X\right)\mathbb{E}\left(d_1^Y d_{k+1}^Y\right)\\ &= \frac{2d_Y}{1-b_Y-c_Y+2d_Y} + \frac{4d_Y}{1-b_Y-c_Y+2d_Y}\sum_{k=1}^{\infty}(2a_X-1)^k(b_Y-c_Y)^k\\ &= \left(\frac{2d_Y}{1-b_Y-c_Y+2d_Y}\right)\times\\ &\quad\left(1+2\lim_{k\to\infty}\frac{(2a_X-1)(b_Y-c_Y)-(2a_X-1)^{k+1}(b_Y-c_Y)^{k+1}}{1-(2a_X-1)(b_Y-c_Y)}\right)\\ &= \left(\frac{2d_Y}{1-b_Y-c_Y+2d_Y}\right)\left(\frac{1+(2a_X-1)(b_Y-c_Y)}{1-(2a_X-1)(b_Y-c_Y)}\right).\end{aligned} \tag{14}$$

Thus, when $t \neq 0$ and the circumstance arises that d_i^X and d_i^Y are not both three-state Markov chains, the p value of the local trend score $LT(D)$ is expressed as

$$P(LT(D) \geq s_D) = \mathcal{L}_D\left(\frac{s_D}{\sigma\sqrt{n}}\right), \tag{15}$$

where s_D represents the local trend score of X_i and Y_i, σ is obtained using the **Eq. 14**, and $\mathcal{L}_D(x)$ is defined as **Eq. 6**.

In summary, the p value approximation formula has been obtained for the local trend score of a two-state, three-state or mixed-state Markov chain. Despite a lack of rigorous mathematical proof for the aforementioned p value approximation method, it is still discovered that the p value obtained using this algorithm is approximately equal to the given significance level by simulation, especially when the sample size is large. Therefore, the results obtained using this method are deemed approximately valid.

2.2.5 Estimation of Markov Chain Transition Probability Matrix

In order to calculate the p value of the local trend score, it is essential to estimate the variance σ^2, and the estimation of the variance depends only on the transition probability matrix of the Markov chain. With the original sequence considered as independent and identically distributed, Xia et al. (2015) deduced the value of parameter in transition probability matrix of the two-state ($t = 0$) and three-state ($t = 0.5$) Markov chain. When the original series are non-independent and identically distributed, however, the estimate is inaccurate. It is detailed below how to estimate the transition probability matrix of a two-state or three-state Markov chain under normal circumstances.

For a two-state Markov chain, since both $T_{-1,-1}$ and $T_{1,1}$ are equal to a, the mean of $n_{-1,-1}/n_{-1,\cdot}$ and $n_{1,1}/n_{1,\cdot}$ is taken as the final estimate of a, that is, $\hat{a} = \frac{1}{2}\left(\frac{n_{-1,-1}}{n_{-1,\cdot}} + \frac{n_{1,1}}{n_{1,\cdot}}\right)$, where $n_{-1,\cdot} = n_{-1,-1} + n_{-1,1}$, $n_{1,\cdot} = n_{1,-1} + n_{1,1}$, $n_{u,v}$ represents the number of $(d_i, d_{i+1}) = (u, v)$, $u, v \in (-1, 1)$, $i = 1, 2, \ldots, n-2$.

Likewise, for a three-state Markov chain, since both $T_{-1,-1}$ and $T_{1,1}$ are equal to b, the mean of $n_{-1,-1}/n_{-1,\cdot}$ and $n_{1,1}/n_{1,\cdot}$ is treated as the final estimate of b, that is, $\hat{b} = \frac{1}{2}\left(\frac{n_{-1,-1}}{n_{-1,\cdot}} + \frac{n_{1,1}}{n_{1,\cdot}}\right)$, where $n_{-1,\cdot} = n_{-1,-1} + n_{-1,0} + n_{-1,1}$, $n_{1,\cdot} = n_{1,-1} + n_{1,0} + n_{1,1}$, and $n_{u,v}$ represents the number of $(d_i, d_{i+1}) = (u, v)$, $u, v \in (-1, 0, 1)$, $i = 1, 2, \ldots, n-2$. Similarly, the estimate of c is $\hat{c} = \frac{1}{2}\left(\frac{n_{1,1}}{n_{1,\cdot}} + \frac{n_{1,-1}}{n_{1,\cdot}}\right)$, and the estimate of d is $\hat{d} = \frac{1}{2}\left(\frac{n_{0,-1}+n_{0,1}}{n_{0,\cdot}}\right)$, where $n_{0,\cdot} = n_{0,-1} + n_{0,0} + n_{0,1}$.

In this article, the method put forward by Xia et al. is denoted as TLTA (Theoretical Local Trend Analysis), while the method proposed in this paper is referred to as STLTA (Stationary Theoretical Local Trend Analysis).

3 RESULTS AND DISCUSSION

3.1 Simulation

The effects on the correlation test of time series data are explored by conducting Permutation test, TLTA and STLTA respectively. The following three models are commonly used and familiar to researchers, which can better reflect the correlation between two time series, especially the correlation of two time series can be adjusted by changing the coefficient values. In order to study the difference in type I error rate and significance level among different methods under the original hypothesis, simulation data is generated using the following three models: The effects on the correlation test of time series data are explored by conducting Permutation test, TLTA and STLTA respectively. In order to study the difference in type I error rate and significance level among different methods under the original hypothesis, simulation data is generated using the following three models:

1) AR(1) model:

$$\begin{aligned}X_t &= \rho_1 X_{t-1} + \varepsilon_t^X,\\ Y_t &= \rho_2 Y_{t-1} + \varepsilon_t^Y.\end{aligned} \tag{16}$$

2) ARMA(1,1) model:

$$\begin{aligned}X_t &= \rho_1 X_{t-1} + \varepsilon_t^X + 0.5\varepsilon_{t-1}^X,\\ Y_t &= \rho_2 Y_{t-1} + \varepsilon_t^Y + 0.5\varepsilon_{t-1}^Y.\end{aligned} \tag{17}$$

3) ARMA(1,1)-TAR(1) model:

$$\begin{aligned}X_t &= \rho_1 X_{t-1} + \varepsilon_t^X + 0.5\varepsilon_{t-1}^X,\\ Y_t &= \begin{cases}\rho_2 Y_{t-1} + \varepsilon_t^Y, Y_{t-1} \leq -1\\ 0.5Y_{t-1} + \varepsilon_t^Y, Y_{t-1} > -1.\end{cases}\end{aligned} \tag{18}$$

Where $0 < |\rho_1|, |\rho_2| < 1$, ε_t^X and ε_t^Y are independent standard normal random variables. All the three models are stationary. For each model, it starts by generating X_1 and Y_1 through the standard normal distribution, before the generation of X_t and Y_t, $i = 2, \ldots, 100, +, n$ using the above-mentioned model. Finally, the first 100 samples are discarded, and the remaining n samples are treated as real X_t and Y_t. This data generation process is effective in ensuring the stationarity of the time series.

With consideration given to the impact of autoregressive coefficients ρ_1, ρ_2 and sample size n on the type I error rate for the different methods with the three models, we choose six different combinations of autoregressive coefficients ρ_1, ρ_2, and respectively take the values of −0.5, −0.5; 0, 0; 0.3, 0.3; 0.3, 0.5; 0.5, 0.5; 0.5, 0.8. For each combination of autoregressive coefficients,

TABLE 1 | Type I error rate for different methods (the third to fifth columns) in the AR(1) model when $t = 0$. The first and second columns represent different combinations of autoregressive coefficients and sample sizes. The number of permutation tests is 1,000, the number of repeated simulations is 10,000, and the significance level is $\alpha = 0.05$.

ρ_1, ρ_2	n	Permutation test	TLTA	STLTA
−0.5, −0.5	20	0.1413	0.0470	0.0040
	40	0.1444	0.0764	0.0128
	60	0.1378	0.0880	0.0169
	80	0.1472	0.1040	0.0213
	100	0.1380	0.1046	0.0238
	200	0.1465	0.1059	0.0283
0, 0	20	0.0610	0.0170	0.0119
	40	0.0613	0.0270	0.0209
	60	0.0605	0.0311	0.0257
	80	0.0545	0.0363	0.0282
	100	0.0551	0.0360	0.0300
	200	0.0581	0.0367	0.0357
0.3, 0.3	20	0.0518	0.0109	0.0136
	40	0.0451	0.0177	0.0272
	60	0.0475	0.0179	0.0285
	80	0.0408	0.0238	0.0310
	100	0.0435	0.0260	0.0349
	200	0.0428	0.0254	0.0371
0.3, 0.5	20	0.0459	0.0092	0.0135
	40	0.0397	0.0165	0.0288
	60	0.0379	0.0181	0.0314
	80	0.0407	0.0233	0.0334
	100	0.0359	0.0237	0.0354
	200	0.0345	0.0221	0.0424
0.5, 0.5	20	0.0398	0.0091	0.0159
	40	0.0414	0.0159	0.0284
	60	0.0365	0.0176	0.0314
	80	0.0369	0.0199	0.0343
	100	0.0355	0.0213	0.0374
	200	0.0344	0.0215	0.0428
0.5, 0.8	20	0.0412	0.0088	0.0161
	40	0.0388	0.0134	0.0277
	60	0.0338	0.0145	0.0342
	80	0.0319	0.0165	0.0357
	100	0.0337	0.0214	0.0411
	200	0.0314	0.0170	0.0402

TABLE 2 | Type I error rate for different methods (the third to fifth columns) in the ARMA(1,1) model when $t = 0$. The first and second columns represent different combinations of autoregressive coefficients and sample sizes. The number of permutation tests is 1,000, the number of repeated simulations is 10,000, and the significance level is $\alpha = 0.05$.

ρ_1, ρ_2	n	Permutation test	TLTA	STLTA
−0.5, −0.5	20	0.0617	0.0166	0.0112
	40	0.0609	0.0262	0.0219
	60	0.0557	0.0323	0.0289
	80	0.0562	0.0333	0.0267
	100	0.0538	0.0354	0.0311
	200	0.0572	0.0338	0.0329
0, 0	20	0.0444	0.0109	0.0210
	40	0.0463	0.0170	0.0380
	60	0.0455	0.0213	0.0404
	80	0.0422	0.0270	0.0464
	100	0.0397	0.0242	0.0444
	200	0.0428	0.0260	0.0539
0.3, 0.3	20	0.0472	0.0109	0.0240
	40	0.0497	0.0168	0.0426
	60	0.0413	0.0187	0.0404
	80	0.0395	0.0222	0.0421
	100	0.0421	0.0261	0.0545
	200	0.0418	0.0250	0.0559
0.3, 0.5	20	0.0483	0.0095	0.0218
	40	0.0447	0.0172	0.0410
	60	0.0438	0.0198	0.0427
	80	0.0453	0.0230	0.0432
	100	0.0399	0.0240	0.0515
	200	0.0420	0.0231	0.0505
0.5, 0.5	20	0.0503	0.0097	0.0220
	40	0.0409	0.0186	0.0403
	60	0.0455	0.0191	0.0417
	80	0.0445	0.0235	0.0460
	100	0.0399	0.0271	0.0509
	200	0.0342	0.0257	0.0591
0.5, 0.8	20	0.0492	0.0093	0.0202
	40	0.0430	0.0158	0.0337
	60	0.0399	0.0193	0.0372
	80	0.0435	0.0206	0.0366
	100	0.0359	0.0204	0.0418
	200	0.0381	0.0199	0.0462

the sample size n is set to 20, 40, 60, 80, 100, 200. For simplicity, we select the time delay $D = 0$. In all simulations, the significance level is set to 0.05.

When $t = 0$, the original sequence is converted into a two-state Markov chain, and the type I error rates in the AR(1) model of different methods are presented in **Table 1**. The results show that when $\rho_1 = -0.5$, $\rho_2 = -0.5$, neither Permutation test nor TLTA can control the type I error rate even if the sample size n is small, and their type I error rates are getting bigger as the sample size increases. At this time, the type I error rate of STLTA gradually approaches the significance level 0.05 with the increase of sample size. When $\rho_1 = 0$, $\rho_2 = 0$, X_t and Y_t are all independent and identically distributed sequences, the type I error rates of the three methods are very close to the given significance level, and are getting closer as the sample size increases. When $\rho_1 > 0, \rho_2 > 0$, the type I error rate of Permutation test decreases with the increase of sample size n, and gradually deviates from the significance level 0.05, while the type I error rate of STLTA is closer to the significance level than that of TLTA. For different autocorrelation coefficients, the type I error rates of Permutation test and TLTA show a declining trend with the increase of ρ, and they are increasingly deviant from the significance level. By contrast, STLTA shows an upward trend with the rise of ρ, and it gradually approaches the significance level, suggesting that STLTA is more suitable for stationary time series data. The performances of these three methods in ARMA(1,1) and ARMA(1,1)-TAR(1) models are shown in the **Tables 2**, **3** respectively, which are similar to that in the AR(1)

TABLE 3 | Type I error rate for different methods (the third to fifth columns) in the ARMA(1,1)-TAR(1) model when$t$ = 0. The first and second columns represent different combinations of autoregressive coefficients and sample sizes. The number of permutation tests is 1,000, the number of repeated simulations is 10,000, and the significance level is α = 0.05.

ρ_1, ρ_2	n	Permutation test	TLTA	STLTA
−0.5, −0.5	20	0.0563	0.0127	0.0119
	40	0.0527	0.0194	0.0220
	60	0.0463	0.0247	0.0282
	80	0.0481	0.0279	0.0285
	100	0.0481	0.0264	0.0291
	200	0.0437	0.0277	0.0341
0, 0	20	0.0437	0.0083	0.0147
	40	0.0436	0.0150	0.0270
	60	0.0393	0.0177	0.0350
	80	0.0412	0.0212	0.0377
	100	0.0354	0.0210	0.0382
	200	0.0362	0.0221	0.0435
0.3, 0.3	20	0.0395	0.0076	0.0172
	40	0.0382	0.0126	0.0332
	60	0.0393	0.0136	0.0349
	80	0.0363	0.0183	0.0385
	100	0.0353	0.0195	0.0411
	200	0.0296	0.0186	0.0470
0.3, 0.5	20	0.0372	0.0068	0.0199
	40	0.0345	0.0128	0.0328
	60	0.0356	0.0137	0.0336
	80	0.0328	0.0174	0.0382
	100	0.0315	0.0208	0.0437
	200	0.0354	0.0184	0.0448
0.5, 0.5	20	0.0343	0.0067	0.0170
	40	0.0338	0.0130	0.0337
	60	0.0305	0.0130	0.0367
	80	0.0319	0.0196	0.0400
	100	0.0309	0.0160	0.0399
	200	0.0251	0.0163	0.0463
0.5, 0.8	20	0.0410	0.0061	0.0176
	40	0.0316	0.0127	0.0322
	60	0.0330	0.0142	0.0354
	80	0.0323	0.0170	0.0377
	100	0.0273	0.0181	0.0414
	200	0.0294	0.0189	0.0466

TABLE 4 | Type I error rate for different methods (the third to fifth columns) in the AR(1) model when$t$ = 0.5. The first and second columns represent different combinations of autoregressive coefficients and sample sizes. The number of permutation tests is 1,000, the number of repeated simulations is 10,000, and the significance level is α = 0.05.

ρ_1, ρ_2	n	Permutation test	TLTA	STLTA
−0.5, −0.5	20	0.2236	0.0275	0.0400
	40	0.2155	0.0520	0.0134
	60	0.2210	0.0508	0.0119
	80	0.2158	0.0665	0.0166
	100	0.2159	0.0682	0.0178
	200	0.2213	0.0702	0.0226
0, 0	20	0.0737	0.0039	0.0263
	40	0.0628	0.0059	0.0188
	60	0.0594	0.0075	0.0220
	80	0.0572	0.0089	0.0247
	100	0.0552	0.0084	0.0246
	200	0.0580	0.0107	0.0325
0.3, 0.3	20	0.0379	0.0009	0.0276
	40	0.0296	0.0012	0.0216
	60	0.0296	0.0011	0.0277
	80	0.0229	0.0025	0.0304
	100	0.0270	0.0017	0.0324
	200	0.0241	0.0021	0.0398
0.3, 0.5	20	0.0243	0.0006	0.0229
	40	0.0174	0.0010	0.0246
	60	0.0170	0.0013	0.0263
	80	0.0184	0.0018	0.0337
	100	0.0184	0.0012	0.0334
	200	0.0152	0.0011	0.0355
0.5, 0.5	20	0.0196	0.0002	0.0175
	40	0.0149	0.0005	0.0221
	60	0.0102	0.0006	0.0282
	80	0.0105	0.0003	0.0311
	100	0.0124	0.0005	0.0350
	200	0.0104	0.0003	0.0430
0.5, 0.8	20	0.0099	0.0001	0.0159
	40	0.0052	0.0001	0.0194
	60	0.0036	0.0002	0.0286
	80	0.0032	0.0001	0.0303
	100	0.0033	0.0000	0.0325
	200	0.0017	0.0000	0.0377

model. Under these two models, when $\rho_1 = -0.5$, $\rho_2 = -0.5$, X_t is an independent and identically distributed sequence, so the type I error rates of Permutation test, TLTA and STLTA are close to the significance level. In other cases, the type I error rate of STLTA is closer to the significance level than that of TLTA, while the type I error rate of Permutation test gradually gets away from the significance level as the sample size increases.

When t = 0.5, the original sequence is converted into a three-state Markov chain, and the type I error rates in the AR(1) model of different methods are presented in **Table 4**. In the AR(1) model, when $\rho_1 = -0.5$, $\rho_2 = -0.5$, the type I error rate of Permutation test still far exceeds the given significance level 0.05 even if the sample size is very small (n = 20), and TLTA cannot control the type I error rate even when the sample size is large. When $\rho_1 = 0$, $\rho_2 = 0$, the type I error rate of Permutation test is closer to the significance level than that of TLTA and STLTA, and the type I error rate of TLTA is far less than the significance level. When $\rho_1 > 0$, $\rho_2 > 0$, similar to the case of t = 0, the type I error rate of Permutation test also decreases with the increase of sample size n, and gradually deviates from the significance level. The type I error rate of TLTA is much smaller than the significance level, while that of STLTA shows an upward trend with the rise of the sample size n and gradually approaches the significance level. For different combinations of autocorrelation coefficients, the type I error rates of permutation test and TLTA decline with the increase of ρ, with a gradual deviation from the significance level, with TLTA in particular. Even though the autocorrelation is extremely weak, the type I error rate is far less than 0.05, even below 0.01. While STLTA performs

TABLE 5 | Type I error rate for different methods (the third to fifth columns) in the ARMA(1,1) model when t = 0.5. The first and second columns represent different combinations of autoregressive coefficients and sample sizes. The number of permutation tests is 1,000, the number of repeated simulations is 10,000, and the significance level is α = 0.05.

ρ_1, ρ_2	n	Permutation test	TLTA	STLTA
−0.5, −0.5	20	0.0767	0.0033	0.0269
	40	0.0609	0.0047	0.0166
	60	0.0595	0.0070	0.0212
	80	0.0566	0.0082	0.0229
	100	0.0542	0.0094	0.0284
	200	0.0552	0.0104	0.0343
0, 0	20	0.0300	0.0008	0.0251
	40	0.0211	0.0008	0.0354
	60	0.0187	0.0013	0.0429
	80	0.0201	0.0012	0.0442
	100	0.0185	0.0018	0.0456
	200	0.0190	0.0016	0.0533
0.3, 0.3	20	0.0137	0.0001	0.0239
	40	0.0112	0.0004	0.0395
	60	0.0115	0.0008	0.0424
	80	0.0083	0.0004	0.0453
	100	0.0100	0.0003	0.0489
	200	0.0073	0.0007	0.0579
0.3, 0.5	20	0.0109	0.0001	0.0208
	40	0.0073	0.0002	0.0306
	60	0.0044	0.0001	0.0431
	80	0.0044	0.0003	0.0456
	100	0.0048	0.0004	0.0473
	200	0.0037	0.0003	0.0565
0.5, 0.5	20	0.0076	0.0000	0.0206
	40	0.0050	0.0000	0.0360
	60	0.0052	0.0002	0.0406
	80	0.0041	0.0000	0.0442
	100	0.0041	0.0002	0.0511
	200	0.0028	0.0001	0.0509
0.5, 0.8	20	0.0020	0.0000	0.0148
	40	0.0010	0.0000	0.0249
	60	0.0011	0.0000	0.0288
	80	0.0008	0.0000	0.0333
	100	0.0007	0.0000	0.0333
	200	0.0003	0.0000	0.0470

TABLE 6 | Type I error rate for different methods (the third to fifth columns) in the ARMA(1,1)-TAR(1) model when t = 0.5. The first and second columns represent different combinations of autoregressive coefficients and sample sizes. The number of permutation tests is 1,000, the number of repeated simulations is 10,000, and the significance level is α = 0.05.

ρ_1, ρ_2	n	Permutation test	TLTA	STLTA
−0.5, −0.5	20	0.0521	0.0013	0.0241
	40	0.0421	0.0034	0.0201
	60	0.0375	0.0040	0.0257
	80	0.0364	0.0049	0.0264
	100	0.0370	0.0049	0.0282
	200	0.0330	0.0049	0.0338
0, 0	20	0.0276	0.0005	0.0234
	40	0.0189	0.0009	0.0245
	60	0.0186	0.0009	0.0311
	80	0.0188	0.0009	0.0360
	100	0.0174	0.0011	0.0340
	200	0.0150	0.0016	0.0440
0.3, 0.3	20	0.0169	0.0003	0.0207
	40	0.0113	0.0005	0.0294
	60	0.0097	0.0007	0.0301
	80	0.0108	0.0006	0.0351
	100	0.0091	0.0007	0.0386
	200	0.0072	0.0004	0.0440
0.3, 0.5	20	0.0140	0.0000	0.0209
	40	0.0089	0.0005	0.0283
	60	0.0077	0.0000	0.0317
	80	0.0072	0.0006	0.0340
	100	0.0079	0.0003	0.0375
	200	0.0067	0.0004	0.0439
0.5, 0.5	20	0.0090	0.0001	0.0198
	40	0.0047	0.0001	0.0271
	60	0.0054	0.0000	0.0296
	80	0.0039	0.0002	0.0360
	100	0.0038	0.0002	0.0370
	200	0.0045	0.0000	0.0450
0.5, 0.8	20	0.0072	0.0000	0.0184
	40	0.0045	0.0001	0.0251
	60	0.0024	0.0001	0.0328
	80	0.0024	0.0001	0.0323
	100	0.0016	0.0000	0.0338
	200	0.0013	0.0000	0.0440

well in controlling the type I error rate across all autocorrelation coefficient combinations. The performances of these three methods in ARMA(1,1) and ARMA(1,1)-TAR(1) models are shown in the **Tables 5**, **6**. In these two models, the type I error rate of TLTA is always far less than the significance level. When $\rho_1 = -0.5$, $\rho_2 = -0.5$, the type I error rate of Permutation test is closer to the significance level than that of STLTA. But in other cases, the type I error rate of Permutation test is much smaller than the significance level, and it increasingly deviants from the significance level with the increase of sample size and autocorrelation, while the type I error rate of STLTA gradually approaches the significance level as the sample size increases.

According to the analysis of the results, it can be figured out that STLTA is capable to control the type I error rate under different models, while the permutation test and TLTA are ineffective in this respect, which evidences that STLTA is more effective in utilizing the internal properties of time series than the other two methods, and that it can achieve a more accurate approximation of the local trend score p value.

3.2 Empirical Analysis

3.2.1 Data set of Moving Pictures of Human Microbiome

The STLTA method is applied to the Moving Pictures of Human Microbiome (MPHM) data set, for comparison with the results as obtained from DDLSA, TLTA and Permutation test. The data set of MPHM was collected from two healthy subjects, one male ("M3") and one female ("F4"). Both individuals were sampled

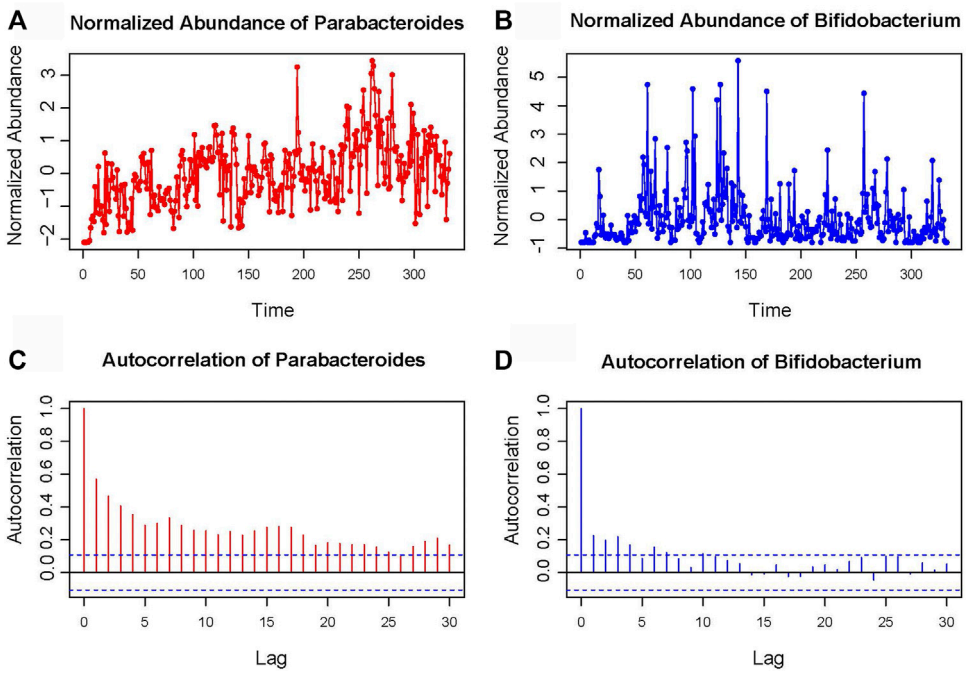


FIGURE 1 | Standardized abundance map of *Parabacteroides* **(A)** and *Bifidobacterium* **(B)** in MPHM "M3" sample fecal data set. The autocorrelation graph **(C,D)** shows the autocorrelation coefficient of the time series at different delays.

daily at three body sites: gut (feces), mouth(tongue), and skin (left and right palms) (Caporaso et al. (2011)). The data set consists of 130, 135 and 133 daily samples from "F4", and 332, 372 and 357 samples from "M3". There are 335, 373 and 1,295 operational taxonomic units (OTUs) from feces, tongue and palm (both left and right) sites of "F4" and "M3", where the taxonomic level is Genus. We selected 59 "core" OTUs that were observed in at least 60% samples from the feces of "M3" and analyzed their relationships. Then, metagenomic analysis is conducted to obtain a time series of OTU abundance. As shown in **Figure 1**, there are two OTUs chosen to display their time series graphs and autocorrelation graphs. It can be found that the abundance sequence of Parabacteroides shows more significant autocorrelation compared to Bifidobacterium, and that their Box-Ljung test p values are all very close to 0, indicating that their autocorrelation relationship is of much significance.

TABLE 7 | The numbers of significant correlations between OTUs found by permutation tests, TLTA, STLTA and DDLSA for different data sets and significance levels.

	—	t = 0.5		t = 0	
Dataset	—	MPHM	PML	MPHM	PML
# of factors	—	59	75	59	75
$p \le 0.05$ $q \le 0.05$	Permutation	589	87	727	29
—	TLTA	165	75	532	13
—	STLTA	739	50	667	13
—	DDLSA	685	371	685	371
$p \le 0.01$ $q \le 0.01$	Permutation	489	84	549	29
—	TLTA	86	74	436	11
—	STLTA	621	16	514	4
—	DDLSA	549	227	549	227

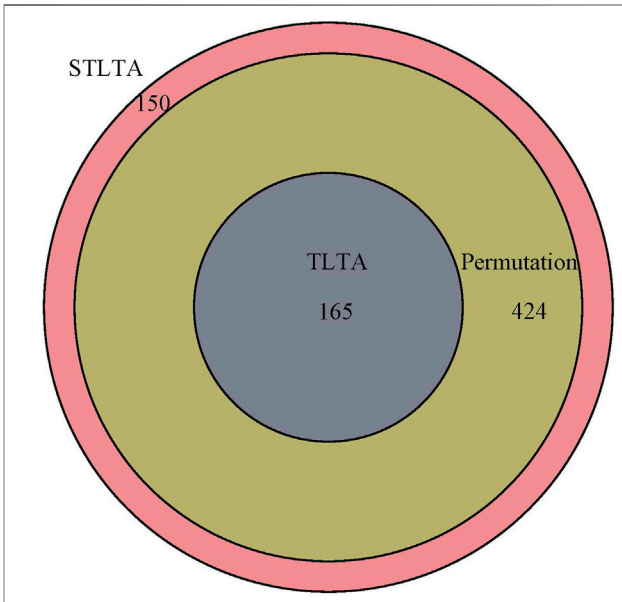


FIGURE 2 | The Venn diagram of the significant relationships found in permutation test, TLTA and STLTA for the MPHM "M3" fecal data set. Green, blue, and red indicate the number of significant relationships found by permutation test, TLTA, and STLTA, respectively.

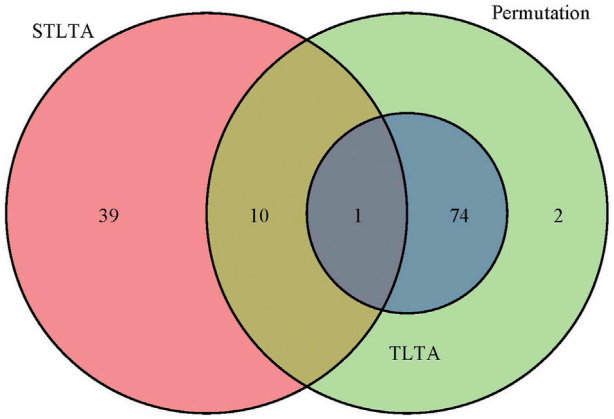


FIGURE 3 | The Venn diagram of the significant relationships found in permutation test, TLTA and STLTA for the PML data set. Green, blue, and red indicate the number of significant relationships found by permutation test, TLTA, and STLTA, respectively.

The significance level is set to 0.05 and 0.01, based on which a comparison is drawn in the significant relationship between the OTUs found by permutation test, TLTA, STLTA and DDLSA with the time delay of $D = 3$. The results are presented in **Table 7**. When $t = 0.5$ and the significance level $p = 0.05$, $Q = 0.05$, in all 1711 pairs of OTU relationships in the "M3" feces sample, it was found that 589, 165, 739 and 685 pairs of significant relationships by Permutation test, TLTA, STLTA and DDLSA respectively, which were 34.4, 9.6, 43.2 and 40% of the total. STLTA found the most significant relationship, followed by DDLSA, and TLTA the least. This is very similar to the simulation results obtained earlier: when $t = 0.5$ and the sample time point is 300, if the samples have autocorrelation relationship, the simulation results show that the type I error rates of Permutation test and TLTA are far less than the given significance level, while the type I error rate of STLTA is close to the given significance level. Therefore when there is correlation between autocorrelation samples, it is possible that permutation test and TLTA fail to identify many significant relationships that actually exist, but STLTA can do this. Although the permutation test can also find many significant relationships, most of them are between samples without autocorrelation. In addition, the numbers of significant correlations between OTUs found by STLTA and DDLSA are approximate, shown that STLSA can discover most significant relationships found by DDLSA.

Venn diagram (**Figure 2**) shows the relationship among the results obtained using different methods in the "M3" stool sample. All of the significant relationships identified by TLTA are discovered by permutation test, and all of the significant relationships identified by permutation test are discovered by STLTA. For more stringent standards $p = 0.01$ and $Q = 0.01$ as well as different thresholds, the results are listed in **Table 7**. By comparing the results of $t = 0$ and $t = 0.5$, it can be found out that the permutation test and TLTA can identify more significant relationships at $t = 0$ then at $t = 0.5$, especially for TLTA. However, STLTA is just the opposite, with the significant relationship found at $t = 0$ less then at $t = 0.5$.

3.2.2 Data set of Plymouth Marine Laboratory

The STLTA method is applied to the Plymouth Marine Laboratory (PML) data set, for comparison with the results as obtained from DDLSA, TLTA and Permutation test. The PML data set is one of the longest microbial time series consisting of monthly samples taken over 6 years at a temperate marine coastal site off Plymouth, United Kingdom (Gilbert et al. (2012)). These samples were sequenced using high-resolution 16S rRNA tag NGS sequencing. A total of 155 bacterial OTUs were identified with the taxonomic level of Order. Among them, we chose 62 abundant OTUs that were present in at least 50% of the time points, and 13 environment factors to analyze their association network. We filled the missing values in the environment data using linear interpolation.

Given time delay $D = 3$ and significance level $p = 0.05$, $Q = 0.05$, when $t = 0.5$ among all the relationships between OTUs and between OTU and environmental factors, permutation test, TLTA, STLTA and DDLSA identified 87, 75, 50 and 371 pairs of significant relationships, as shown in **Table 7**. Venn diagram (**Figure 3**) reveals the relationships among the results as obtained using different methods in the PML samples. All of the significant relationships identified by TLTA are discovered by permutation tests. Among all these significant relationships, however, only 11 pairs of relationships are found out by both permutation test and STLTA. This is because there are only 33 (~44%) factors showing autocorrelation, with more than half of the factors bearing no autocorrelation. Therefore, permutation test can be conducted to find out about the significant relationships between many time series without autocorrelation. However, there are as few as 72 sample time points, since STLTA is conservative to some extent when there are a small number of time points. Among the significant relationships discovered by the permutation test, there are 76 pairs not identified by STLTA. In addition, it is suspected that 39 pairs of significant relationships which are found out by STLTA but fail to be detected by permutation test are between autocorrelation sequences, and these relationships can be discovered by neither permutation test nor TLTA. For more stringent standards $p = 0.01$ and $Q = 0.01$ as well as different thresholds, the results are shown in **Table 7**. It can be found out from the table that when $t = 0$, the number of significant relationships identified by all methods is smaller than that of relationships discovered when $t = 0.5$. As the PML data set has only 72 time points, there is a massive information loss in STLTA. Thus, the number of significant correlations between OTUs found by STLTA is far from that by DDLSA.

4 CONCLUSION

In this paper, a theoretical evaluation method was proposed for the statistical significance of local trend scores, STLTA. First of all, the original sequence was discretized into a changing trend

sequence and the local trend score was calculated. Then, according to the spectral decomposition theory of the matrix, the variance of the trend sequence was estimated for different state spaces. Finally, in combination with the limit theory of Markov chain local similarity analysis, the limit distribution of the local trend score was obtained, and the approximate p value of the local trend score was calculated. By means of simulation, it was discovered in a given stationary time series model that the type I error rate of STLTA can be made significantly closer to the given significance level, with the type I error rates of permutation test and TLTA increasingly deviant from the given significance level over time, especially when $t = 0.5$. It is suggested that STLTA method is more effective than permutation test and TLTA method. Then, these three methods were applied to the MPHM and PML data sets. In the relatively long data set MPHM "M3" fecal data set, STLTA detected the most significant relationships, and all of the significant relationships discovered by permutation tests and TLTA were identified by STLTA. In the PML data set with relatively short time points, STLTA discovered some relationships that cannot be found out by permutation tests and TLTA, with these relationships resulting from the autocorrelation of the sequence.

Compared with local similarity analysis, however, local trend analysis converts a continuous original time series into a discrete trend series, which may cause the loss of some information from the original series, thus limiting the practical application of local trend analysis. Nonetheless, the discretization of the original sequence may lead to the transformation of some non-stationary time series into a stationary Markov sequence, which is a major advantage of local trend analysis. In addition, the DDLSA based on non-parametric kernel estimation and the MBBLSA based on moving block bootstrap can be applied to the statistical significance analysis as part of local trend analysis, which provides another direction of further research.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The "MPHM" datasets used during the current study are publicly available in the supplementary of Gilbert et al. (2012), whose link is https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-5-r50#additional-information. The "PML" data can be found here: https://vamps2.mbl.edu/.

AUTHOR CONTRIBUTIONS

AS gave the main writing of the manuscript. FZ gave the main data analysis program of the manuscript. YL gave some idea and proofreading of the manuscript.

FUNDING

This work was supported by the National Science Foundation of China (Grant Number: 11971264) and the National Key R&D program of China (Grant Number: 2018YFA0703900).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.729011/full#supplementary-material

REFERENCES

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., and Kämper, J. (2005). Clustering of Gene Expression Data Using a Local Shape-Based Similarity Measure. *Bioinformatics* 21 (7), 1069–1077. doi:10.1093/bioinformatics/bti095

Beman, J. M., Steele, J. A., and Fuhrman, J. A. (2011). Co-occurrence Patterns for Abundant marine Archaeal and Bacterial Lineages in the Deep Chlorophyll Maximum of Coastal California. *ISME J.* 5 (7), 1077–1085. doi:10.1038/ismej.2010.204

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving Pictures of the Human Microbiome. *Genome Biol.* 12 (5), R50. doi:10.1186/gb-2011-12-5-r50

Cram, J. A., Xia, L. C., Needham, D. M., Sachdeva, R., Sun, F., and Fuhrman, J. A. (2015). Cross-depth Analysis of marine Bacterial Networks Suggests Downward Propagation of Temporal Changes. *ISME J.* 9 (12), 2573–2586. doi:10.1038/ismej.2015.76

Daudin, J.-J., Etienne, M. P., and Vallois, P. (2003). Asymptotic Behavior of the Local Score of Independent and Identically Distributed Random Sequences. *Stochastic Process. their Appl.* 107 (1), 1–28. doi:10.1016/s0304-4149(03)00061-9

Etienne, M. P., and Vallois, P. (2004). Approximation of the Distribution of the Supremum of a Centered Random Walk. Application to the Local Score. *Methodol. Comput. Appl. Probab.* 6 (3), 255–275. doi:10.1023/b:mcap.0000026559.87023.ec

Feller, W. (1951). The Asymptotic Distribution of the Range of Sums of Independent Random Variables. *Ann. Math. Statist.* 22 (3), 427–432. doi:10.1214/aoms/1177729589

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., et al. (2012). Defining Seasonal marine Microbial Community Dynamics. *Isme J.* 6 (2), 298–308. doi:10.1038/ismej.2011.107

Goncalves, J. P., and Madeira, S. C. (2014). LateBiclustering: Efficient Heuristic Algorithm for Time-Lagged Bicluster Identification. *Ieee/acm Trans. Comput. Biol. Bioinf.* 11 (5), 801–813. doi:10.1109/tcbb.2014.2312007

Gonçalves, J. P., Aires, R. S., Francisco, A. P., and Madeira, S. C. (2012). Regulatory Snapshots: Integrative Mining of Regulatory Modules from Expression Time Series and Regulatory Networks. *Plos One* 7 (5), e35977. doi:10.1371/journal.pone.0035977

He, F., Chen, H., Probst-Kepper, M., Geffers, R., Eifes, S., del Sol, A., et al. (2012). PLAU Inferred from a Correlation Network Is Critical for Suppressor Function of Regulatory T Cells. *Mol. Syst. Biol.* 8 (1), 624. doi:10.1038/msb.2012.56

He, F., and Zeng, A.-P. (2006). In Search of Functional Association from Time-Series Microarray Data Based on the Change Trend and Level of Gene Expression. *BMC Bioinformatics* 7, 69. doi:10.1186/1471-2105-7-69

Ji, L., and Tan, K.-L. (2004). Mining Gene Expression Data for Positive and Negative Co-regulated Gene Clusters. *Bioinformatics* 20 (16), 2711–2718. doi:10.1093/bioinformatics/bth312

Madeira, S. C., Teixeira, M. C., Sá-Correia, I., and Oliveira, A. L. (2010). Identification of Regulatory Modules in Time Series Gene Expression Data

Using a Linear Time Biclustering Algorithm. *Ieee/acm Trans. Comput. Biol. Bioinform* 7 (1), 153–165. doi:10.1109/TCBB.2008.34

Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., and Gerstein, M. (2001). Beyond Synexpression Relationships: Local Clustering of Time-Shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. *J. Mol. Biol.* 314 (5), 1053–1066. doi:10.1006/jmbi.2000.5219

Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A., and Sun, F. (2006). Local Similarity Analysis Reveals Unique Associations Among marine Bacterioplankton Species and Environmental Factors. *Bioinformatics* 22 (20), 2532–2538. doi:10.1093/bioinformatics/btl417

Seno, S., Takenaka, Y., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., et al. (2006). A Method for Similarity Search of Genomic Positional Expression Using CAGE. *Plos Genet.* 2 (4), e44. doi:10.1371/journal.pgen.0020044

Skreti, G., Bei, E. S., Kalantzaki, K., and Zervakis, M. (2014). Temporal and Spatial Patterns of Gene Profiles during Chondrogenic Differentiation. *IEEE J. Biomed. Health Inform.* 18 (3), 799–809. doi:10.1109/jbhi.2014.2305770

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., et al. (2011). Marine Bacterial, Archaeal and Protistan Association Networks Reveal Ecological Linkages. *ISME J.* 5 (9), 1414–1425. doi:10.1038/ismej.2011.24

Wu, L.-C., Huang, J.-L., Horng, J.-T., and Huang, H.-D. (2010). An Expert System to Identify Co-regulated Gene Groups from Time-Lagged Gene Clusters Using Cell Cycle Expression Data. *Expert Syst. Appl.* 37 (3), 2202–2213. doi:10.1016/j.eswa.2009.07.053

Xia, L. C., Steele, J. A., Cram, J. A., Cardon, Z. G., Simmons, S. L., Vallino, J. J., et al. (2011). Extended Local Similarity Analysis (eLSA) of Microbial Community and Other Time Series Data with Replicates. *BMC Syst. Biol.* 5 Suppl 2, S15. doi:10.1186/1752-0509-5-S2-S15

Xia, L. C., Ai, D., Cram, J. A., Liang, X., Fuhrman, J. A., and Sun, F. (2015). Statistical Significance Approximation in Local Trend Analysis of High-Throughput Time-Series Data Using the Theory of Markov Chains. *BMC Bioinformatics* 16, 301. doi:10.1186/s12859-015-0732-8

Zhang, F., Shan, A., and Luan, Y. (2018). A Novel Method to Accurately Calculate Statistical Significance of Local Similarity Analysis for High-Throughput Time Series. *Stat. Appl. Genet. Mol. Biol.* 17 (6), 20180019. doi:10.1515/sagmb-2018-0019

Zhang, F., Sun, F., and Luan, Y. (2019). Statistical Significance Approximation for Local Similarity Analysis of Dependent Time Series Data. *BMC Bioinformatics* 20, 53. doi:10.1186/s12859-019-2595-x

Conflict of Interest: Author AG is employed by Postdoctoral Programme of Zhongtai Securities Co. Ltd, Jinan, China

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership