INTEGRATIVE ANALYSIS OF GENOME-WIDE ASSOCIATION STUDIES AND SINGLE-CELL SEQUENCING STUDIES

EDITED BY: Sheng Yang, Shiquan Sun, Xiang Zhou and Yang Zhao
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-467-4
DOI 10.3389/978-2-88971-467-4

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

INTEGRATIVE ANALYSIS OF GENOME-WIDE ASSOCIATION STUDIES AND SINGLE-CELL SEQUENCING STUDIES

Topic Editors:
Sheng Yang, Nanjing Medical University, China
Shiquan Sun, Xi'an Jiaotong University, China
Xiang Zhou, University of Michigan, United States
Yang Zhao, Nanjing Medical University, China

Citation: Yang, S., Sun, S., Zhou, X., Zhao, Y., eds. (2021). Integrative Analysis of Genome-Wide Association Studies and Single-Cell Sequencing Studies. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-467-4

Table of Contents

frontiers in Genetics

EDITORIAL
published: 27 August 2021
doi: 10.3389/fgene.2021.752057

Editorial: Integrative Analysis of Genome-Wide Association Studies and Single-Cell Sequencing Studies

Shiquan Sun[1] *and Sheng Yang*[2]*

[1] School of Public Health, Health Science Center of Xi'an Jiaotong University, Xi'an, China, [2] Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, China

Keywords: genome-wide association analysis, single-cell RNA sequencing, integrative analysis, Mendelian randomization, transcriptome association analysis

Editorial on the Research Topic

Integrative Analysis of Genome-Wide Association Studies and Single-Cell Sequencing Studies

Genome-wide association studies (GWAS) have identified thousands of genetic loci that are significantly associated with complex traits and diseases status. However, the functions/roles of the majority (∼90%) of these associations remain poorly understood. Systematic characterization of their function is challenging because the function of variants of most traits likely acts in a tissue or cell-type specific fashion. The recent advances of single-cell sequencing technologies that enable characterization of epigenetic, proteomic, and transcriptomic profiles at individual cell, providing an unprecedented opportunity, alongside computational challenges, to comprehensively understand the functions/roles of associations in complex traits within the cellular perspective.

Therefore, integrating known functional cell-type specific annotations (e.g., cell-type specific expression levels etc.) into GWAS can potentially prioritize functional genetic variants and improve the performance of genomic predictions. Although various integrative analysis methods have been developed for such analyses, there is a pressing need to develop computationally scalable tools for large-scale GWAS, such as UK Biobank, China Kadoorie Biobank and FINNGEN. To address this need, this Research Topic focuses on integrative analysis to highlight the interpretation of genome-wide associations by leveraging the recent advances in single-cell sequencing studies.

In this special issue, we accepted 9 manuscripts on genome-wide association studies and/or single-cell sequencing in both methodology development and data analysis. We summarized the main contribution of these studies as follows:

Li et al. developed multiple computational approaches to deconvolve the bulk transcriptome data from whole kidney tissue with lupus nephritis (LN) into immune cell type-specific fractions and revealed that intrarenal mononuclear phagocytes might be an adjunctive histology marker for forecasting LN onset and retarded remission induction, which may facilitate on treatment and monitoring of LN patients.

Xiao et al. performed transcriptome-wide association study (TWAS) analysis on amyotrophic lateral sclerosis (ALS) and applied summary data-based Cauchy Aggregation TWAS (SCAT), a flexible p-value combination strategy, to integrate association signals from multiple brain tissues, and identified 5 new ALS-associated genes. Extensive simulations demonstrated that the proposed method can produce well-calibrated p-value for the control of type I error and more powerful to identify trait-association signals against single-tissue TWAS analysis.

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

****Correspondence:***
Sheng Yang
yangsheng@njmu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *02 August 2021*
Accepted: *09 August 2021*
Published: *27 August 2021*

Citation:
Sun S and Yang S (2021) Editorial: Integrative Analysis of Genome-Wide Association Studies and Single-Cell Sequencing Studies.
Front. Genet. 12:752057.
doi: 10.3389/fgene.2021.752057

Chen et al. performed genetic correlation analysis, gene-based association analysis, and pleiotropy-informed informatics analysis with coronary artery disease (CAD) and chronic kidney disease (CKD) related GWAS summary data, and identified common genetic architectures between the CAD and CKD, which may help to understand of the molecular mechanisms underlying the comorbidity of both diseases.

Gong et al. performed an integrative analysis of TWAS and mRNA expression profiles for idiopathic pulmonary fibrosis (IPF), and identified multiple novel candidate genes, GO terms and pathways for IPF, which would potentially contribute to the understanding of the genetic mechanism of IPF.

He et al. developed a new computational tool, single cell mixed model score tests (scMMSTs), to identify differentially expressed (DE) genes in single cell RNA sequencing (scRNA-seq) data with zero-inflation using the generalized linear mixed model (GLMM). Both simulations and real data analysis indicated that scMMSTs have more powerful performance in defining DE genes of zero-inflated scRNA-seq data with batch effects compared with the existing methods.

Ye et al. performed an integrative analysis on several GWAS and scRNA-seq data from chronic liver diseases (CLD), and identified B cell and NK cell as potential HCC-related cell types, which may supply clues for understanding the pathogenesis of CLD from a new angle.

Liu et al. developed a new agglomerative nesting clustering method for phenotypic dimensionality reduction analysis (AGNEP), which integrates agglomerative nesting clustering algorithm (AGNES) and principal component analysis (PCA) to detect genetic associations between SNPs and multiple phenotypes in GWAS. With extensive simulations and real data applications, AGNES shows more powerful performance in statistical power, computing time, and the number of quantitative trait nucleotides (QTNs).

Zhang et al. developed a flexible and scalable mixed linear model (MLM)-based method, the fast multi-locus ridge regression (FastRR), for QTNs dissection in GWAS. With simulations and real data applications, the results showed that the FastRR is more powerful for both large and small QTN detection, more accurate in QTN effect estimation, and has more stable results under various polygenic backgrounds.

Wang et al. developed a new deep convolutional neural network (CNN) of residual neural network (ResNet) on the whole-slide pathology features of breast cancer H&E stains and the patients' gBRCA mutation status, and the results demonstrated that the proposed method largely improve the prediction accuracy, which may potentially improve the cancer prognosis and therapeutics by utilizing biological markers currently imperceptible to clinicians.

With the further development of omics techniques and related analytical methods, integrative analysis on multiple omics data from the perspective of all in one will help to comprehensively understand the mechanism of complex traits and diseases status in large extent.

AUTHOR CONTRIBUTIONS

SS and SY wrote the editorial. Both authors have approved the submission.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 81703321 and 61902319).

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.

ORIGINAL RESEARCH
published: 09 November 2020
doi: 10.3389/fgene.2020.577040

Immune Cell Landscape Identification Associates Intrarenal Mononuclear Phagocytes With Onset and Remission of Lupus Nephritis in NZB/W Mice

Bin Li[1,2], *Yanlai Tang*[3], *Xuhao Ni*[4] *and Wei Chen*[1,2]*

[1] *Department of Nephrology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China,* [2] *Key Laboratory of Nephrology, National Health Commission and Guangdong Province, Guangzhou, China,* [3] *Department of Pediatrics, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China,* [4] *Department of Pancreato-Biliary Surgery, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China*

OPEN ACCESS

Edited by:
Sheng Yang,
Nanjing Medical University, China

Reviewed by:
Ni Zhao,
Johns Hopkins University,
United States
Ping Zeng,
Xuzhou Medical University, China

Correspondence:
Wei Chen
vvchen66@qq.com

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *28 June 2020*
Accepted: *07 October 2020*
Published: *09 November 2020*

Citation:
Li B, Tang Y, Ni X and Chen W (2020) Immune Cell Landscape Identification Associates Intrarenal Mononuclear Phagocytes With Onset and Remission of Lupus Nephritis in NZB/W Mice.
Front. Genet. 11:577040.
doi: 10.3389/fgene.2020.577040

Objective: A challenging issue in the clinical management of lupus nephritis (LN) is the resistance to immunosuppressive therapy. We postulated that perturbed intrarenal immune cell landscape affected LN onset and remission induction, and shedding light on the characteristics of intrarenal immune cell infiltration could cultivate more efficient treatment regimens.

Materials and Methods: Genome-wide expression profiles of microarray datasets were downloaded from the Gene Expression Omnibus database. The CIBERSORT algorithm was used to analyze the intrarenal immune cell landscape, followed by Pearson correlation analysis and principal component analysis. The differentially expressed genes were identified and subjected to Gene Ontology (GO) enrichment analyses and protein-protein interaction network establishment, being visualized by Cytoscape and further analyzed by CytoHubba to extract hub genes. Hub genes were also validated in the genomic dataset from kidney biopsy-proven LN patients.

Results: In addition to memory B cells, monocytes and M1 macrophages were identified as two predominantly increased intrarenal immune cell types in LN-prone NZB/W mice upon nephritis onset. Most interestingly, apart from memory B cells, monocytes and M1 macrophages proportions in kidney tissue were significantly lower in early remission mice compared with late remission mice. Furthermore, GO analysis showed that intrarenal mononuclear phagocytes triggered nephritis onset mainly via the initiation of adaptive immune response and inflammatory reaction, but this functional involvement was mitigated upon remission induction. Hub genes related to LN onset in NZB/W mice were validated in the genomic dataset from kidney biopsy-proven LN patients.

Conclusion: LN characterizes aberrant mononuclear phagocytes abundance and signature upon disease onset, of which the reversal is associated with early remission induction in LN-prone NZB/W mice. Mononuclear phagocytes might be an adjunctive histology marker for monitoring disease onset and stratifying LN patients in terms of response to remission induction therapy.

Keywords: lupus nephritis, immune cell landscape, mononuclear phagocytes, NZB/W mice, CIBERSORT

INTRODUCTION

Lupus nephritis (LN) predominantly manifests as immune complex-mediated glomerular and tubulointerstitial immune complex deposition and inflammation. LN occurs in 40–60% of systemic lupus erythematosus (SLE) patients and accounts for one of the most prevalent and serious organ complications during the course of their disease (Lisnevskaia et al., 2014; Hanly et al., 2016). Introduction of corticosteroids and other immunosuppressants have profoundly contributed to improved treatment of LN. Nevertheless, a challenging issue in clinical practice is that 20% to 70% of patients diagnosed with LN are documented to be resistant to standard immunosuppressive regimens (Ginzler et al., 2005). Despite the absence of consensus denotation for a complete response after induction therapy, refractory LN is commonly denoted as failure to achieve clinical remission following proper induction immunosuppressive treatment (Chen et al., 2008). During the past decade, some clinical trials aimed to optimize therapeutic approaches by prolonging therapeutic course, increasing glucocorticoid dosage, or adding a calcineurin inhibitor ended up with varying success. In the meantime, although biomarkers for nephritis occurrence have increasingly being identified, a reliable way of predicting or determining which kind of patients will respond to induction therapy is scarce. Therefore, a more comprehensive understanding toward the mechanisms of refractory LN is demanded to generate highly specific predictive biomarkers and highly efficacious therapeutic approaches. Murine models that spontaneously develop SLE contribute a lot to our understanding of human disease and are extensively employed for the identification of effective therapeutics. Notably, the NZB/W F1 mice (F1 hybrid between New Zealand Black mouse and New Zealand White mouse, hereafter referred to as NZB/W mice) is featured by hypercellular renal impairment and fibrinoid necrosis, resembling the lesions that occur in human LN kidney biopsies. Intriguingly, induction therapy using a single dose of cyclophosphamide (CYC) administered in combination with six doses of CTLA4Ig (cytotoxic T-lymphocyte-associated protein 4 immunoglobulin G) and six doses of anti-CD154 (triple therapy) promptly reverses albuminuria and stabilizes kidney function in NZB/W mice with established nephritis (Schiffer et al., 2003), which has made it possible to dig deeper into the underlying mechanisms linked with remission induction of glomerulonephritis.

Efforts over the past decade have emphasized the critical role of innate immune cells in promoting and potentiating LN. For example, numerous studies indicate that glomerulonephritis in LN is attributable to a systemic breakdown of B cell tolerance that results in the local precipitation of immune complexes; thus, B cell targeted therapeutic strategies such as depleting B cell or blocking B cell survival factors are theoretically promising and have also been developed accordingly. Nevertheless, the efficacy of therapies targeting B cells still remains disputable (Liossis and Staveri, 2017), because several studies documented favorable outcomes in LN, while some other studies observed no clinical refinement (Melander et al., 2009; Duxbury et al., 2013). Therefore, a more thorough understanding of immune cell in the pathogenesis of LN is needed for yielding therapeutic choice with higher efficiency. Current knowledge of the subpopulations of infiltrating immune cell in LN comes mainly from the immunohistochemistry and flow cytometry studies of kidney biopsies; however, the heterogeneity of subpopulations in different disease stages remains enigmatic. Particularly, the role and mechanism of these subpopulations of infiltrating immune cell in the progress of remission induction remain unrevealed, although accumulating evidence proposes the intrarenal infiltrating immune cell as an essential factor associated with response upon immunosuppressive therapy (Melander et al., 2009; Duxbury et al., 2013; Liossis and Staveri, 2017). Hence, there is an imperative urge to unravel the immunologic mechanisms bridging immune cell state with LN progression and remission induction. These efforts may replenish key insights to precipitate better disease predictors and better-designed drugs targeting to tame LN autoimmunity.

Bioinformatics is emerging as a new interdisciplinary subject that has enabled the high-throughput and high-efficacy collection of biological information. Remarkably, the deconvolution techniques can yield surplus insight into the abundant variation of specific cell types that arise at different disease stages throughout onset, development, and treatment, thus allowing earlier and more accurate diagnosis of comorbidities and prediction of therapeutic response. For example, a newly developed bioinformatic approach, Cell-type Identification By Estimating Relative Subsets Of known RNA Transcripts (CIBERSORT) deconvolution algorithm method[1], has been successfully applied to assess the levels of 22 kinds of immune cell types in large amounts of heterogeneous samples based on gene expression profiles, allowing large-scale interpretation of mRNA compound for defining novel cellular biomarkers and therapeutic targets. On the other hand, even though several recent studies have contributed to defining the immune cell infiltration state milieu of LN by integrated bioinformatic analyses (Arazi et al., 2019; Cao et al., 2019), transcriptional signatures of infiltrating immune cell landscape that distinguish

[1] https://cibersort.stanford.edu/

LN subgroups with the varied response to remission induction have not yet been described.

We hypothesized that intrarenal immune infiltration state possibly has predictive value in delineating the response to therapy. Thus, to clarify the association between immune infiltration and LN onset as well as remission induction, the microarray datasets of LN-prone NZB/W mice at various disease stages in Gene Expression Omnibus (GEO) were dissected by a variety of bioinformatics techniques including CIBERSORT to define distinct subgroups. Findings from this study correlate LN onset with exuberant mononuclear phagocytes abundance and signature, whose reversal is associated with early remission induction. This explorative study extends our knowledge about mononuclear phagocytes as a future platform for diagnosis and precision medicine in LN.

MATERIALS AND METHODS

Overview of Microarray Datasets Collection

The diagram of the overall study design and analysis process is displayed in **Figure 1**. We screened the qualified datasets that contained comprehensive intrarenal gene-expression profiles of kidneys from SLE-prone murine models or human LN patients, because these dataset types can help determine intrarenal immune cell landscape at different disease stages of LN. As a result, four independent LN gene expression profiles (GSE32583, GSE49898, GSE27045, and GSE32591) were downloaded from the GEO database and exploited to identify or validate differentially expressed genes (DEGs). **Supplementary Table S1** provides additional information about all of the above four datasets. A detailed description of murine and human RNA extraction, microarray preparation and processing, as well as gene-expression data processing and analysis could be retrieved from the corresponding original literature (Schiffer et al., 2003; Reddy et al., 2008; Bethunaickan et al., 2011, 2014; Berthier et al., 2012).

In GSE32583 dataset, the NZB/W mice were allocated into pre-nephritis control group (without proteinuria) and nephritis group (proteinuria > 300 mg/dl). Therefore, GSE32583 was used to discover and compare intrarenal gene expression between pre-nephritis and nephritis mice. The expression matrix of 19 pre-nephritis mice (GSM807484–GSM807502) and 16 nephritis mice (GSM807503–GSM807518) were obtained to profile the infiltrating immune cell (Schiffer et al., 2003; Berthier et al., 2012).

In GSE49898 dataset, nephritic NZB/W mice (proteinuria > 300 mg/dl) were treated with a single dose of CYC and 6 doses of CTLA4Ig and anti-CD154. Mice that attained proteinuria ≤ 30 mg/dl within 3–4 weeks post induction treatment fell into the early remission group and showed complete histologic remission. By contrast, mice that obtained proteinuria ≤ 30 mg/dl more than 5–14 weeks post induction treatment fell into the late remission group and displayed only partial histologic remission by light microscopy (Chan et al., 1997). Therefore, GSE49898 was employed to compare the intrarenal gene expression among nephritis, early remission, and late remission NZB/W mice. After data processing, the expression matrix of 7 nephritis mice (GSM1209137–GSM1209143), 7 early remission mice (GSM1209145–GSM1209151), and 11 late remission mice (GSM1209152–GSM1209162) were obtained to profile the infiltrating immune cell (Bethunaickan et al., 2014).

Similarly, in GSE27045, nephritic NZB/W mice were treated with a single dose of CYC and 6 doses of CTLA4Ig and anti-CD154 if proteinuria > 300 mg/dl occurred. Remission was defined as proteinuria ≤ 30 mg/dl, and some young NZB/W mice were allocated into pre-nephritis group. The F4/80^{hi} (a classic mononuclear phagocyte marker) mononuclear phagocytes, which acquire an activated phenotype during active nephritis and reverse upon remission induction, were sorted and isolated by flow cytometry from single-cell suspensions of perfused kidneys. Therefore, GSE27045 was utilized to compare the gene expression of F4/80^{hi} intrarenal mononuclear phagocytes among pre-nephritis, nephritis, and remission NZB/W mice. The expression matrix of 6 pre-nephritis mice (GSM667532–GSM667537), 7 nephritis mice (GSM667538–GSM667544) and 4 remission mice (GSM667545–GSM667548) were obtained for defining mononuclear phagocytes-derived genes associated with LN onset and remission induction (Bethunaickan et al., 2011).

In GSE32591, a total of 47 renal biopsies from the European Renal cDNA Bank (Schmid et al., 2006) were collected according to the guidelines of the respective local ethics committees. The demographic, clinical, and histologic characteristics of the included patients could be retrieved from the original literature (Berthier et al., 2012). Therefore, GSE32591 was exploited to investigate and compare the gene expression of glomeruli and tubulointerstitial compartments of renal biopsies from LN patients ($n = 32$) and pretransplant healthy living donors ($n = 15$). The expression matrix of renal tubulointerstitial compartment including 32 LN patients (GSM807842–GSM807873) and 15 healthy living donors (GSM807874–GSM807888), as well as renal glomeruli compartment including 32 LN patients (GSM807889–GSM807920) and 14 healthy living donors (GSM807921–GSM807934), were obtained for validation analyses in our current study (Berthier et al., 2012).

Evaluation of Immune Cell Infiltration by CIBERSORT Analyses

To determine the immune cell landscape in kidney tissues, the analytical platform CIBERSORT (see footnote 1) with the reference of 1000 permutations and LM22 signature was employed. The CIBERSORT deconvolution algorithm has been validated to accurately and reliably calculate 22 types of immune cell fractions dependent on microarray expression data. These immune cells are composed of naive B cells, memory B cells, plasma cells, CD8+ T cells, naive CD4+ T cells, resting memory CD4+ T cells, activated memory CD4+ T cells, follicular helper T cells, regulatory T cells (Tregs), gamma delta T cells, resting NK cells, activated NK cells, monocytes, M0 macrophages, M1 macrophages, M2 macrophages, resting dendritic cells,

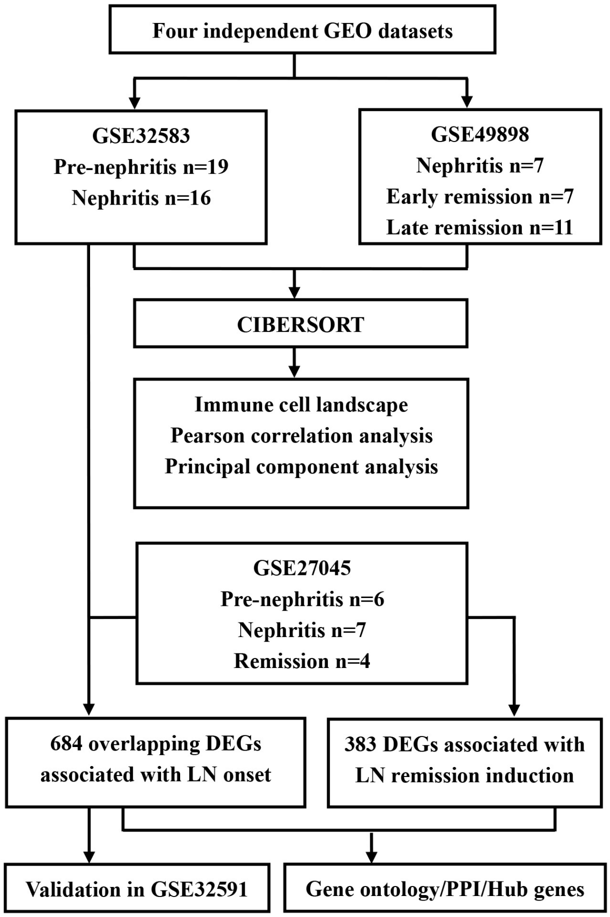
Four independent GEO datasets
GSE32583
Pre-nephritis n=19
Nephritis n=16
GSE49898
Nephritis n=7
Early remission n=7
Late remission n=11
CIBERSORT
Immune cell landscape
Pearson correlation analysis
Principal component analysis
GSE27045
Pre-nephritis n=6
Nephritis n=7
Remission n=4
684 overlapping DEGs associated with LN onset
383 DEGs associated with LN remission induction
Validation in GSE32591
Gene ontology/PPI/Hub genes

FIGURE 1 | Flowchart of the analyses used in this study. GEO, Genome Expression Omnibus; DEGs, differentially expressed genes; PPI, protein-protein interaction.

activated dendritic cells, resting mast cells, activated mast cells, eosinophils, and neutrophils. The significant alteration of immune cell fractions was recognized according to the threshold of the Wilcoxon test at p-value < 0.05. Associations between different immune cell subtypes were evaluated via Pearson correlation coefficient.

Principal Component Analysis

Principal component analysis (PCA) is often used as a technique in exploratory data analysis for variable dimensionality reduction. Therefore, PCA was utilized in the current study to ascertain primary sources of variance in the fraction of diverse infiltrating immune cell types among different groups, and the prominent sources of variance can likely be the diagnostic clues for LN onset or predictive biomarkers for early LN remission induction. To be specific, log-ratio PCA is proposed as an efficient tool for the exploration of compositional data (Graffelman et al., 2019); thus, we followed that approach by applying the centered log-ratio (*clr*) transformation to the compositional data. In brief, the compositional data of immune cell fractions derived from CIBERSORT was expressed in isometric coordinates. Afterward, PCA was performed to decompose the normalized, log10-transformed immune cell composition matrix by using the dudi.pca function in R. Resulting loadings and scores were back-transformed to the *clr* space where the compositional biplot could be shown.

Identification of DEGs and Functional Enrichment Analyses

An R-based web application, GEO2R, was employed to obtain DEGs in GSE datasets by comparing the expression values among different subgroups and using the GEOquery and the linear models for microarray data (LIMMA) package of R. The adjusted p-values (calculated by Benjamini and Hochberg false discovery rate method) via GEO2R tool were adopted to avoid the occurrence of false-positive results. DEGs (adjusted p-value < 0.05) between pre-nephritis group and nephritis group mice were identified in GSE32583 and GSE27045 dataset, respectively. GSE27045 dataset contained the genomic profile of kidney-isolated F4/80^{hi} mononuclear phagocytes, hence the overlapping DGEs (adjusted p-value < 0.05) between GSE27045 and GSE32583 datasets were further identified to define mononuclear phagocytes-specific DEGs associated with LN onset. On the other hand, DEGs (adjusted p-value < 0.05) between nephritis and remission group mice were identified in GSE49898 and GSE27045 dataset, respectively. However, statistical analyses showed that there were no DEGs (adjusted p-value < 0.05) between nephritis and early remission NZB/W mice in GSE49898 dataset, hence F4/80^{hi} mononuclear phagocytes-specific DEGs (adjusted p-value < 0.05, and $|\log2$ fold change (FC)$| > 1$) between nephritis and remission NZB/W mice in GSE27045 dataset were further pinpointed and deemed as mononuclear phagocytes-derived genes associated with LN remission induction. These DEGs or overlapping DEGs were included for Gene Ontology (GO) enrichment analyses through WebGestalt (WEB-based Gene SeT AnaLysis Toolkit), and Ggplot2 of R was applied to draw heatmap for visualization of the overlapping DEGs. Furthermore, protein-protein interaction networks of the overlapping DEGs were established via STRING (Search Tool for the Retrieval of Interacting Genes database) online tool and visualized in Cytoscape software. Hub genes with a high degree of connectivity were extracted by applying the plug-in of CytoHubba. LN onset-related hub genes were also validated in GSE32591 dataset that included kidney biopsies from LN patients.

RESULTS

Composition of Immune Cell Between Pre-nephritis and Nephritis NZB/W Mice by CIBERSORT

To determine whether renal infiltrating immune cell corresponded with nephritis onset, CIBERSORT was utilized to quantify the immune cell proportions within kidney samples (GSE32583 dataset). Compared to pre-nephritis mice, nephritis NZB/W mice were characterized by obviously lower proportions in naïve B cells, follicular helper T cells, and activated NK cells, but notably higher proportions in memory B cells, monocytes, and M1 macrophages (**Figures 2A,B** and **Supplementary Table S2**). Particularly, monocytes accounted for the highest proportion and the most pronounced elevation among all the immune cell types in the nephritic kidney from NZB/W mice (**Figures 2A,B**). This result is in line with a very recent bioinformatic study that recognized monocytes as the most significantly increased and the most abundant infiltrating immune cell type in kidney biopsy from human LN subjects (Cao et al., 2019). Furthermore, a significantly positive correlation between monocytes and M1 macrophages (r = 0.40) was presented by the correlation analyses (**Figure 2C**). Intriguingly, both the percentages of monocytes and M1 macrophages were significantly negatively correlated with naïve B cells ($r = -0.38$ and -0.48 respectively), follicular helper T cells ($r = -0.47$ and -0.34 respectively), activated NK cells ($r = -0.57$ and -0.44 respectively), and resting mast cells ($r = -0.56$ and -0.54 respectively), but significantly positively correlated with gamma delta T cells ($r = 0.39$ and 0.59 respectively) (**Figure 2C**). The broadly notable correlations between monocytes/M1 macrophages and other immune cell types underpin the critical role of mononuclear phagocytes in orchestrating LN occurrence.

PCA was subsequently performed to evaluate if the fractions of infiltrating immune cell could be exploited to distinguish the diagnosis of LN onset. M1 macrophages, memory B cells, gamma delta T cells, and monocytes were found to be the major components of principal component (PC) 1, and they were positively associated with LN onset (**Figure 3A**). Nevertheless, data visualization by PCA in **Figure 3B** shows that the first two PCs explained only 8.5% of the variance, indicating the mere composition of immune cells in LN kidney tissue was insufficient to discriminate pre-nephritis from nephritis mice. Taken together, the above results suggested aberrant immune infiltration in LN kidney tissues as a tightly regulated process that influenced the pathogenesis of LN onset. Specifically, it is worth noting that mononuclear phagocytes including monocytes and M1 macrophages were dramatically augmented in the nephritic kidney from NZB/W mice, making the mononuclear phagocytes abundance a potential adjuvant diagnostic biomarker for LN occurrence.

Composition of Immune Cell Between Early Remission and Late Remission NZB/W Mice by CIBERSORT

In order to examine whether the renal infiltrating immune cells affected nephritis remission upon immunosuppressive therapy, CIBERSORT was further utilized to profile the immune cell landscape within kidney samples from LN mice at different disease stages (GSE49898 dataset). Consistent with the result from CIBERSORT analyses in GSE32583 (**Figures 2A,B**), monocytes still accounted for the highest proportion among all the immune cells in nephritic kidney (**Figures 4A,B** and **Supplementary Table S3**). More importantly, in comparison with late remission mice, early remission mice attained noticeably higher proportions of naïve B cells, activated NK cells, but significantly lower proportions of memory B cells, monocytes, and M1 macrophages (**Figures 4A,B** and **Supplementary Table S3**). Despite the suppressed memory B cells, the notably decreased percentages of mononuclear phagocytes including monocytes and M1 macrophages in early remission kidney samples indicated an essential role of attenuating mononuclear phagocytes abundance in contributing to early response upon immunosuppressive therapy. Besides, the positive correlation between monocytes and M1 macrophages was also present in the GSE49898 dataset (**Figure 4C**), consistent with the results from GSE32583 as mentioned earlier. The percentages of monocytes and M1 macrophages were found negatively correlated with naïve B cells ($r = -0.67$ and -0.64 respectively), plasma cells ($r = -0.67$ and -0.74 respectively), regulatory T cells (Tregs) ($r = -0.41$ and -0.54 respectively), activated NK cells ($r = -0.66$ and -0.68 respectively), resting mast cells ($r = -0.61$ and -0.58 respectively), but positively correlated with memory B cells ($r = 0.73$ and 0.75 respectively) and activated memory CD4 T cells ($r = 0.41$ and 0.54 respectively) (**Figure 4C**). The positive correlation between mononuclear phagocytes and B or T cells was in line with previous evidence showing the full capacity of mononuclear phagocytes to facilitate B and T cell responses and orchestrate adaptive autoimmune response (Gkirtzimanaki et al., 2018).

Furthermore, PCA results demonstrated that memory B cells, M1 macrophages, monocytes, and gamma delta T cells were the major components of PC1 that were negatively associated with remission induction upon immunosuppressive therapy (**Figure 5A**). These results coordinated with findings in **Figure 3A** to suggest those four immune cell types as the major components that contributed to LN onset and hampered early remission induction. However, the first two PCs only explained 10.3% variation (**Figure 5B**), suggesting that the mere composition of immune cells in LN kidney tissue was not compelling to distinguish early remission mice from nephritis or late remission mice. Collectively, the above results implied that dysregulated immune infiltration in LN might convey important meanings for predicting the response to immunosuppressive therapy. Particular attention should be paid that the fraction of mononuclear phagocytes including monocytes and M1 macrophages were significantly lower in early remission LN mice, indicating the possibility of proposing mononuclear phagocytes abundance as a predictive marker for discovering potential refractory LN patients.

GO Analyses of Mononuclear Phagocytes-Specific DEGs Associated With LN Onset

Considering the notable amplification of mononuclear phagocytes abundance upon nephritis onset, we sought to examine the functional involvement of mononuclear phagocytes in the development of LN through bioinformatic analyses. Interestingly, GSE27045 encompassed the microarray data from NZB/W mice kidney-isolated F4/80^{hi} (a broadly accepted mononuclear phagocyte marker) mononuclear phagocytes (Bethunaickan et al., 2011; Waddell et al., 2018), which were the

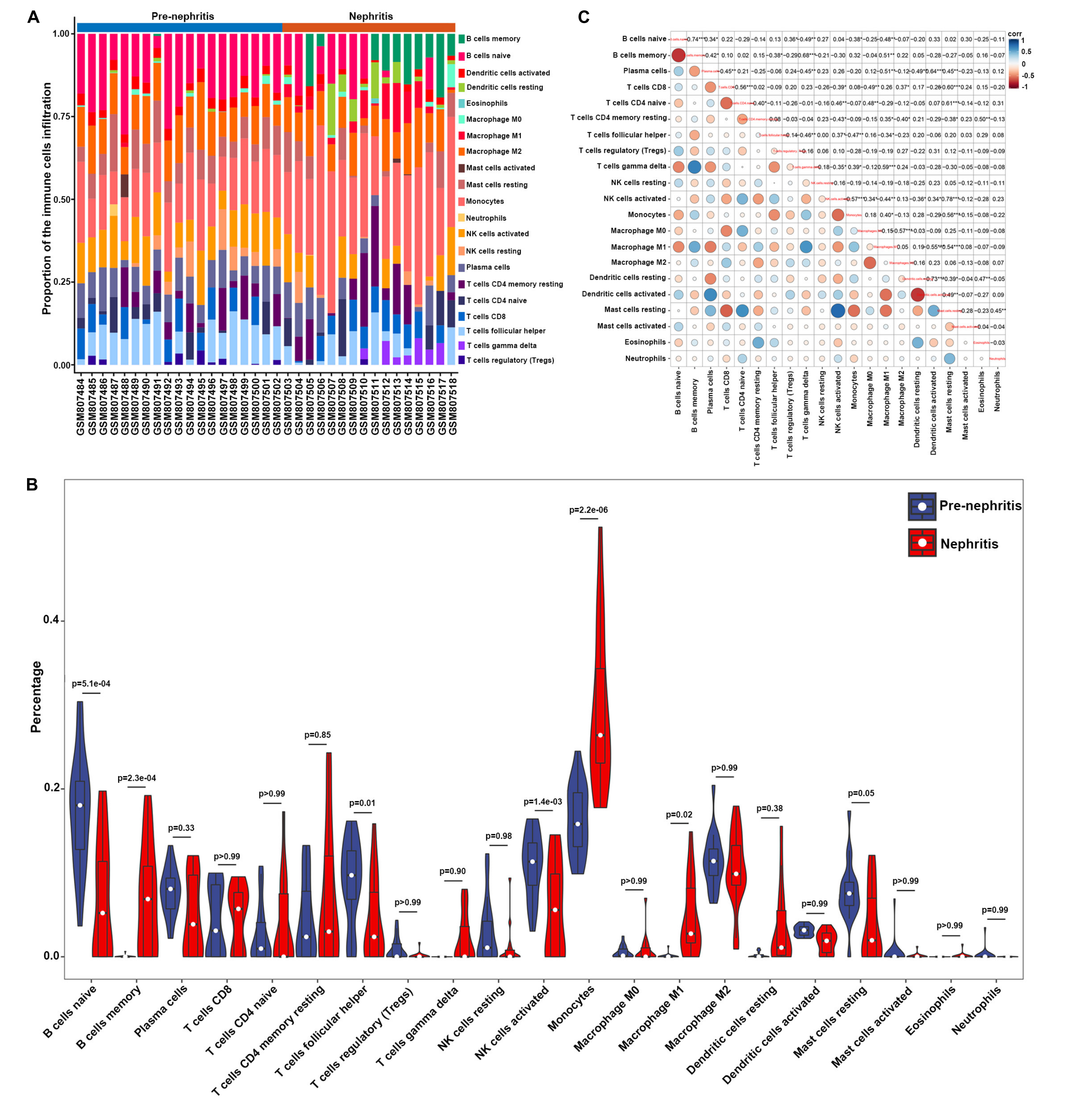

FIGURE 2 | Composition of infiltrating immune cell subpopulations in kidney tissues from pre-nephritis and nephritis NZB/W mice in GSE32583 dataset. **(A)** The fraction of infiltrating immune cell subpopulations was determined by CIBERSORT. **(B)** Comparison of renal immune infiltration between pre-nephritis and nephritis NZB/W mice. **(C)** Correlation among infiltrating immune cell subpopulations.

dominant intrarenal origin of proinflammatory cytokines and chemokines. These F4/80^{hi} mononuclear phagocytes acquired an activated phenotype during active nephritis and reversed upon remission induction (Bethunaickan et al., 2011). Hence we assumed that the overlapping DEGs between GSE27045 and GSE32583 could represent the mononuclear phagocytes-specific genes closely associated with LN onset. Accordingly, 684 overlapping DEGs (adjusted p-value < 0.05) between GSE27045 and GSE32853 (**Supplementary Table S4**) were identified for subsequent functional enrichment analysis.

GO enrichment analysis found that the overlapping DEGs between GSE27045 and GSE32583 were mainly enriched in the

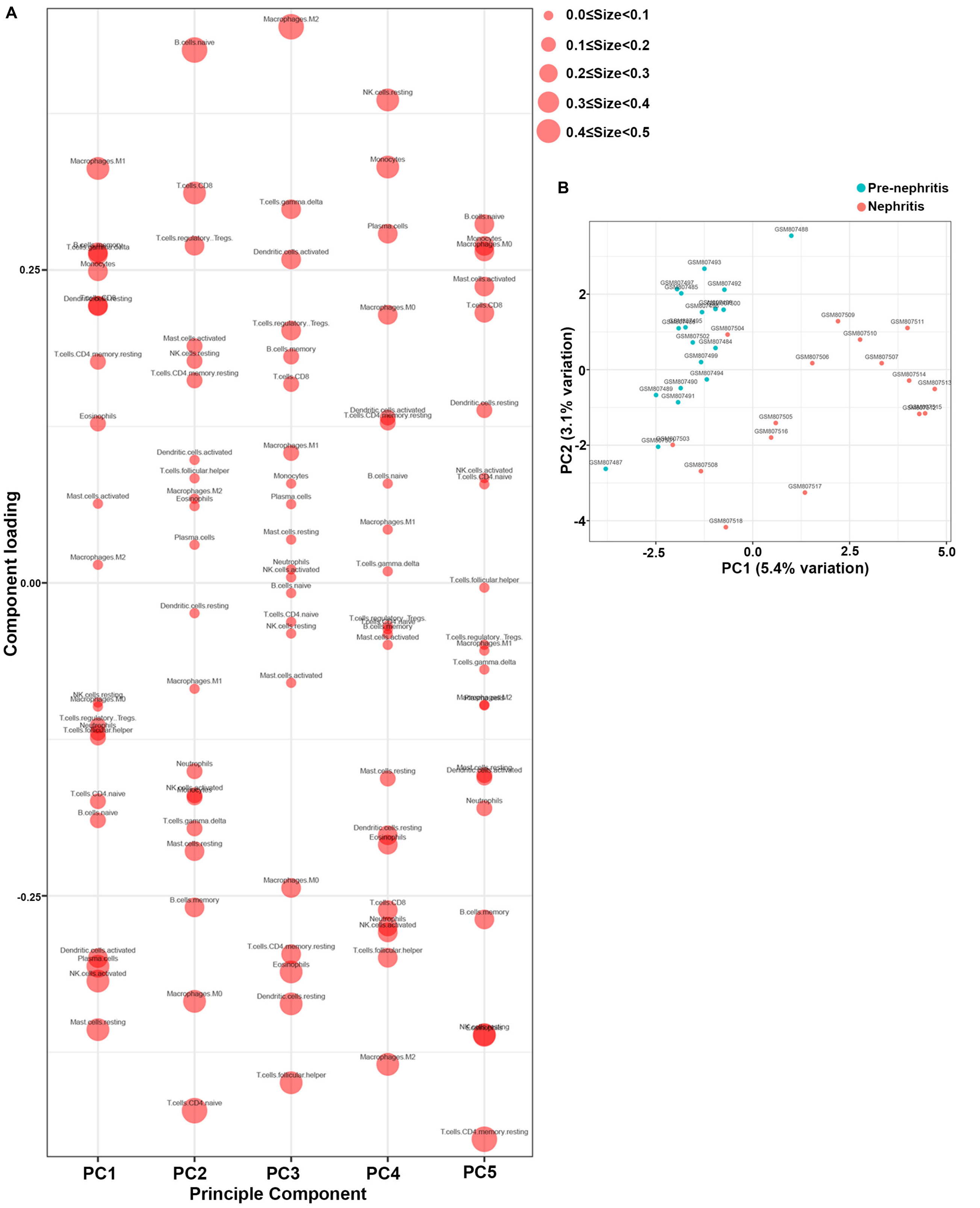

FIGURE 3 | Principal component analysis (PCA) was performed to reveal differences in immune cell landscape between pre-nephritis and nephritis NZB/W mice in GSE32583 dataset. **(A)** Component loading in PCA results. **(B)** Score plot for PC1 and PC2. The percentages of variance explained by PC1 and PC2 are in the axis labels. PC, principal component.

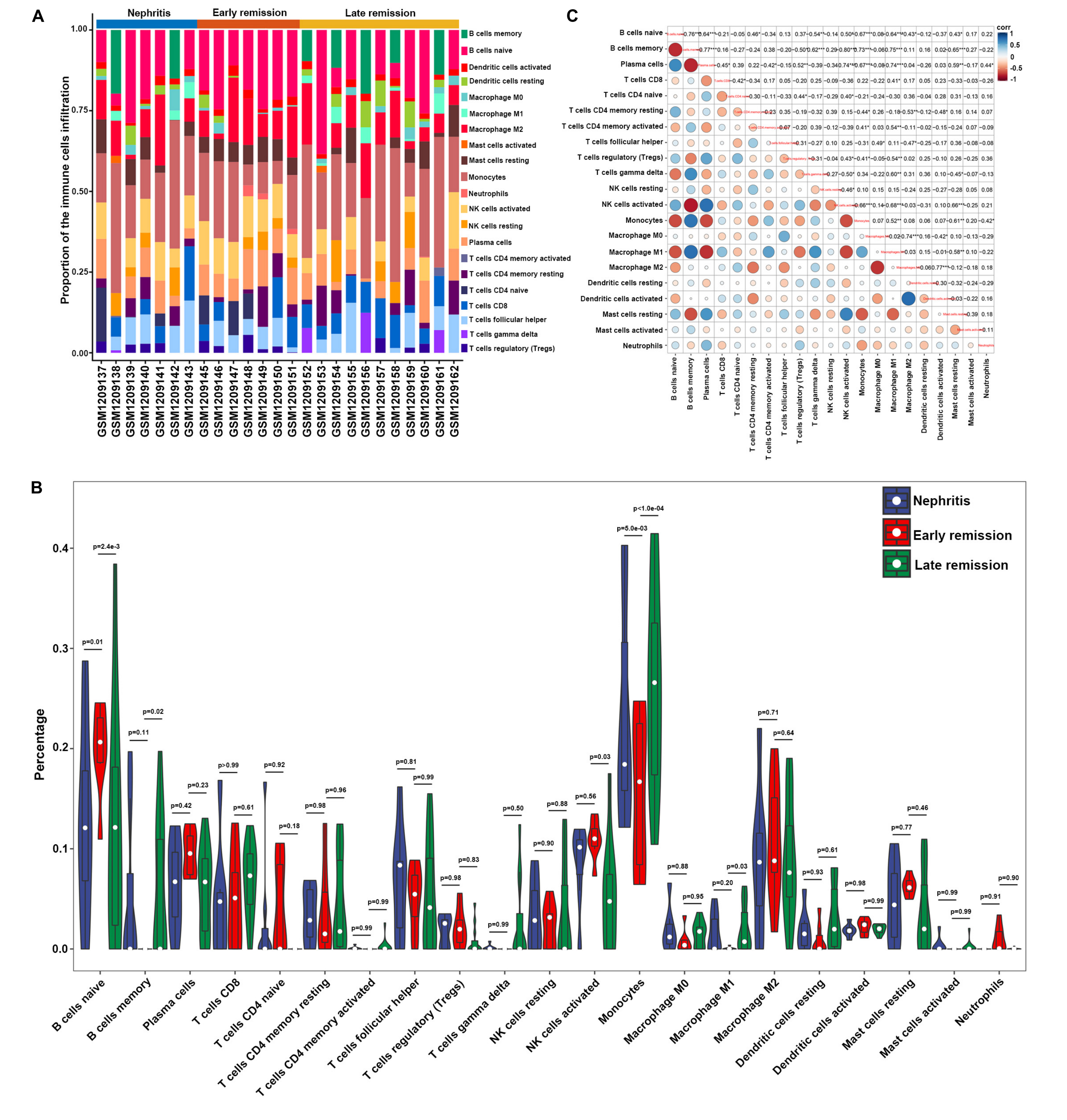

FIGURE 4 | Composition of infiltrating immune cell subpopulations in kidney tissues from nephritis, early remission, and late remission NZB/W mice in GSE49898 dataset. **(A)** The fraction of infiltrating immune cell subpopulations was determined by CIBERSORT. **(B)** Comparison of renal immune infiltration among nephritis, early remission, and late remission NZB/W mice. **(C)** Correlation among infiltrating immune cell subpopulations.

mobilization of adaptive immune response and proinflammatory reaction, with the top three enriched GO terms being T cell activation, regulation of immune effector process, and positive regulation of cytokine production (**Figure 6A**). This result underpinned the vital role of mononuclear phagocytes signature in mobilizing the intrarenal adaptive immune response, thus cultivating the inflammatory reaction during the development of LN onset. In addition, 20 hub genes were extracted from a constructed protein-protein interaction network established by these overlapping DEGs (**Figure 6B** and **Supplementary Table S5**). Heatmap showed that most of these hub genes were markedly increased in nephritic kidney samples, except Cd28

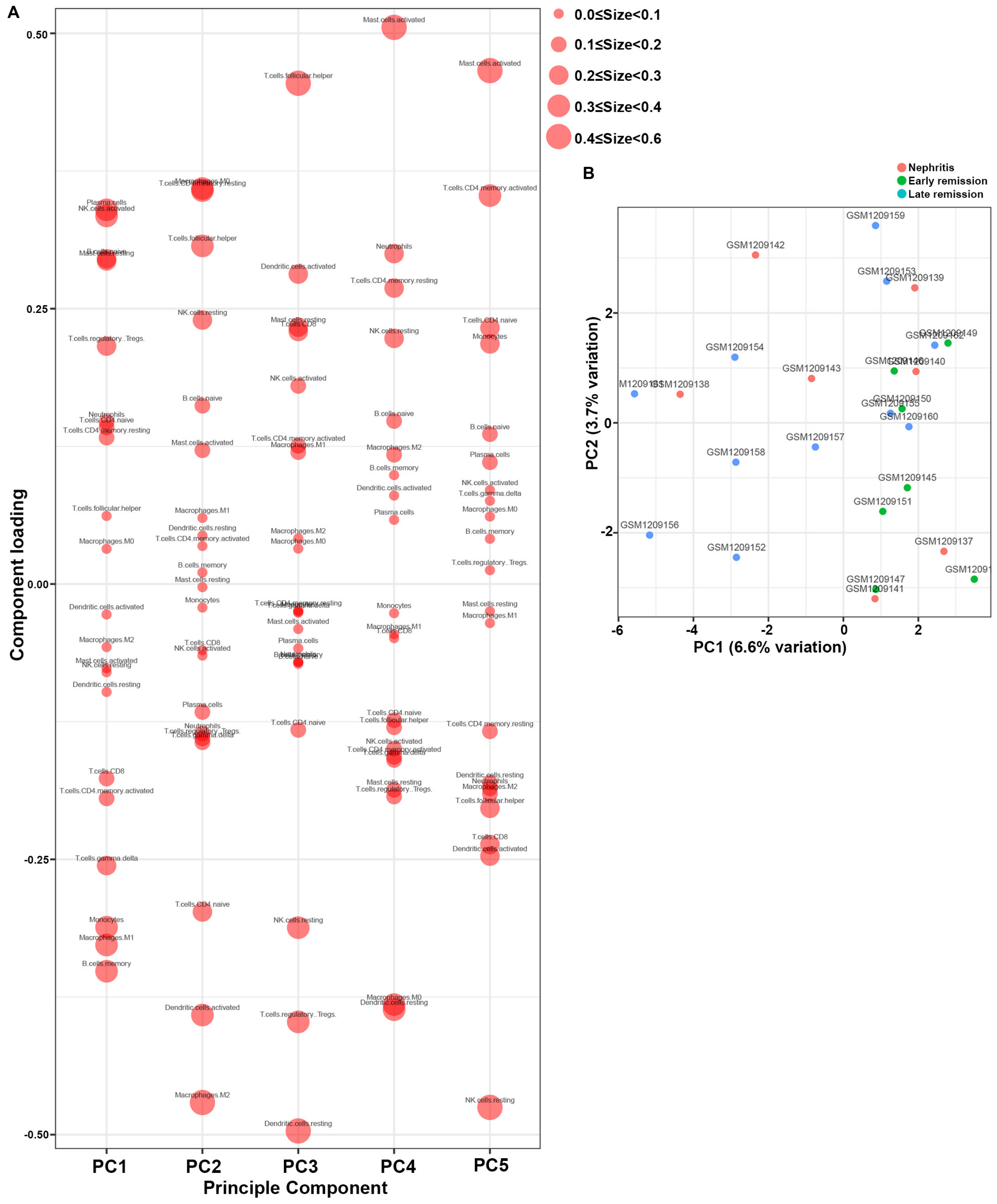

FIGURE 5 | Principal component analysis (PCA) was applied to reveal differences of immune cell landscape among nephritis, early remission, and late remission NZB/W mice in GSE49898 dataset. **(A)** Component loading in PCA results. **(B)** Score plot for PC1 and PC2. The percentages of variance explained by PC1 and PC2 are in the axis labels. PC, principal component.

that were downregulated (**Figure 6C**). In particular, dramatically increased hub genes like CD40 (cluster of differentiation 40), Itgam (integrin subunit alpha M), C3 (complement 3), and Myd88 (myeloid differentiation primary response 88) have previously been proven to play essential roles in the expansion and activation of B and T cells (Zarnegar et al., 2004;

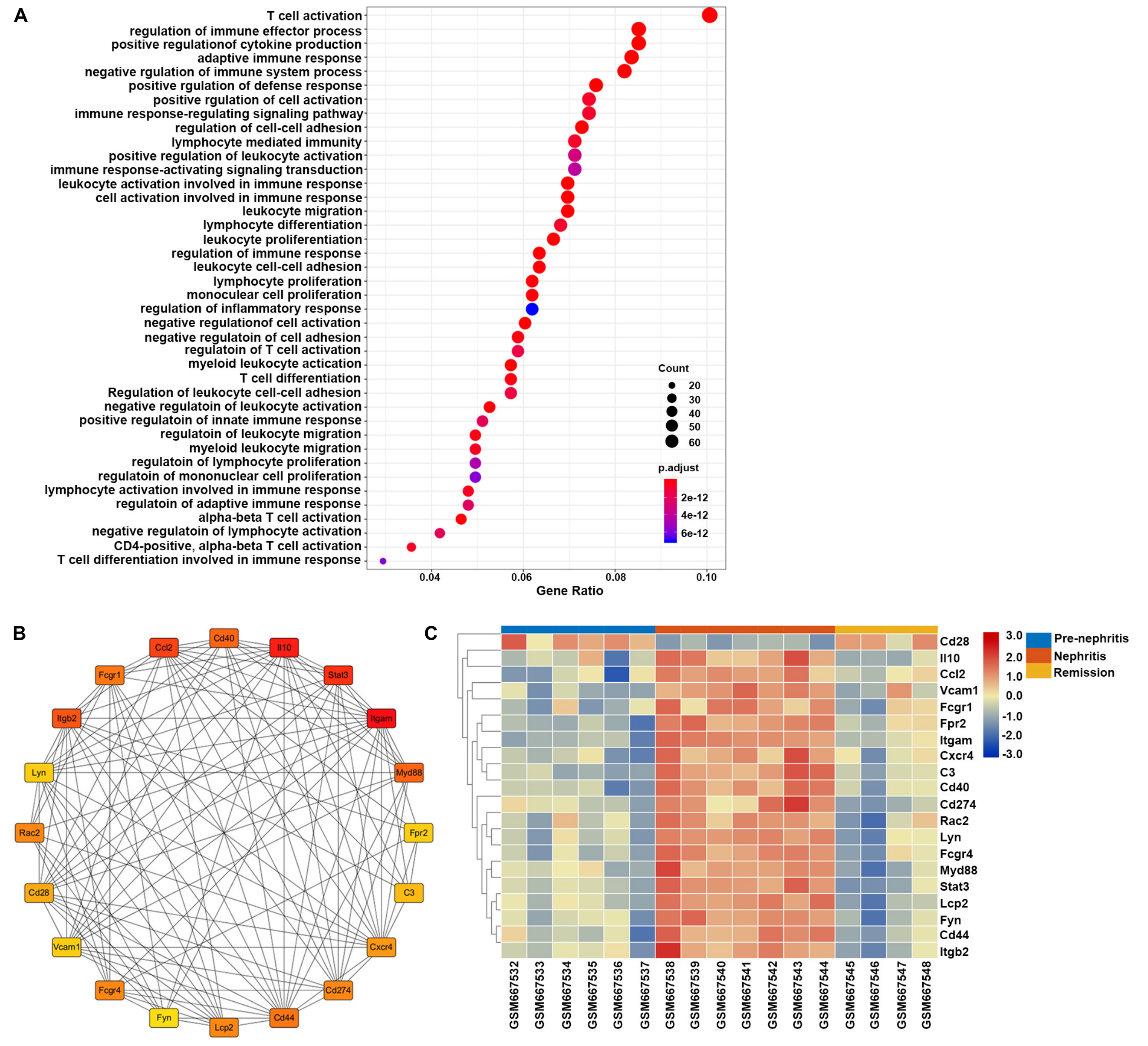

FIGURE 6 | Gene set enrichment analysis of the overlapping DEGs between GSE27045 and GSE32583, which were deemed as genes linked with LN onset in NZB/W mice. **(A)** GO enrichment analysis of overlapping DEGs between GSE27045 and GSE32583. The *y*-axis labels represent clustered GO terms, and the Gene Ratio represents the ratio of the number of genes enriched in one GO term to the number of DEGs. **(B)** Network analysis of identified hub genes from overlapping DEGs between GSE27045 and GSE32583. **(C)** Heatmap of identified hub genes from overlapping DEGs between GSE27045 and GSE32583. GO, gene ontology; DEGs, differentially expressed genes.

Quigley et al., 2009; Griffin and Rothstein, 2011; Liszewski et al., 2013), supporting the role of mononuclear phagocytes signature in bridging innate immune response with the adaptive autoimmune response in the context of LN. Meanwhile, elevated hub genes like Ccl2 (C-C motif ligand 2), Cxcr4 (C-X-C chemokine receptor type 4), Il10 (interleukin 10), and Vcam1 (vascular cell adhesion molecule 1) have already been proven to play critical roles in recruiting immune cells for triggering and driving inflammation (Ishida et al., 1994; Daly and Rollins, 2003; Kong et al., 2018; Garcia-Cuesta et al., 2019), strengthening the notion of targeting mononuclear phagocytes signature to ameliorate renal pathology in the context of LN.

GO Analyses of Mononuclear Phagocytes-Specific DEGs Associated With LN Remission Induction

Although we postulated that the overlapping DEGs between GSE27045 and GSE49898 could be ascertained and acknowledged as the potential intrarenal mononuclear phagocytes-specific DEGs highly associated with LN remission induction, statistical analysis found no DEGs (adjusted p-value < 0.05) between nephritis and early remission group in GSE49898. Therefore, 383 intrarenal F4/80^{hi} mononuclear phagocytes-derived DEGs (adjusted p-value < 0.05, and $|\log 2$ fold change (FC)$| > 1$, **Supplementary Table S6**) between nephritis and remission NZB/W mice were directly recognized and deemed as mononuclear phagocytes-specific and LN remission induction-related genes for the following functional enrichment analysis.

GO enrichment analysis found that the top enriched GO term of these DEGs was negative regulation of immune system process, which indicated that immunosuppressive therapy initiated the functional signature in mononuclear phagocytes to counteract autoimmune response with LN kidney (**Figure 7A**). This result together with the GO enrichment analysis of overlapping DEGs between GSE27045 and GSE32583 suggested that mononuclear phagocytes-mediated immune response and inflammatory reaction were eased upon immunosuppressive therapy in NZB/W mice. This reversal possibly played the

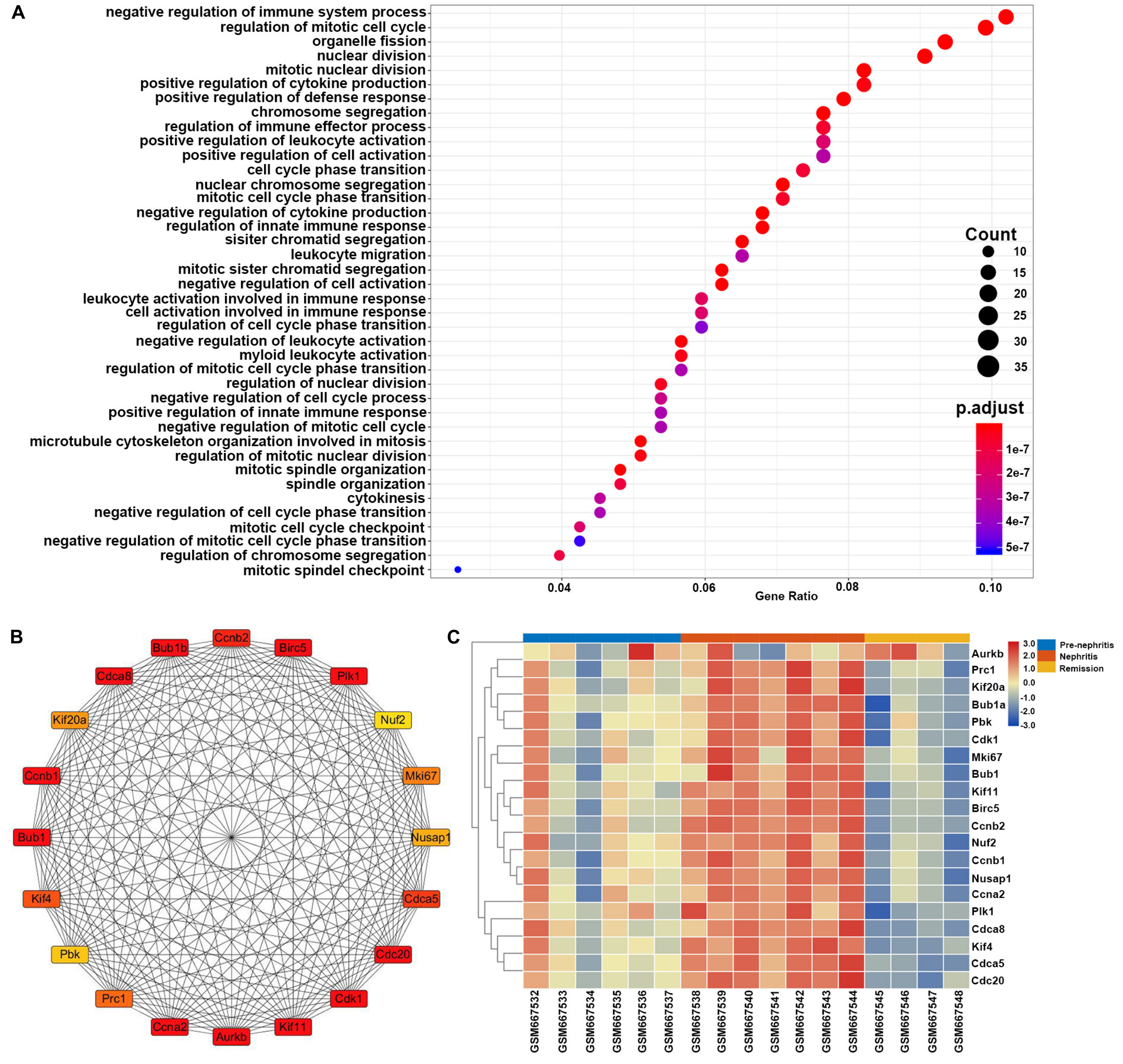

FIGURE 7 | Gene set enrichment analysis of the LN remission-related DEGs in NZB/W mice from GSE27045. The y-axis labels represent clustered GO terms, and the Gene Ratio represents the ratio of the number of genes enriched in one GO term to the number of DEGs. **(A)** GO analysis of LN remission-associated DEGs in GSE27045. **(B)** Network analysis of identified hub genes from LN remission-associated DEGs in GSE27045. **(C)** Heatmap of identified hub genes from LN remission-associated DEGs in GSE27045.

dominant role in achieving remission induction, in agreement with the previous consensus that successfully induced remission was linked with reversal of renal autoimmunity and inflammation signature. Furthermore, heatmap of 20 hub genes extracted from DEGs between nephritis and remission mice in GSE27045 (**Figure 7B** and **Supplementary Table S7**) demonstrated that almost all these hub genes were markedly elevated in the nephritic kidney, but significantly reversed or downregulated upon remission induction (**Figure 7C**). This result highlighted the potential role of these critical genes in impeding remission induction; despite how these hub genes interfere the response to immunosuppressive regimens in the context of LN remains elusive. Intriguingly, in line with our findings, some previous lines of evidence have already shown that some of these hub genes like Cdk1 (cyclin-dependent kinase 1) (Wu et al., 2016) and Mki67 (Rahbar et al., 2018) were tightly involved in the pathogenesis of SLE or LN.

Validation of LN Onset-Related Hub Genes in Human Kidney Biopsies From LN Patients

To determine the relevance to human disease, LN onset-related hub genes identified from overlapping DEGs between GSE27045 and GSE23852 were further validated in kidney biopsies from LN subjects in GSE32591. Results demonstrated that these hub genes were apparently upregulated in glomeruli instead of tubulointerstitial compartments, indicating the functional difference of mononuclear phagocytes in these two intrarenal microenvironments. Moreover, of the total 20 mouse LN onset-related hub genes (corresponding to 18 human genes), most demonstrated a striking level of concordance as that in LN-prone NZB/W mice, as evidenced by notably higher expression level in glomeruli from LN patients compared to healthy living donors. The top three dramatically upregulated hub genes were ITGB2 (integrin subunit beta 2), LYN (LYN proto-oncogene, Src family tyrosine kinase), and CXCR4 (C-X-C motif chemokine receptor 4) (**Figure 8** and **Supplementary Table S8**). It is worth noting that the hub gene CXCR4 was previously found abundant in kidney biopsies from SLE patients, and CXCR4 antagonist administration could improve disease severity and nephritis in murine lupus models (Chong and Mohan, 2009). Therefore, this overlap of a substantial subset of molecular markers with those in human LN kidneys highlighted mononuclear phagocytes signature as a cross-species shared feature (Olaru et al., 2018).

DISCUSSION

High frequency of being resistant to immunosuppressive therapy contributes to adverse renal outcomes in LN (Ginzler et al., 2005). Only 50–70% of LN patients achieved remission by the current therapeutic regimens, and progression to end-stage renal disease still occurs to 10–20% patients over 5–10 years (Tektonidou et al., 2016). Recent advances have pinpointed the involvement of a diverse range of immune cells in LN progression, despite the fact that the underlying pathophysiology is not fully understood. In the current study, through the in-depth bioinformatic analyses of LN transcriptomics data, we implemented a comprehensive deconstruction of intrarenal immune cell landscape in LN-prone NZB/W mice at different disease stages. In consequence, our findings highlighted that amplified mononuclear phagocytes abundance and signature were attributable to not only disease onset but also the failure of early remission induction. Besides, functional enrichment analyses delineated the functional profile of mononuclear phagocytes in driving nephritis onset and hampering the early remission induction. Collectively, our research revealed the characteristic of mononuclear phagocytes abundance and signature to segregate pre-nephritis mice from nephritis mice, as well as early remission mice from late remission mice. These findings suggest the promise of utilizing mononuclear phagocytes as prognostic and predictive biomarkers, as well as potential therapeutic targets for LN administration.

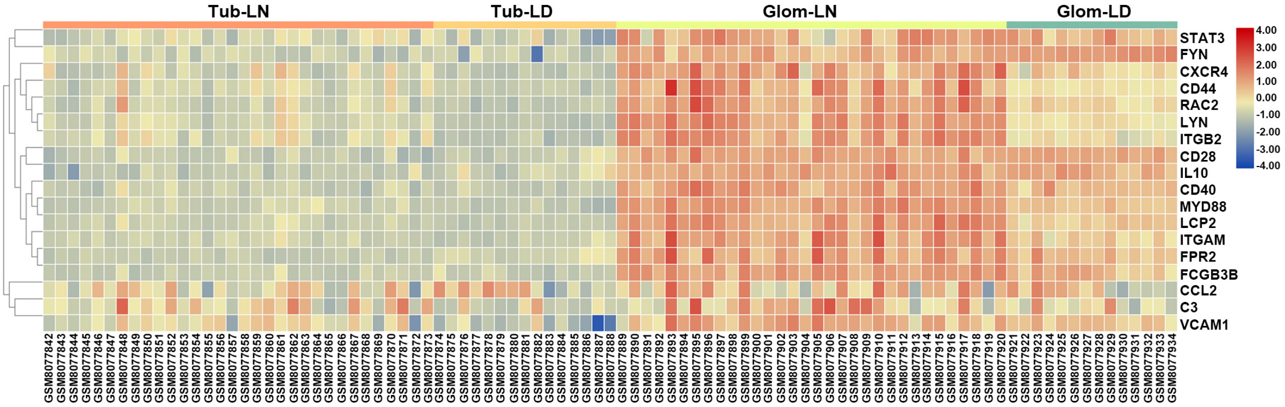

FIGURE 8 | Validation of LN onset-related hub genes from overlapping DEGs between GSE27045 and GSE32583 in the GSE32591 dataset. A heatmap of LN onset-related hub genes identified from the overlapping DEGs between GSE27045 and GSE32583 in the GSE32591 genomic dataset which included kidney biopsies from LN patients. Tub, tubulointerstitial; LN, lupus nephritis; LD, healthy living donor; Glom, glomeruli.

Abnormalities in mononuclear phagocytes phenotype, function, and activation are increasingly being associated with the pathogenesis of autoimmune diseases including SLE (Zhang et al., 2016; Burbano et al., 2018) and rheumatoid arthritis (Ma et al., 2019), despite that the underlying regulatory mechanism of mononuclear phagocytes in the context of LN has not been fully elucidated. Although the strong correlation of mononuclear phagocytes with disease activity or organ damage in LN has been recognized for over a decade, only recently has the research been directed toward the understanding of the cellular and molecular mechanism. For example, LN subjects with severer forms (Class III and Class IV) or LN-prone murine models with severer histological impairment displayed more significant intrarenal monocytes infiltration (Bergtold et al., 2006; Yoshimoto et al., 2007; Menke et al., 2011; Bignon et al., 2014; Barrera Garcia et al., 2016); however, only lately have several lines of evidence shown that three different LN-prone murine models and LN patients were all characterized by the glomeruli-specific accumulation of monocytes with a unique capacity to trigger early immune complex-induced inflammation (Olaru et al., 2018; Kuriakose et al., 2019). Meanwhile, activated renal macrophage has been acknowledged as the hallmark of LN onset and failed remission induction in both LN-prone murine models and human LN subjects (Schiffer et al., 2008; Menke et al., 2009; Triantafyllopoulou et al., 2010; Olmes et al., 2016; Kim et al., 2020), while blockade of macrophage infiltration ameliorated renal inflammation and proteinuria in LN-prone murine models (Kishimoto et al., 2018; Luan et al., 2019).

The failure of most clinical trials of rationally designed therapies in both SLE and LN pleads for an imperative need to dissect the potential mechanisms that impel LN (Arazi et al., 2019). Interpreting the relevance of mononuclear phagocytes in LN and the parallel mechanisms could drive the future identification of more potent therapeutic strategies. Over the last decade, newly emerging technologies like omics-based techniques (e.g., genomics, transcriptomics, and proteomics) offer a promising path toward this goal by expanding our understanding of the molecular basis of LN. Furthermore, multiple computational algorithms like CIBERSORT enable the direct enumeration of immune cell subsets linked with LN kidney conditions. Consistent with previous preclinical and clinical evidence, our current study conducted by comprehensive bioinformatic analyses for the first time depicted that mononuclear phagocytes abundance and signature could be a robust biological marker of LN progression and predictor of failed early remission induction. More importantly, functional enrichment analyses further strengthened the essential role of mononuclear phagocytes in triggering adaptive immune response and inflammation within the kidney upon LN onset, which was however substantially quenched after initiation of immunosuppressive therapy.

Besides, correlation analyses together with GO functional analyses indicated that mononuclear phagocytes signature and other immune cell signals were interwound. This crosstalk orchestrated the adaptive immunity like the differentiation of monocyte to macrophage with increased capacity to drive B and T cell response. For example, findings from Pearson correlation analyses verified the close correlation between the fractions of intrarenal mononuclear phagocytes and the proportions of B cells and T cells. This result was consistent with GO functional enrichment analyses results showing that LN onset-related DEGs in intrarenal mononuclear phagocytes were intensively enriched in lymphocyte differentiation, proliferation, and activation. Last but not least, the majority of the LN onset-associated hub genes identified in NZB/W mice were validated in human LN kidney genomic profile. These validating hub genes demonstrated a similar elevating trend in the glomerular compartment of LN patients, indicating the shared common and unique features between LN-prone murine model and human LN.

CONCLUSION

In the current study, multiple computational approaches were performed to deconvolve the bulk transcriptome data from whole kidney tissue into immune cell type-specific fractions. These results delineated the intrarenal immune cell landscape and estimated the percentages alterations associated with LN onset and remission induction in NZB/W mice. Specifically, our findings identify the significantly amplified mononuclear phagocytes abundance and signature as the source of biological markers that forecast LN onset and retarded remission induction. These discoveries may be extremely pivotal for clinical trial designs and management of novel immunosuppressive therapies in patients with different remission period by shedding light on the suitability of combining mononuclear phagocytes-targeted adjuvant regimen against LN.

DATA AVAILABILITY STATEMENT

The LN-related microarray datasets GSE32583, GSE49898, GSE 27045, and GSE32591 were downloaded from the GEO database (https://www-ncbi-nlm-nih-gov.eproxy.lib.hku.hk/gds).

AUTHOR CONTRIBUTIONS

BL and WC designed this study. BL, YT, and XN retrieved and analyzed the data. BL, YT, and XN drafted the manuscript. WC edited and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was supported by grants from the Chinese National Key Technology R&D Program, Ministry of Science and Technology (Nos. 2017YFC0907601, 2017YFC0907602, and 2017YFC0907603); National Natural Science Foundation of China (No. 81970599); Guangzhou Municipal Program of Science and Technology (No. 201804020049); and Key Laboratory of National Health Commission, and Key Laboratory of Nephrology, Guangdong Province, China (Nos. 2002B60118 and 2017B030314019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.577040/full#supplementary-material

Supplementary Table 1 | Characteristic of all the datasets used in this study.

Supplementary Table 2 | CIBERSORT fraction of infiltrating immune cells in GSE32583.

Supplementary Table 3 | CIBERSORT fraction of infiltrating immune cells in GSE49898.

Supplementary Table 4 | LN onset-related DEGs overlapped between GSE27045 and GSE32583.

Supplementary Table 5 | LN onset-related hub genes identified from DEGs overlapped between GSE27045 and GSE32583.

Supplementary Table 6 | LN remission-related DEGs in GSE27045.

Supplementary Table 7 | LN remission-related hub genes in GSE27045.

Supplementary Table 8 | Validation of LN onset-related hub genes in GSE32591.

REFERENCES

Arazi, A., Rao, D. A., Berthier, C. C., Davidson, A., Liu, Y., Hoover, P. J., et al. (2019). The immune cell landscape in kidneys of patients with lupus nephritis. *Nat. Immunol.* 20, 902–914. doi: 10.1038/s41590-019-0398-x

Barrera Garcia, A., Gomez-Puerta, J. A., Arias, L. F., Burbano, C., Restrepo, M., Vanegas, A. L., et al. (2016). Infiltrating CD16(+) are associated with a reduction in peripheral CD14(+)CD16(++) monocytes and severe forms of lupus nephritis. *Autoimmune Dis.* 2016:9324315. doi: 10.1155/2016/9324315

Bergtold, A., Gavhane, A., D'Agati, V., Madaio, M., and Clynes, R. (2006). FcR-bearing myeloid cells are responsible for triggering murine lupus nephritis. *J. Immunol.* 177, 7287–7295. doi: 10.4049/jimmunol.177.10.7287

Berthier, C. C., Bethunaickan, R., Gonzalez-Rivera, T., Nair, V., Ramanujam, M., Zhang, W., et al. (2012). Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. *J. Immunol.* 189, 988–1001. doi: 10.4049/jimmunol.1103031

Bethunaickan, R., Berthier, C. C., Ramanujam, M., Sahu, R., Zhang, W., Sun, Y., et al. (2011). A unique hybrid renal mononuclear phagocyte activation phenotype in murine systemic lupus erythematosus nephritis. *J. Immunol.* 186, 4994–5003. doi: 10.4049/jimmunol.1003010

Bethunaickan, R., Berthier, C. C., Zhang, W., Eksi, R., Li, H. D., Guan, Y., et al. (2014). Identification of stage-specific genes associated with lupus nephritis and response to remission induction in (NZB x NZW)F1 and NZM2410 mice. *Arthritis Rheumatol.* 66, 2246–2258. doi: 10.1002/art.38679

Bignon, A., Gaudin, F., Hemon, P., Tharinger, H., Mayol, K., Walzer, T., et al. (2014). CCR1 inhibition ameliorates the progression of lupus nephritis in NZB/W mice. *J. Immunol.* 192, 886–896. doi: 10.4049/jimmunol.1300123

Burbano, C., Villar-Vesga, J., Orejuela, J., Munoz, C., Vanegas, A., Vasquez, G., et al. (2018). Potential involvement of platelet-derived microparticles and microparticles forming immune complexes during monocyte activation in patients with systemic lupus erythematosus. *Front. Immunol.* 9:322. doi: 10.3389/fimmu.2018.00322

Cao, Y., Tang, W., and Tang, W. (2019). Immune cell infiltration characteristics and related core genes in lupus nephritis: results from bioinformatic analysis. *BMC Immunol.* 20:37. doi: 10.1186/s12865-019-0316-x

Chan, O., Madaio, M. P., and Shlomchik, M. J. (1997). The roles of B cells in MRL/lpr murine lupus. *Ann. N. Y. Acad. Sci.* 815, 75–87. doi: 10.1111/j.1749-6632.1997.tb52046.x

Chen, Y. E., Korbet, S. M., Katz, R. S., Schwartz, M. M., Lewis, E. J., and Collaborative Study, G. (2008). Value of a complete or partial remission in severe lupus nephritis. *Clin. J. Am. Soc. Nephrol.* 3, 46–53. doi: 10.2215/cjn.03280807

Chong, B. F., and Mohan, C. (2009). Targeting the CXCR4/CXCL12 axis in systemic lupus erythematosus. *Expert Opin. Ther. Targets* 13, 1147–1153. doi: 10.1517/14728220903196761

Daly, C., and Rollins, B. J. (2003). Monocyte chemoattractant protein-1 (CCL2) in inflammatory disease and adaptive immunity: therapeutic opportunities and controversies. *Microcirculation* 10, 247–257. doi: 10.1080/713773639

Duxbury, B., Combescure, C., and Chizzolini, C. (2013). Rituximab in systemic lupus erythematosus: an updated systematic review and meta-analysis. *Lupus* 22, 1489–1503. doi: 10.1177/0961203313509295

Garcia-Cuesta, E. M., Santiago, C. A., Vallejo-Diaz, J., Juarranz, Y., Rodriguez-Frade, J. M., and Mellado, M. (2019). The role of the CXCL12/CXCR4/ACKR3 axis in autoimmune diseases. *Front. Endocrinol. (Lausanne)* 10:585. doi: 10.3389/fendo.2019.00585

Ginzler, E. M., Dooley, M. A., Aranow, C., Kim, M. Y., Buyon, J., Merrill, J. T., et al. (2005). Mycophenolate mofetil or intravenous cyclophosphamide for lupus nephritis. *N. Engl. J. Med.* 353, 2219–2228. doi: 10.1056/NEJMoa043731

Gkirtzimanaki, K., Kabrani, E., Nikoleri, D., Polyzos, A., Blanas, A., Sidiropoulos, P., et al. (2018). IFNalpha impairs autophagic degradation of mtDNA promoting autoreactivity of SLE monocytes in a STING-dependent fashion. *Cell Rep.* 25:e925. doi: 10.1016/j.celrep.2018.09.001

Graffelman, J., Galvan Femenia, I., de Cid, R., and Barcelo Vidal, C. (2019). A log-ratio biplot approach for exploring genetic relatedness based on identity by state. *Front. Genet.* 10:341. doi: 10.3389/fgene.2019.00341

Griffin, D. O., and Rothstein, T. L. (2011). A small CD11b(+) human B1 cell subpopulation stimulates T cells and is expanded in lupus. *J. Exp. Med.* 208, 2591–2598. doi: 10.1084/jem.20110978

Hanly, J. G., O'Keeffe, A. G., Su, L., Urowitz, M. B., Romero-Diaz, J., Gordon, C., et al. (2016). The frequency and outcome of lupus nephritis: results from an international inception cohort study. *Rheumatology (Oxford)* 55, 252–262. doi: 10.1093/rheumatology/kev311

Ishida, H., Muchamuel, T., Sakaguchi, S., Andrade, S., Menon, S., and Howard, M. (1994). Continuous administration of anti-interleukin 10 antibodies delays onset of autoimmunity in NZB/W F1 mice. *J. Exp. Med.* 179, 305–310. doi: 10.1084/jem.179.1.305

Kim, J., Jeong, J. H., Jung, J., Jeon, H., Lee, S., Lim, J. S., et al. (2020). Immunological characteristics and possible pathogenic role of urinary CD11c+ macrophages in lupus nephritis. *Rheumatology (Oxford)* 59, 2135–2145. doi: 10.1093/rheumatology/keaa053

Kishimoto, D., Kirino, Y., Tamura, M., Takeno, M., Kunishita, Y., Takase-Minegishi, K., et al. (2018). Dysregulated heme oxygenase-1(low) M2-like macrophages augment lupus nephritis via Bach1 induced by type I interferons. *Arthritis Res. Ther.* 20:64. doi: 10.1186/s13075-018-1568-1

Kong, D. H., Kim, Y. K., Kim, M. R., Jang, J. H., and Lee, S. (2018). Emerging roles of vascular cell adhesion molecule-1 (VCAM-1) in immunological disorders and cancer. *Int. J. Mol. Sci.* 19:1057. doi: 10.3390/ijms19041057

Kuriakose, J., Redecke, V., Guy, C., Zhou, J., Wu, R., Ippagunta, S. K., et al. (2019). Patrolling monocytes promote the pathogenesis of early lupus-like glomerulonephritis. *J. Clin. Invest.* 129, 2251–2265. doi: 10.1172/jci125116

Liossis, S. C., and Staveri, C. (2017). B cell-based treatments in SLE: past experience and current directions. *Curr. Rheumatol. Rep.* 19:78. doi: 10.1007/s11926-017-0707-z

Lisnevskaia, L., Murphy, G., and Isenberg, D. (2014). Systemic lupus erythematosus. *Lancet* 384, 1878–1888. doi: 10.1016/S0140-6736(14)60128-8

Liszewski, M. K., Kolev, M., Le Friec, G., Leung, M., Bertram, P. G., Fara, A. F., et al. (2013). Intracellular complement activation sustains T cell homeostasis and mediates effector differentiation. *Immunity* 39, 1143–1157. doi: 10.1016/j.immuni.2013.10.018

Luan, J., Fu, J., Chen, C., Jiao, C., Kong, W., Zhang, Y., et al. (2019). LNA-anti-miR-150 ameliorated kidney injury of lupus nephritis by inhibiting renal fibrosis and macrophage infiltration. *Arthritis Res. Ther.* 21:276. doi: 10.1186/s13075-019-2044-2

Ma, W. T., Gao, F., Gu, K., and Chen, D. K. (2019). The role of monocytes and macrophages in autoimmune diseases: a comprehensive review. *Front. Immunol.* 10:1140. doi: 10.3389/fimmu.2019.01140

Melander, C., Sallee, M., Trolliet, P., Candon, S., Belenfant, X., Daugas, E., et al. (2009). Rituximab in severe lupus nephritis: early B-cell depletion affects long-term renal outcome. *Clin. J. Am. Soc. Nephrol.* 4, 579–587. doi: 10.2215/cjn.04030808

Menke, J., Iwata, Y., Rabacal, W. A., Basu, R., Stanley, E. R., and Kelley, V. R. (2011). Distinct roles of CSF-1 isoforms in lupus nephritis. *J. Am. Soc. Nephrol.* 22, 1821–1833. doi: 10.1681/asn.2011010038

Menke, J., Rabacal, W. A., Byrne, K. T., Iwata, Y., Schwartz, M. M., Stanley, E. R., et al. (2009). Circulating CSF-1 promotes monocyte and macrophage phenotypes that enhance lupus nephritis. *J. Am. Soc. Nephrol.* 20, 2581–2592. doi: 10.1681/asn.2009050499

Olaru, F., Dobel, T., Lonsdorf, A. S., Oehrl, S., Maas, M., Enk, A. H., et al. (2018). Intracapillary immune complexes recruit and activate slan-expressing CD16+ monocytes in human lupus nephritis. *JCI Insight* 3:e96492. doi: 10.1172/jci.insight.96492

Olmes, G., Buttner-Herold, M., Ferrazzi, F., Distel, L., Amann, K., and Daniel, C. (2016). CD163+ M2c-like macrophages predominate in renal biopsies from patients with lupus nephritis. *Arthritis Res. Ther.* 18:90. doi: 10.1186/s13075-016-0989-y

Quigley, M., Martinez, J., Huang, X., and Yang, Y. (2009). A critical role for direct TLR2-MyD88 signaling in CD8 T-cell clonal expansion and memory formation following vaccinia viral infection. *Blood* 113, 2256–2264. doi: 10.1182/blood-2008-03-148809

Rahbar, M. H., Rahbar, M. R., Mardanpour, N., and Mardanpour, S. (2018). The potential diagnostic utility of coexpression of Ki-67 and P53 in the renal biopsy in pediatric lupus nephritis. *Int. J. Nephrol. Renovasc. Dis.* 11, 343–350. doi: 10.2147/ijnrd.s175481

Reddy, P. S., Legault, H. M., Sypek, J. P., Collins, M. J., Goad, E., Goldman, S. J., et al. (2008). Mapping similarities in mTOR pathway perturbations in mouse lupus nephritis models and human lupus nephritis. *Arthritis Res. Ther.* 10:R127. doi: 10.1186/ar2541

Schiffer, L., Bethunaickan, R., Ramanujam, M., Huang, W., Schiffer, M., Tao, H., et al. (2008). Activated renal macrophages are markers of disease onset and disease remission in lupus nephritis. *J. Immunol.* 180, 1938–1947. doi: 10.4049/jimmunol.180.3.1938

Schiffer, L., Sinha, J., Wang, X., Huang, W., von Gersdorff, G., Schiffer, M., et al. (2003). Short term administration of costimulatory blockade and cyclophosphamide induces remission of systemic lupus erythematosus nephritis in NZB/W F1 mice by a mechanism downstream of renal immune complex deposition. *J. Immunol.* 171, 489–497. doi: 10.4049/jimmunol.171.1.489

Schmid, H., Boucherot, A., Yasuda, Y., Henger, A., Brunner, B., Eichinger, F., et al. (2006). Modular activation of nuclear factor-kappaB transcriptional programs in human diabetic nephropathy. *Diabetes* 55, 2993–3003. doi: 10.2337/db06-0477

Tektonidou, M. G., Dasgupta, A., and Ward, M. M. (2016). Risk of end-stage renal disease in patients with lupus nephritis, 1971-2015: a systematic review and bayesian meta-analysis. *Arthritis Rheumatol.* 68, 1432–1441. doi: 10.1002/art.39594

Triantafyllopoulou, A., Franzke, C. W., Seshan, S. V., Perino, G., Kalliolias, G. D., Ramanujam, M., et al. (2010). Proliferative lesions and metalloproteinase activity in murine lupus nephritis mediated by type I interferons and macrophages. *Proc. Natl. Acad. Sci. U.S.A.* 107, 3012–3017. doi: 10.1073/pnas.0914902107

Waddell, L. A., Lefevre, L., Bush, S. J., Raper, A., Young, R., Lisowski, Z. M., et al. (2018). ADGRE1 (EMR1, F4/80) is a rapidly-evolving gene expressed in mammalian monocyte-macrophages. *Front. Immunol.* 9:2246. doi: 10.3389/fimmu.2018.02246

Wu, L., Qin, Y., Xia, S., Dai, M., Han, X., Wu, Y., et al. (2016). Identification of cyclin-dependent kinase 1 as a novel regulator of type I interferon signaling in systemic lupus erythematosus. *Arthritis Rheumatol.* 68,1222–1232. doi: 10.1002/art.39543

Yoshimoto, S., Nakatani, K., Iwano, M., Asai, O., Samejima, K., Sakan, H., et al. (2007). Elevated levels of fractalkine expression and accumulation of CD16+ monocytes in glomeruli of active lupus nephritis. *Am. J. Kidney Dis.* 50, 47–58. doi: 10.1053/j.ajkd.2007.04.012

Zarnegar, B., He, J. Q., Oganesyan, G., Hoffmann, A., Baltimore, D., and Cheng, G. (2004). Unique CD40-mediated biological program in B cell activation requires both type 1 and type 2 NF-kappaB activation pathways. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8108–8113. doi: 10.1073/pnas.0402629101

Zhang, H., Fu, R., Guo, C., Huang, Y., Wang, H., Wang, S., et al. (2016). Anti-dsDNA antibodies bind to TLR4 and activate NLRP3 inflammasome in lupus monocytes/macrophages. *J Transl Med.* 14:156. doi: 10.1186/s12967-016-0911-z

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: 20 November 2020
doi: 10.3389/fgene.2020.587243

Multiple-Tissue Integrative Transcriptome-Wide Association Studies Discovered New Genes Associated With Amyotrophic Lateral Sclerosis

Lishun Xiao[1]†, Zhongshang Yuan[2]†, Siyi Jin[1], Ting Wang[1], Shuiping Huang[1,3] and Ping Zeng[1,3]*

[1] Department of Epidemiology and Biostatistics, Xuzhou Medical University, Xuzhou, China, [2] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China, [3] Center for Medical Statistics and Data Analysis, School of Public Health, Xuzhou Medical University, Xuzhou, China

Genome-wide association studies (GWAS) have identified multiple causal genes associated with amyotrophic lateral sclerosis (ALS); however, the genetic architecture of ALS remains completely unknown and a large number of causal genes have yet been discovered. To full such gap in part, we implemented an integrative analysis of transcriptome-wide association study (TWAS) for ALS to prioritize causal genes with summary statistics from 80,610 European individuals and employed 13 GTEx brain tissues as reference transcriptome panels. The summary-level TWAS analysis with single brain tissue was first undertaken and then a flexible *p*-value combination strategy, called summary data-based Cauchy Aggregation TWAS (SCAT), was proposed to pool association signals from single-tissue TWAS analysis while protecting against highly positive correlation among tests. Extensive simulations demonstrated SCAT can produce well-calibrated *p*-value for the control of type I error and was often much more powerful to identify association signals across various scenarios compared with single-tissue TWAS analysis. Using SCAT, we replicated three ALS-associated genes (i.e., *ATXN3*, *SCFD1,* and *C9orf72*) identified in previous GWASs and discovered additional five genes (i.e., *SLC9A8*, *FAM66D*, *TRIP11*, *JUP,* and *RP11-529H20.6*) which were not reported before. Furthermore, we discovered the five associations were largely driven by genes themselves and thus might be new genes which were likely related to the risk of ALS. However, further investigations are warranted to verify these results and untangle the pathophysiological function of the genes in developing ALS.

Keywords: transcriptome-wide association study (TWAS), amyotrophic lateral sclerosis (ALS), genome-wide association studies (GWAS), brain tissue, type I error control

OPEN ACCESS

Edited by:
Yang Zhao,
Nanjing Medical University, China

Reviewed by:
Lihong Huang,
Fudan University, China
Yongyue Wei,
Nanjing Medical University, China

****Correspondence:***
Ping Zeng
zpstat@xzhmu.edu.cn

†These authors share first authorship

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *25 July 2020*
Accepted: *26 October 2020*
Published: *20 November 2020*

Citation:
Xiao L, Yuan Z, Jin S, Wang T, Huang S and Zeng P (2020) Multiple-Tissue Integrative Transcriptome-Wide Association Studies Discovered New Genes Associated With Amyotrophic Lateral Sclerosis. Front. Genet. 11:587243. doi: 10.3389/fgene.2020.587243

BACKGROUND

Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, is an adult-onset progressive and fatal neurodegenerative disease (Kiernan et al., 2011). Although its prevalence rate is not high worldwide (Vazquez, 2008; Marin et al., 2017; Mehta et al., 2018), ALS can lead to severe clinical consequence (Chio et al., 2009) and economic burden (Larkindale et al., 2014;

Gladman and Zinman, 2015). One of the greatest challenges with regards to ALS is that few effective therapeutic interventions have been confirmed and nearly no cure is available in clinic (Mehta et al., 2018; Zeng et al., 2019a). In addition, it is evaluated that the ALS cases across the globe will elevate up to ∼400K in the coming 20 years owing to aging of the population (Arthur et al., 2016), which will further aggravate the socioeconomic threat of ALS.

Prior work has revealed that ALS is highly heritable, with the heritability ranging from 0.52 (95%CI 0.43–0.62) for the ordinary population, to 0.37 (95%CI 0.20–0.54) for those without genetic risks according to population-based studies, and to 0.66 (95%CI 0.59–0.74) based on mother-daughter pairings (Ryan et al., 2019) or 0.61 (95%CI 0.38–0.78) in terms of twin studies (Al-Chalabi et al., 2010). Therefore, understanding the genetic etiology of ALS and identifying risk genes are crucial for early prevention and also have the potential to discover effective therapeutic targets. Indeed, in the past decade dozens of genome-wide association studies (GWAS) have identified multiple single nucleotide polymorphisms (SNPs) and genes causally associated with ALS (McMahon et al., 2019) (**Table 1** and **Supplementary Table S1**). However, the genetic architecture of ALS remains largely unknown and the functional influences of those genetic variants are also not completely clear. For example, the SNP-based heritability estimated in GWAS is only 21%, which is much smaller than that reported above (Keller et al., 2014), implying a large amount of causal genes have not yet been identified and the effort to find causative genes for ALS should continue.

The importance of gene expression regulation in complex diseases motivates us to apply novel statistical tools prioritizing causal genes of ALS through the integration of expression quantitative trait loci (eQTL) into GWAS (Nica et al., 2010; Nicolae et al., 2010; GTEx Consortium, 2015; Li et al., 2016; Wen et al., 2016; GTEx Consortium, 2017; Mancuso et al., 2019). Transcriptome-wide association study (TWAS) is exactly one of such approaches popular in genomic integrative analysis (Gusev et al., 2016; Hu et al., 2019; Mancuso et al., 2019; Wainberg et al., 2019). Methodologically, TWAS can be viewed as a relatively independent two-stage inference procedure to discover causal genes (**Figure 1**). Briefly, in the first stage weights (i.e., the joint effect sizes) of *cis*-SNPs of a given gene are computed from external tissue-related transcriptome reference datasets; and then the association between the imputed expression and the disease of interest is examined for that gene in the second stage. The original TWAS analysis needs large scale individual-level data sets (Gusev et al., 2016), which limits its applicability due to unavailability of such data sets because of privacy concerns in data sharing among various research groups (Gusev et al., 2016; Pasaniuc and Price, 2016). Fortunately, such limitation is already eliminated with the development of summary-level TWAS (Gusev et al., 2016; Barbeira et al., 2018), for which only pre-estimated weights of QTL and summary statistics of GWAS are necessary.

Moreover, because it has been shown that spurious associations may be generated if integrating gene expression from tissues that are not biologically related to the disease (Wainberg et al., 2019), a strongly recommended strategy in TWAS analysis is that one should calculate weights of *cis*-SNPs with expression measurements from the most relevant tissues in the first stage. For instance, the breast-cancer TWAS analysis employs transcriptome datasets of the breast tissue (Wu et al., 2018) and the prostate-cancer TWAS analysis applies transcriptome datasets of the prostate tissue (Mancuso et al., 2018; Wu et al., 2019). Therefore, it is the natural choice of brain tissues when implementing TWAS for ALS. There are 13 GTEx brain tissues that can be employed as reference transcriptome panels (GTEx Consortium, 2015, 2017) (**Table 2**). The rich transcriptome datasets offer an unprecedented opportunity to comprehensively integrating QTL information into the GWAS of ALS. In the meantime, they also propose a great statistical challenge for such integration.

Performing ALS TWAS analysis from one brain tissue to another and then adjusting for multiple comparisons is a conventional approach. However, doing this may be underpowered because of the multiple testing burden; and such a manipulation is not optimal as it ignores useful information of shared eQTLs across brain tissues (GTEx Consortium, 2017). Therefore, it is important to integrate associations from all available brain tissues in the TWAS analysis of ALS with a more efficient manner, which would have the potential to improve power and discover newly genes associated with ALS. However, in terms of our literature view there is little existing work on how to aggregate such evidence efficiently when only summary-level eQTL and GWAS marginal statistics are utilizable. It is hence desirable to construct feasible omnibus tests to handle this problem.

The Fisher's method (Fisher, 1934), one commonly used omnibus test, may be the first choice. Unfortunately, the Fisher's method is only valid for independent multiple tests and thus cannot be employed due to highly positive correlation among individual TWAS tests (see simulations below for details). In fact, as we will demonstrate later, the Fisher's method is overinflated and can lead to too many spurious associations when the TWAS test statistics are not independent. Alternatively, one may take the minimum *p*-value as the significance measure (Conneely and Boehnke, 2007). However, due to the same issue of unknown positive dependence, the null distribution of the minimum *p*-value may be extremely complicated and the computation is often time-consuming since numerical permutation/bootstrap is involved (Conneely and Boehnke, 2007; Sun and Lin, 2019).

Therefore, it is of substantial interest to develop omnibus tests that are robust against correlation. To achieve this objective, herein we propose a novel *p*-values integrative strategy called summary data-based Cauchy Aggregation TWAS (SCAT). Compared to previous approaches, SCAT owns an attractive strength that it takes the summary of a set of *p*-values as test statistic and evaluates the significance analytically without the knowledge of correlation structure. Consequently, SCAT is extraordinarily flexible and computationally fast. With extensive simulation studies we demonstrated that SCAT can produce well-calibrated *p*-value for the control of type I error and is often much more powerful compared with single-tissue TWAS analysis. Finally, using SCAT we discovered several new ALS-associated genes that would be missed by existing statistical strategies.

TABLE 1 | Previous association studies for ALS in terms of the GWAS catalog.

Year	Pop	cases/controls (discover + replication)	*m*	References
2007	EUR	276/271	3	Schymick et al., 2007
2007	EUR	461/450 + 876/906	1	Van Es et al., 2007
2007	EUR	221/211 + 737/721	1	Cronin et al., 2007
2008	EUR	737/721 + 1,030/1,195	3	Van Es et al., 2008
2009	EUR	958/932 + 309/404	1	Cronin et al., 2009
2009	EUR	1,821/2,258 + 538/556	14	Landers et al., 2009
2009	EUR	2,323/9,013 + 2,532/5,940	3	van Es et al., 2009
2010	EUR	405/497	4	Laaksovirta et al., 2010
2010	EUR	4,857/8,987	0	Shatunov et al., 2010
2010	EUR	639/6,257 + 183/961	2	Kwee et al., 2012
2013	EUR	4,243/5,112	19	The Alsgen Consortium, 2013
2013	EUR	6,100/7,125 + 2,074/2,556	3	Fogh et al., 2013
2014	EUR	4,377 + 435/14,431 + 4,056/3,958	10	Diekstra et al., 2014
2015	EUR	25/1,179	1	McLaughlin et al., 2015
2016	EUR	12,577/23,475 + 2,579/2,767	4	van Rheenen et al., 2016
2018	EUR	20,806/59,804 + 4,159/18,650	10	Nicolas et al., 2018
2019	EUR	4,244/3,106	1	Dekker et al., 2019
2013	CHI	506/1,859 + 706/1,777	4	Deng et al., 2013
2013	CHI	4,243 (age of ALS on-set)	15	The Alsgen Consortium, 2013
2013	CHI	250/250	174	Xie et al., 2014
2016	CHI	94/376	1	Chen C.J. et al., 2016
2017	CHI	1,234/2,850 + 576/683	7	Benyamin et al., 2017

Pop denotes which populations the GWAS was performed on, with EUR representing the European population and CHI representing the Chinese Han population; the third column is the sample size of GWAS in the discover stage and in the replication stage if conducted; m denotes the number of unique genes mapped by associated SNPs; these results are overviewed in terms of the GWAS catalog at https://www.ebi.ac.uk/gwas (until 2020-02-02). Of note, some of GWASs had only limited sample sizes, which might influence the validity of the discovered genetic variants and mapped genes in these studies. Therefore, the associations need to interpret in caution.

MATERIALS AND METHODS

GWAS Summary Statistics for ALS

We obtained marginal summary statistics (e.g., Z scores) of ALS from the largest ALS GWAS to date (Nicolas et al., 2018). This study included several previous ALS cohorts such as the work of van Rheenen et al. (2016). For each SNP the logistic regression was first implemented per cohort with individual-level genotypes while incorporating several top principal components, age, and gender as covariates. Then, the inverse-variance weighted fixed-effect meta-analysis was implemented to pool association results across cohorts. Finally, after quality control approximately 8.6 million SNPs on 20,806 cases and 59,804 controls of European ancestry were left for our TWAS analysis.

TWAS Analysis With Single Brain Tissue

To be self-contained, we first introduce TWAS approach for individual-level dataset. Suppose that $\mathbf{G}$ is an $n \times m$ matrix of genotypes of *cis*-SNPs for a gene, n is the sample size for ALS and m is the number of genetic variants and generally changes from gene to gene; $\mathbf{E}$ is an n-vector for *unmeasured* gene expression in the ALS GWAS and $\mathbf{y}$ is an n-vector of binary variable for ALS cases and controls. In addition, assume $\mathbf{g}$ is a $d \times m$ genotype matrix of *cis*-SNPs and $\mathbf{e}$ is a d-vector of gene expression from one of the GTEx brain tissues for the same gene, with d the sample size of the reference panel. The individual-level TWAS analysis can be implemented as

$$\begin{gathered}\textbf{stage 1}: \_\text{weights estimation with genetic prediction models} \\ \_\mathbf{e} = f_{\mathbf{w}}(\text{gw}) \;\Rightarrow\; \hat{\mathbf{w}} \\ \textbf{stage 2}: \_\text{ gene expression imputation and association analysis} \\ \_\text{logit}(\mu) = \hat{E}\theta \text{ with } \hat{E} = G\hat{w}\end{gathered} \tag{1}$$

where $\mathbf{w} = (w_1,\ w_2,\ \ldots,\ w_m)$ is the m-vector of effect sizes for *cis*-SNPs and can be estimated (denoted by $\hat{\mathbf{w}}$) with some genetic prediction model (denoted by f_w) (Zeng and Zhou, 2017); ε is a normal residual and μ is the expectation of $\mathbf{y}$; and θ is the effect size for imputed gene expression. In the TWAS analysis we aim to test for the null hypothesis H_0: $\theta = 0$. It is seen that TWAS bridges the gap between QTL and GWAS in a conceptually simple fashion.

FUSION: A Summary-Level TWAS With Single Tissue

When only summary-level datasets are available (as the case in our analysis of ALS), under the condition of no association between SNP and ALS we have

$$\begin{cases} \hat{\mathbf{z}}^{\text{ALS}} \_\!\sim \mathbf{MVN}(0,\ \mathbf{R}) \\ \hat{\mathbf{z}}^{\text{ALS}}\hat{\mathbf{w}}^{T} \_\!\sim \mathbf{MVN}\left\{0,\ \hat{\mathbf{w}}\mathbf{R}\hat{\mathbf{w}}^{T}\right\} \end{cases} \tag{2}$$

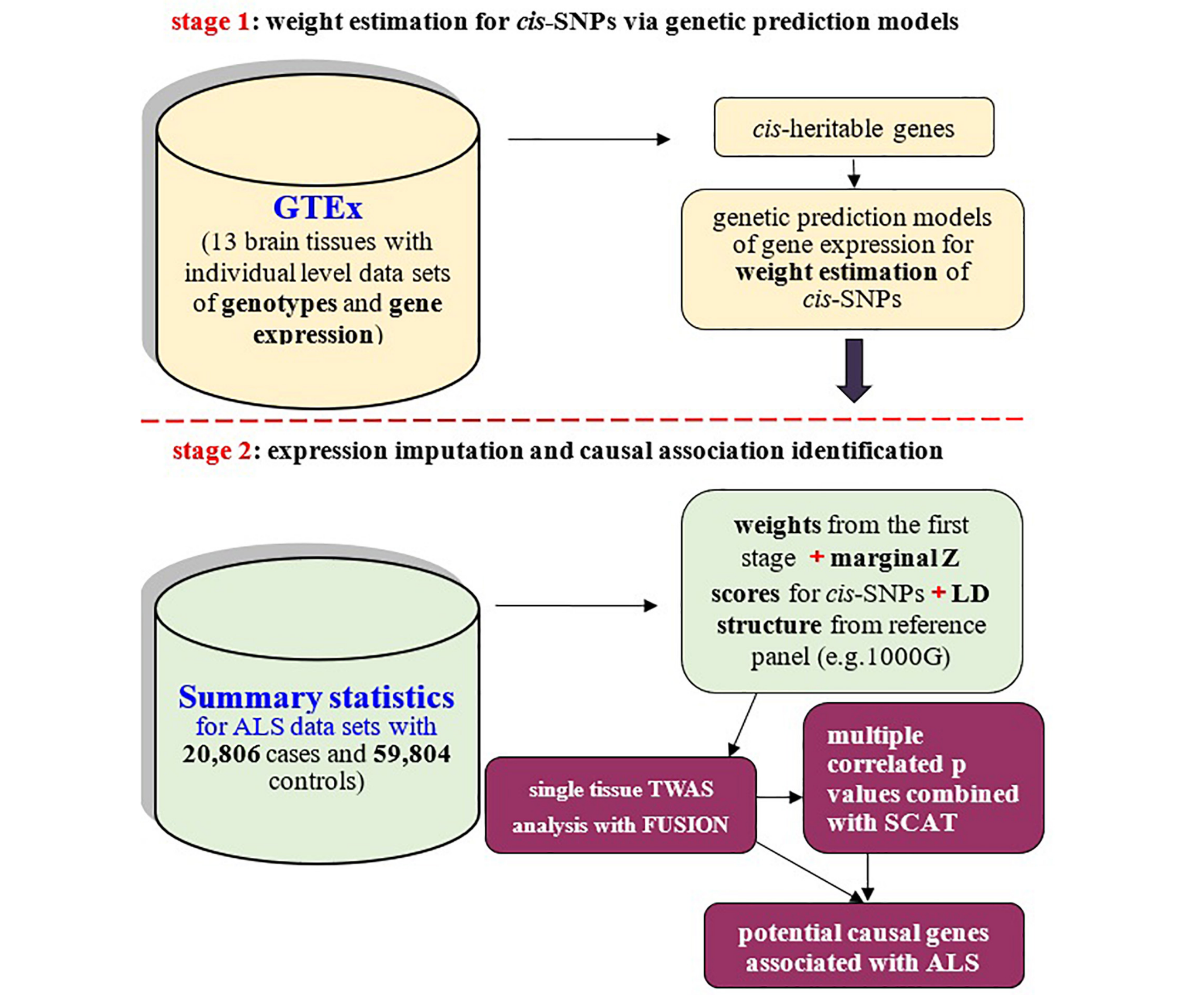
stage 1: weight estimation for cis-SNPs via genetic prediction models
GTEx
(13 brain tissues with individual level data sets of genotypes and gene expression)
cis-heritable genes
genetic prediction models of gene expression for weight estimation of cis-SNPs
stage 2: expression imputation and causal association identification
Summary statistics for ALS data sets with 20,806 cases and 59,804 controls)
weights from the first stage + marginal Z scores for cis-SNPs + LD structure from reference panel (e.g.1000G)
single tissue TWAS analysis with FUSION
multiple correlated p values combined with SCAT
potential causal genes associated with ALS

FIGURE 1 | Schematic framework of TWAS with FUSION and SCAT based on only summary-level datasets and reference panel for linkage disequilibrium (LD) structure of SNPs. TWAS can be viewed to be a relatively independent two-stage inference procedure: the first stage is to estimate weights for *cis*-SNPs with GTEx brain transcriptome reference panel (the **top panel**); the second stage is to examine causal association between genes and ALS with weights obtained from the first stage (the **bottom panel**).

where $\hat{\mathbf{z}}^{\text{ALS}}$ is an m-vector of marginal Z scores of *cis*-SNPs and often generated with single SNP regression (Zeng et al., 2015); **MVN** denotes the multivariate normal distribution, and **R** is the unknown LD correlation matrix among *cis*-SNPs and can be approximately estimated with reference datasets such as 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). With these in hand we define the TWAS statistic as

$$Z_t = \{\hat{\mathbf{z}}^{\text{ALS}}\hat{\mathbf{w}}^T\}\{\hat{\mathbf{w}}\mathbf{R}\hat{\mathbf{w}}^T\}^{-\frac{1}{2}} \tag{3}$$

The p-value of Z_t can be easily obtained since it asymptotically follows a standard normal distribution. The above TWAS analysis is implemented through the FUSION software (Gusev et al., 2016).

Summary-Level TWAS for Multiple-Tissues With Known Correlation Structure

When the correlation structure among gene expressions is known (but it is in fact unknown), a summary-level TWAS approach combining FUSION results of multiple tissues can be designed assuming no association between the gene and ALS across tissues

$$Q\_ = \mathbf{Z}\mathbf{C}^{-1}\mathbf{Z}^T \sim \chi^2_T \tag{4}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_T)$ approximately follows $\mathbf{MVN}(0, \mathbf{C})$ with $\mathbf{C}$ the correlation matrix of gene expressions from T tissues. The above method is also called multiXcan (Barbeira et al., 2019)

TABLE 2 | ALS-associated genes identified by SCAT or FUSION with 13 GTEx brain tissues.

Tissue	*N*	p_0	p_1 (%)	*FAM66D*	*C9orf72*	*TRIP11*	*RP11-529H20.6*	*ATXN3*	*JUP*	*SCFD1*	*SLC9A8*
Amygdala	81	1,799	0 (0.00)								2.85E-1
Anterior cingulate cortex BA24	102	2,653	4 (0.15)				1.20E-1	4.11E-3		6.90E-4	
Caudate basal ganglia	126	3,586	1 (0.03)		2.71E-8			3.06E-1			
Cerebellar hemisphere	113	4,327	6 (0.14)	3.36E-1	3.93E-10	2.01E-1	2.37E-1	3.45E-1		7.25E-4	4.76E-2
Cerebellum	137	5,752	4 (0.07)		4.97E-4	5.86E-3		1.02E-2		1.15E-3	3.73E-1
Cortex	119	3,943	3 (0.08)	7.79E-3	6.41E-3	2.00E-1	1.22E-1				2.00E-1
Frontal cortex BA9	104	3,080	1 (0.03)	5.87E-1	3.84E-16					1.88E-1	
Hippocampus	99	2,245	1 (0.04)	3.66E-1	1.12E-4			8.61E-2			8.61E-2
Hypothalamus	98	2,257	3 (0.13)	4.94E-1			3.65E-1	3.65E-1	1.82E-2	1.55E-4	6.40E-3
Nucleus accumbens basal ganglia	114	3,172	2 (0.06)	5.53E-1	3.32E-24		4.91E-3				
Putamen basal ganglia	98	2,766	1 (0.04)		6.04E-7		2.07E-1				
Spinal cord cervical c-1	76	1,974	2 (0.10)	4.97E-1	1.26E-7						
Substantia nigra	70	1,568	2 (0.13)								
SCAT		11469	8 (0.07)	4.22E-2	1.08E-22	3.49E-2	4.10E-2	3.68E-2	4.22E-2	1.20E-3	4.22E-2

N is the sample size of gene expression in each tissue; p_0 denotes the number of converged genes with heritability estimation, p_1 (%) is the number (or proportion) of associated genes that have FDR < 0.05 in each tissue before adjustment of the 13 GTEx brain tissues.

and provides an omnibus test for the combination of effect in any brain tissue while accounting for correlation. We refer to the test shown in (4) as the *oracle* TWAS. However, due to the lack of transcriptome reference panels (The 1000 Genomes Project Consortium, 2015), **C** is often unknown or cannot be estimated accurately from expression datasets with small sample sizes (Gusev et al., 2016; GTEx Consortium, 2017).

Combination of TWAS via the Aggregated Cauchy Association Test

We here introduce how SCAT can be adopted in our ALS TWAS analysis. First, we separately implement FUSION for each brain tissue and yield Z_t and p_t (t = 1, 2, . . ., T; with T = 13 here); as expected, these p_ts (or Z_ts) are highly correlated (see also below) (Brown, 1975; Kost and McDermott, 2002; Poole et al., 2016; Heard and Rubin-Delanchy, 2018). As a result, as mentioned before the Fisher's method, which assumes independent tests, is not appropriate. We instead apply SCAT which allows us to aggregate multiple potentially dependent p-values obtained from multiple FUSION analyses into a single well-calibrated p-value that can maintain the type I error correctly. The pooled p-value of SCAT follows a Cauchy distribution regardless whether p-values are correlated or not (Liu et al., 2019; Liu and Xie, 2019). Briefly, with SCAT we have

$$\begin{aligned} T_{SCAT} \_= & \sum_{t=1}^{T} \varpi_t \tan\left\{\left(\tfrac{1}{2} - p_t\right)\pi\right\} \\ p_{T_{SCAT}} \_= & \tfrac{1}{2} - \arctan\left\{T_{SCAT}/\left(\sum_{t=1}^{T} \varpi_t\right)\right\}/\pi \end{aligned} \tag{5}$$

where ϖ_t denotes the non-negative weight for each p_t with $\sum_{t=1}^{T} \varpi_t = 1$, and assume that ϖ_t is independent of p_t. When no prior information is available, equal weights are utilized. Because SCAT only takes a group of p-values as input and no any dependence structure is required, its implementation is thus rather straightforward and fast.

Numerical Simulations

We implement simulation studies to assess the performance of SCAT and compare it with the Fisher's method. As described before because both the two methods used only p-values as input; we thus start our simulations by generating a series of independent or non-independent p-values. This is also the simulation framework used in previous work (Liu and Xie, 2019). Specifically, we first obtained the correlation matrix of **Z** values of FUSION (i.e., the **C** matrix; shown in **Supplementary Figure S1**) and generated a 13-dimentional multivariate random variable which followed **MVN**(μ, **C**). Then, we yielded the p-value for each marginal random variable by assuming it followed a standard normal distribution. Finally, we combined these p-values with SCAT or the Fisher's method.

We set μ = 0 when evaluating the type I error control, but randomly sampled μ from an independent normal distribution with mean zero and variance 2.5 when assessing the statistical power. A total of 10^6 or 10^3 replications were generated for type I error control and power evaluation respectively. Furthermore, to match the application in real-life datasets — not all genes were identified to be *cis*-heritable across all brain tissues with the current sample sizes of transcriptome datasets (see **Supplementary Figure S2** for more information) — in each replication of the power assessment we randomly selected at least five but at most eleven tissues to be missing. Doing this was equivalent to generating missing values in each group of marginal p-values.

In the present analysis genes with false discover rate (FDR) (Benjamini and Hochberg, 1995) less than 0.05 were defined to be associated genes. All analyses were carried out with the R software (version 3.6.2); and the codes to reproduce simulations as well as the FUSION results of ALS can be found at https://github.com/biostatpzeng. In addition, since we only employed summary-level genetic datasets that can be publicly available; therefore, additional ethical review was not needed for our study.

RESULTS

Type I Error Control and Power Evaluation

It is observed that both the Fisher's method and SCAT can correctly control the type I error if the *p*-values are independent (**Figures 2A,B**). However, in the presence of positive dependence among *p*-values, the Fisher's method fails to maintain the type I error control and is rather liberal (**Figures 2C,D**). In contrast, SCAT is robust to the positive correlation structure and still displays a desirable behavior on the control of type I error (**Figures 2C,D**). Because of the failure in the type I error control, in the following we no longer consider the Fisher's method.

The estimated statistical power is shown in **Figures 2E,F**. Here, several pronounced observations need to emphasize. **First**, SCAT substantially outperforms any individual one-tissue FUSION in our simulation settings (**Figure 2E** vs. **Figure 2F**). **Second**, as anticipated, ignoring correlation among *p*-values can indeed lead to power reduction. For example, the oracle TWAS (denoted by oracle in **Figure 2F**), which considers the true correlation among the test statistics, has an approximately 10.1% higher power compared with SCAT (denoted by SCAT13 in **Figure 2F**), and the advantage of the oracle TWAS would be more evident if less FUSION analyses are combined by SCAT (e.g., oracle vs. SCAT4 or oracle vs. SCAT8 in **Figure 2F**). However, as aforementioned, the oracle TWAS cannot be applicable due to unavailability of correlation structure in practice, while SCAT is a universal combination approach without such limitation.

Third, SCAT that combines FUSION with a larger set of tissues is often much more powerful than that contains a smaller set of tissues (e.g., SCAT13 vs. SCAT8 or SCAT4; here the number attached represents the number of tissues used in the SCAT analysis, with a greater number indicating more tissues included); in the extreme case where only one tissue in each group (i.e., SCAT1), SCAT reduces to FUSION and exhibits the similar behavior to FUSION. Note that, this simulation is also equivalent to the case where missing *p*-values emerge. Nevertheless, SCAT is still better than any FUSION analysis with one tissue as long as more than two significant tissues are contained. **Fourth**, however, it is not necessarily the case that SCAT can always improve the power. For example, we find SCAT would encounter a loss of power if some of the combined individual FUSION analyses are non-significant (**Supplementary Figure S3**). **Fifth**, it is shown that SCAT would loss the power as the increase in the correlation under various correlation structures (**Supplementary Figure S4**). For instance, SCAT has a power of 0.241, 0.317, 0.427, or 0.572 when the correlation is 0.9, 0.6, 0.3 or 0 in the exchangeable structure (**Supplementary Figure S4A**). In addition, as can be expected, different correlation structures among the test statistics have various influences on the power of SCAT (**Supplementary Figures S4A–C**).

Associated Genes With ALS Discovered in Previous GWASs

In terms of the GWAS Catalog[1], most of the ALS GWASs (17 out of 22) were performed on European individuals (**Table 1**). Totally, there are 313 SNP association pairs discovered across all chromosomes, especially in chromosomes 1 (i.e., 19 SNPs), 2 (i.e., 19 SNPs), and 9 (i.e., 21 SNPs) (**Figures 3A,B**). Those genetic variants are mapped to 253 unique genetic regions, among which 25 are located within *intergenic* (**Figure 3C**). In particular, *C9orf72* — a famous risk gene of ALS (Renton, 2011; Byrne, 2012; Garcia-Redondo, 2013; Diekstra et al., 2014; Chen Y. et al., 2016) — is the most frequent gene. The remaining genes with high frequency include *UNC13A* and *CPNE4* (**Figure 3C**).

Associated Genes With ALS Discovered by FUSION and SCAT

Now we applied FUSION to ALS using 13 GTEx brain tissues as reference transcriptome datasets and then combined the results with SCAT for the overall significance. The correlation among gene expressions is displayed in **Supplementary Figure S5**. A total of 11,469 unique genes are analyzed but only 361 overlapped genes emerging in all the 13 GTEx brain tissues. It is empirically demonstrated that the *p*-values of FUSION among various GTEx brain tissues exhibit highly positive dependency (**Supplementary Figure S5**), which, together the unavailability of correlation information makes nearly all previous *p*-values combined methods cannot be directly utilized.

For each GTEx brain tissue the number of genes with FDR < 0.05 (before adjustment of the issue of multiple tissues) is shown in **Table 2** and **Supplementary Figure S6**A. The full results of TWAS for ALS are shown in **Figure 4**. It is seen that more genes are discovered in cerebellar hemisphere (i.e., 6 genes), following by anterior cingulate cortex BA24 and cerebellum (e.g., 4 genes for both tissues). Again, we observe that *C9orf72* is discovered to be associated with ALS in almost brain tissues which previously had been kept after screening of heritable genes in FUSION. However, if further considering the issue of multiple testing, many of these genes identified by single-tissue FUSION would be non-significant, leaving only two statistically significant genes (i.e., *SCFD1* and *C9orf72*).

The adjusted associations are displayed in **Table 2** and **Supplementary Figure S6B**. Here, a total of eight genes are found by SCAT (FDR < 0.05), among which three (i.e., *SCFD1* with FDR = 0.001, *ATXN3* with FDR = 0.04 and *C9orf72* with FDR = 1.08E-22) are previously identified (**Supplementary Table S1**), while five (i.e., *SLC9A8* with FDR = 0.04, *FAM66D* with FDR = 0.04, *TRIP11* with FDR = 0.03, *JUP* with FDR = 0.04 and *RP11-529H20.6* with FDR = 0.04) are not. Except for *FAM66D* (antisense) and *RP11-529H20.6* (sense overlapping), all others are protein-coding genes (**Supplementary Table S2**). Furthermore, we find that there are no significant SNPs (with p < 5.00E-8) included within any of these five genes (**Supplementary Figure S7**). Thus, in our analysis *SLC9A8*, *FAM66D*, *TRIP11*, *JUP*,

[1] https://www.ebi.ac.uk/gwas

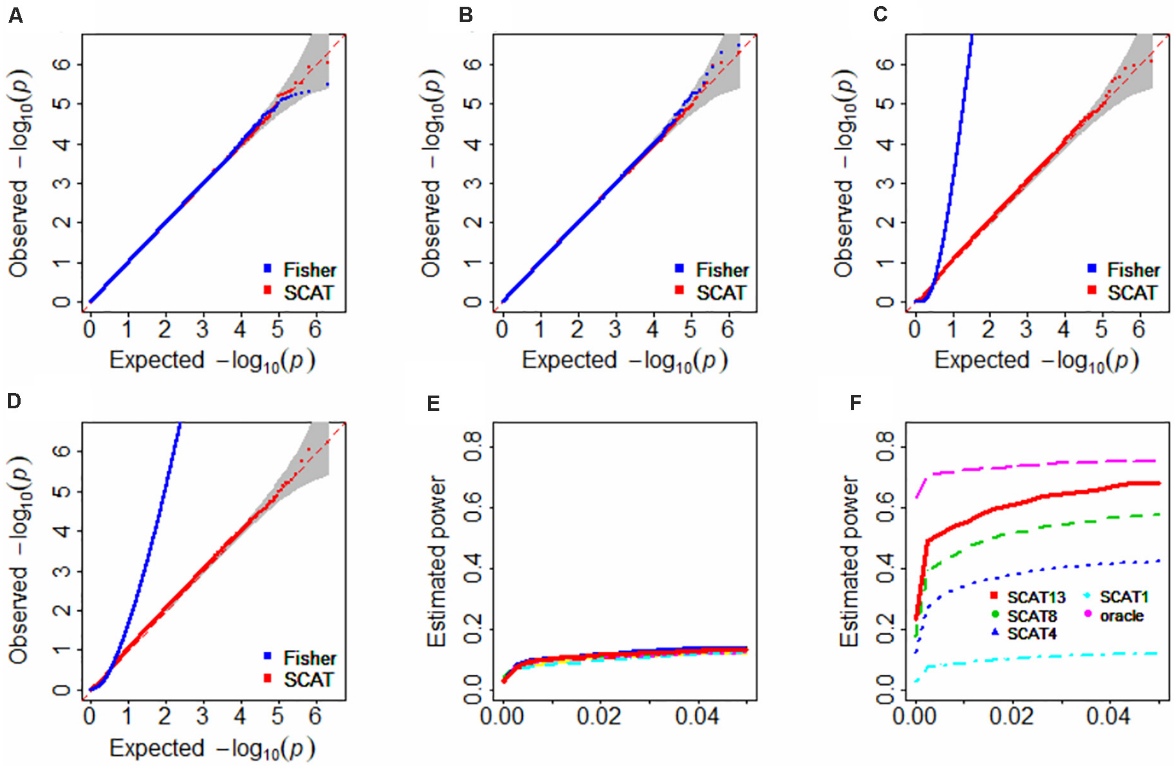

FIGURE 2 | Type I error control **(A–D)** and Estimated statistical power **(E,F)** in the simulation studies. In **(A,B)**, the correlation matrix was independent; in panels **(C,D)**, the correlation matrix was specified with the matrix shown in **Supplementary Figure S2**; in **(E)**, the clustered lines with various colors represent the 13 types of FUSION analysis with one tissue and cannot be clearly separated; in **(F)**, the number attached by SCAT indicates various tissues included; oracle denotes the oracle TWAS approach with the matrix shown in **Supplementary Figure S2**; because the inclusion of all 13 tissues in the oracle TWAS would result in 100% power; thus, here we only considers three tissues that were randomly selected in the oracle TWAS.

and *RP11-529H20.6* can be deemed to be newly genes that are likely associated with ALS.

DISCUSSION

Given the severe health threat and little knowledge of ALS, persistent work should be done to explore genetic and environmental risk factors related to ALS. The present study is one of such efforts with the aim to discover newly causal genes for ALS. To achieve this goal, we conducted the TWAS analysis and integrated association signals from multiple GTEx brain tissues to improve power by borrowing the idea of *p*-values combination. As demonstrated before, the main challenge in our TWAS analysis of ALS emerges in two aspects. First, multiple brain tissues were involved and the statistics of FUSION across tissues exhibited highly positive correlation; second, the dependency structure was unknown in practice because only summary-level statistics results can be available. Those difficulties lead to the failure of the Fisher's method and also hamper the use of other commonly employed methods that can combine dependent *p*-values such as the Brown's method (Brown, 1975), the Kost's method (Kost and McDermott, 2002) and some tests proposed recently (Barnett et al., 2017; Gaynor et al., 2019; Sun et al., 2019; Sun and Lin, 2019), which typically require known covariance among *p*-values.

Our TWAS analysis relies on the newly flexible statistical framework of SCAT for hypothesis testing. Compared with FUSION (i.e., the summary-level TWAS analysis with one tissue each time), SCAT is more efficient as it aggregates individual association signals. With simulation studies we revealed that SCAT produced well-calibrated *p*-value for type I error control and was often much more powerful to identify associated signals across various scenarios compared with FUSION with only single tissue. Using SCAT we replicated three GWAS-discovered genes including *SCFD1* found in van Rheenen et al. (2016) and Nicolas et al. (2018), *ATXN3* identified in Nicolas et al. (2018) and *C9orf72* found in multiple previous GWASs (**Supplementary Table S1**). Among those *C9orf72* is a well-known genetic mutation of ALS previously detected in both European population (The Alsgen Consortium, 2013; Diekstra et al., 2014; McLaughlin et al., 2015; van Rheenen et al., 2016; Nicolas et al., 2018; Dekker et al., 2019) and East Asian population (Benyamin et al., 2017).

More importantly, with SCAT we identified five newly ALS-associated genes that were otherwise missed by existing statistical strategies, including *SLC9A8*, *FAM66D*, *TRIP11*, *JUP*, and *RP11-529H20.6*. Our new findings are also partially

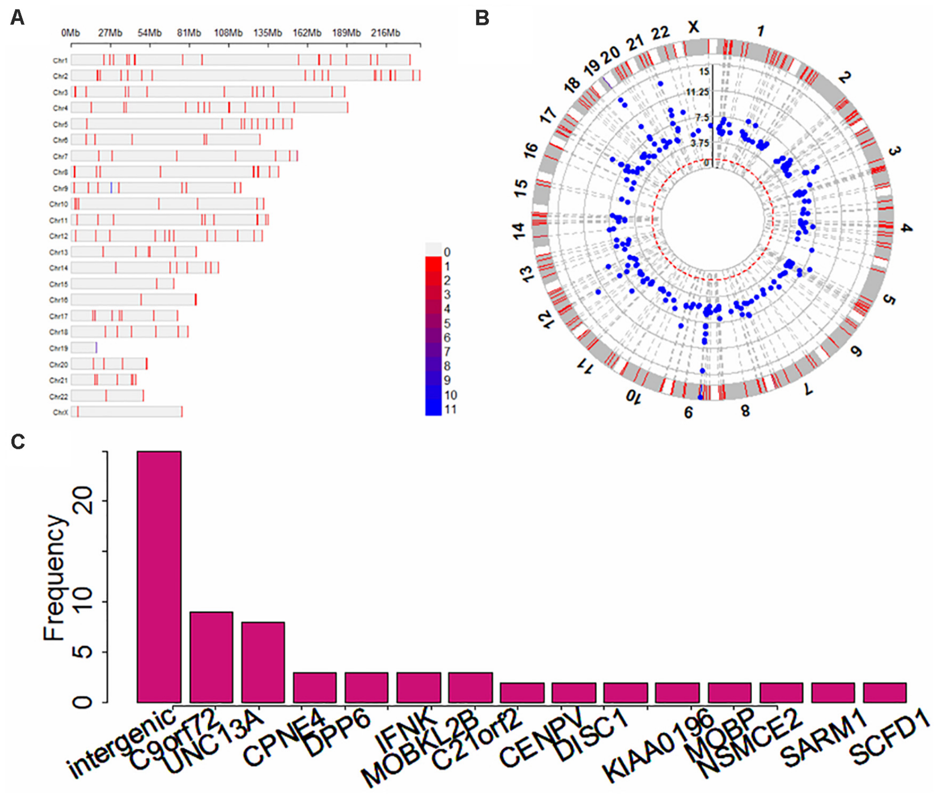

FIGURE 3 | Summary results for ALS-associated SNPs and mapped genes identified in previous GWASs. **(A)** The distribution for associated SNPs across all 22 chromosomes; **(B)** The *p*-values of circle Manhattan plot of associated SNPs for significance; **(C)** The distribution for genes with high frequency.

supported by previous studies. First, in the molecular level one typical pathological hallmark for neurodegeneration of ALS (e.g., tau, amyloid, and beta-protein precursor) is the change in cell cycle control and progression, which can be regulated by *SLC9A8* by inhibiting Na^+/H^+ exchanger activity in epithelia (Hu et al., 1998; Orlowski and Grinstein, 2004). In the population level, *SLC9A8* exhibits widely pleiotropic influence on chronic inflammatory diseases including ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis, and ulcerative colitis (Stuart et al., 2010; Ellinghaus et al., 2016); in addition, *SLC9A8* is also associated with psoriasis (Stuart et al., 2010), gut microbiota (beta diversity) (Wang et al., 2016) and multiple sclerosis (International Multiple Sclerosis Genetics Consortium, 2013).

Second, *TRIP11* can provide instruction for generating a type of protein known as Golgi microtubule-associated protein 210 (GMAP-210) (Infante et al., 1999). This protein is found in the Golgi apparatus, a cell structure in which newly produced proteins are modified so they can be activated. On the other hand, the depletion of Golgi matrix proteins can result in an abnormal, fragmented Golgi morphology, which has been observed in multiple neurodegenerative diseases including ALS (Fujita and Okamoto, 2005), suggesting that the fragmentation of Golgi apparatus may be related to the neuronal degeneration of ALS. In population-based studies, *TRIP11* is identified to be associated with anthropometric traits including height (Gudbjartsson et al., 2008; Lettre et al., 2008; Lango Allen et al., 2010; Wood et al., 2014; He et al., 2015; Tachmazidou et al., 2017; Akiyama et al., 2019) and waist circumference adjusted for body mass index (Shungin et al., 2015; Graff et al., 2017; Justice et al., 2017), which are in turn believed to be relevant to the development of ALS (Desport et al., 1999; Jawaid et al., 2010; Paganoni et al., 2011; Shimizu et al., 2012; O'Reilly et al., 2013; Reich-Slotky et al., 2013; Calvo et al., 2017; Peter et al., 2017; Zeng et al., 2019b).

Third, *JUP* can regulate plakoglobin, a protein plays an important role in signaling within cells as part of the Wingless/Int (Wnt) pathway (Asimaki et al., 2007). The Wnt is a key pathway involved in neural development during embryogenesis (Wang and Wynshaw-Boris, 2004; Harrison-Uy and Pleasure, 2012) and in the maintenance of neuronal homeostasis (Ille and Sommer, 2005; Zhang et al., 2011). In particular, the perturbations of the Wnt pathway have been shown to have a correlation to neurological disorders (De Ferrari and Moon, 2006) as well as neurodegenerative diseases (De Ferrari et al., 2003; Inestrosa and Arenas, 2010).

In addition, in terms of BioSystems *SLC9A8* and *TRIP11* belong to the pathway of GO 0000139 Golgi membrane and *JUP* belongs to the pathway of GO 0000988 transcription factor activity, both of which have a functional role on brain tissues. All those provide evidence that supports the relationship between

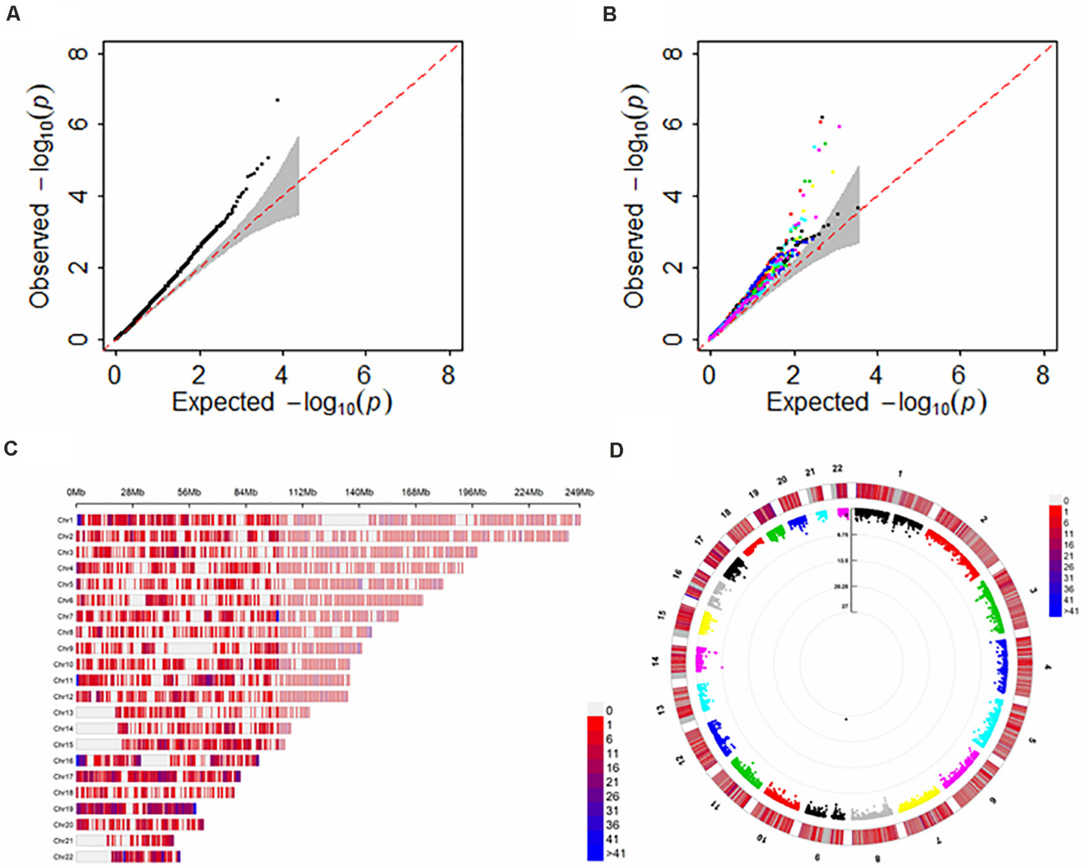

FIGURE 4 | Results of FUSION and SCAT for TWAS analysis of ALS with multiple brain tissues. **(A)** The QQ plot for SCAT; **(B)** The QQ plot for FUSION with each of the GTEx brain tissues as reference dataset; **(C)** The distribution for analyzed genes across all 22 chromosomes; **(D)** The *p*-values of circle Manhattan plot of analyzed genes for significance. Of note, the genomic inflation factor of the p values obtained via SCAT is 1.04, indicating the slight inflation observed in **(A)** might be due to the polygenicity of ALS rather than uncontrolled unknown confounders.

SLC9A8, *JUP*, and *TRIP11* with ALS. It also suggests that those genes may be associated with ALS in a direct, pleiotropic or mediated manner. Those new discoveries are expected to have the potential to advance our understanding of the molecular mechanism with regards to ALS and offer new insight into the etiology of ALS.

Besides discovering new ALS-associated genes, another contribution of the present study exists in the development of SCAT that can integrate a series of correlated association signals efficiently. As illustrated before, SCAT owns the attractive advantage that it takes the summary of a group of *p*-values as test statistic and evaluates the significance analytically without the knowledge of correlation structure (Liu et al., 2019; Liu and Xie, 2019). Therefore, as enthusiastic interest in TWAS continues to grow with more and more genetic and transcriptome data sets collected, especially since large scale individual-level datasets are still unable to obtain for some reasons, we believe that SCAT possesses extensive usefulness to many analogous situations of integrative genomic analyses.

Finally, several limitations of our work need to state. First, among the five new SCAT-identified genes, we do not find reasonable evidence for *FAM66D* and *RP11-529H20.6* in the literature. Second, we cannot replicate those new discoveries in external data sets since such data resources are unavailable for us; we thus simultaneously highlight the need to further validate our findings with additional investigation and experimental follow-up. Third, the used GTEx brain transcriptome reference panels have small samples sizes (ranging from 70 to 137, with the average of 102); as a result, our TWAS analysis may have only limited power. Nevertheless, we note that, in terms of the number of associated genes detected by FUSION with single brain tissue, we believe those new associations are more likely biologically relevant to ALS rather than completely driven by tissues with greater sample size. For example, only 0.07% (i.e., 4) genes were found in brain cerebellum although it has the largest sample size (i.e., 137) and the greatest *cis*-heritable genes (i.e., 5,752); while 0.15% genes were identified in brain anterior cingulate cortex BA24 which has only moderate sample size (i.e., 102) and *cis*-heritable genes (i.e., 2,653). Fourth, because not all genes can be available across all GTEx brain tissues (e.g., **Table 2**), we cannot determine ALS-specific tissues or identify tissue-specific ALS-associated genes, although both are also very interesting and worth of pursuing further (Sonawane et al., 2017; Finucane et al., 2018; Hao et al., 2018). Nevertheless, results displayed in **Table 2** offer some suggestive observations for this issue. For instance, *FAM66D* is likely specially associated with ALS in brain cortex and *RP11-529H20.6* is possibly specifically associated with ALS in brain nucleus accumbens basal ganglia; *ATXN3*, *SCFD1* and *SLC9A8* are relevant to ALS in some brain tissues but not others; while *C9orf72* is associated with ALS across nearly all brain tissues. We note that the step-down inference procedure introduced in Sun et al. (2019) may be a promising approach that can be applied to discriminate which genes drive the observed association signal; but we reserve this problem for investigation in the future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

PZ conceived the idea for the study. PZ, LX, SH, and ZY obtained the data. TW and PZ cleared up the datasets. PZ, LX, SJ, and ZY performed the data analyses. PZ, LX, and ZY interpreted the results of the data analyses. PZ, LX, and ZY drafted the manuscript. All the authors approved the manuscript and provided relevant suggestions.

FUNDING

The research of PZ was supported in part by the Youth Foundation of Humanity and Social Science funded by Ministry of Education of China (18YJC910002), the Natural Science Foundation of Jiangsu Province of China (BK20181472), the China Postdoctoral Science Foundation (2018M630607 and 2019T120465), the QingLan Research Project of Jiangsu Province for Outstanding Young Teachers, the Six-Talent Peaks Project in Jiangsu Province of China (WSN-087), the Training Project for Youth Teams of Science and Technology Innovation at Xuzhou Medical University (TD202008), the Postdoctoral Science Foundation of Xuzhou Medical University, the National Natural Science Foundation of China (81402765), and the Statistical Science Research Project from National Bureau of Statistics of China (2014LY112). The research of ZY was supported in part by the National Natural Science Foundation of China (81673272 and 81872712), the Natural Science Foundation of Shandong Province (ZR2019ZD02), and the Young Scholars Program of Shandong University (2016WLJH23). The research of SH was supported in part by the Social Development Project of Xuzhou City (KC19017). The research of TW was supported in part by the Social Development Project of Xuzhou City (KC20062). The research of LX was supported in part by the China Postdoctoral Science Foundation (2020M671607), the Research Initiation Foundation of Xuzhou Medical University (D2018002), and the National Natural Science Foundation of China (12001470).

ACKNOWLEDGMENTS

We thank the ALS GWAS consortium studies for making the summary data publicly available for us and are grateful of all the investigators and participants contributed to those studies. The data analyses in the present study were implemented with the high-performance computing cluster that was supported by the special central finance project of local universities for Xuzhou Medical University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.587243/full#supplementary-material

REFERENCES

Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* 10:4393. doi: 10.1038/s41467-019-12276-5

Al-Chalabi, A., Fang, F., Hanby, M. F., Leigh, P. N., Shaw, C. E., Ye, W., et al. (2010). An estimate of amyotrophic lateral sclerosis heritability using twin data. *J. Neurol. Neurosurg. Psychiatry* 81, 1324–1326. doi: 10.1136/jnnp.2010.207464

Arthur, K. C., Calvo, A., Price, T. R., Geiger, J. T., Chiò, A., and Traynor, B. J. (2016). Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. *Nat. Commun.* 7:12408. doi: 10.1038/ncomms12408

Asimaki, A., Syrris, P., Wichter, T., Matthias, P., Saffitz, J. E., and McKenna, W. J. (2007). A novel dominant mutation in plakoglobin causes arrhythmogenic right ventricular cardiomyopathy. *Am. J. Hum. Genet.* 81, 964–973. doi: 10.1086/521633

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

Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L., and Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 15:e1007889. doi: 10.1371/journal.pgen.1007889

Barnett, I., Mukherjee, R., and Lin, X. (2017). The generalized higher criticism for testing SNP-set effects in genetic association studies. *J. Am. Statist. Assoc.* 112, 64–76. doi: 10.1080/01621459.2016.1192039

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Benyamin, B., He, J., Zhao, Q., Gratten, J., Garton, F., Leo, P. J., et al. (2017). Cross-ethnic meta-analysis identifies association of the GPX3-TNIP1 locus with amyotrophic lateral sclerosis. *Nat. Commun.* 8:611.

Brown, M. B. (1975). 400: A method for combining non-independent, one-sides tests of significance. *Biometrics* 31, 987–992. doi: 10.2307/2529826

Byrne, S. (2012). Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study. *Lancet Neurol.* 11, 232–240. doi: 10.1016/s1474-4422(12)70014-5

Calvo, A., Moglia, C., Lunetta, C., Marinou, K., Ticozzi, N., Ferrante, G. D., et al. (2017). Factors predicting survival in ALS: a multicenter Italian study. *J. Neurol.* 264, 54–63. doi: 10.1007/s00415-016-8313-y

Chen, C.-J., Chen, C.-M., Pai, T.-W., Chang, H.-T., and Hwang, C.-S. (2016). A genome-wide association study on amyotrophic lateral sclerosis in the Taiwanese Han population. *Biomark. Med.* 10, 597–611. doi: 10.2217/bmm.15.115

Chen, Y., Lin, Z., Chen, X., Cao, B., Wei, Q., Ou, R., et al. (2016). Large C9orf72 repeat expansions are seen in Chinese patients with sporadic amyotrophic lateral sclerosis. *Neurobiol. Aging* 38, .e215–.e217. doi: 10.1016/j.neurobiolaging.2015.11.016

Chio, A., Logroscino, G., Hardiman, O., Swingler, R., Mitchell, D., Beghi, E., et al. (2009). Prognostic factors in ALS: a critical review. *Amyotr. Lateral Sclerosis* 10, 310–323. doi: 10.3109/17482960802566824

Conneely, K. N., and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *Am. J. Hum. Genet.* 81, 1158–1168. doi: 10.1086/522036

Cronin, S., Berger, S., Ding, J., Schymick, J. C., Washecka, N., Hernandez, D. G., et al. (2007). A genome-wide association study of sporadic ALS in a homogenous Irish population. *Hum. Mol. Genet.* 17, 768–774. doi: 10.1093/hmg/ddm361

Cronin, S., Tomik, B., Bradley, D. G., Slowik, A., and Hardiman, O. (2009). Screening for replication of genome-wide SNP associations in sporadic ALS. *Eur. J. Hum. Genet.* 17, 213–218. doi: 10.1038/ejhg.2008.194

De Ferrari, G. V., Chacon, M. A., Barria, M. I., Garrido, J. L., Godoy, J. A., Olivares, G., et al. (2003). Activation of Wnt signaling rescues neurodegeneration and behavioral impairments induced by beta-amyloid fibrils. *Mol. Psychiatry* 8, 195–208. doi: 10.1038/sj.mp.4001208

De Ferrari, G. V., and Moon, R. T. (2006). The ups and downs of Wnt signaling in prevalent neurological disorders. *Oncogene* 25, 7545–7553. doi: 10.1038/sj.onc.1210064

Dekker, A. M., Diekstra, F. P., Pulit, S. L., Tazelaar, G. H. P., van der Spek, R. A., van Rheenen, W., et al. (2019). Exome array analysis of rare and low frequency variants in amyotrophic lateral sclerosis. *Sci. Rep.* 9:5931. doi: 10.1038/s41598-019-42091-3

Deng, M., Wei, L., Zuo, X., Tian, Y., Xie, F., Hu, P., et al. (2013). Genome-wide association analyses in Han Chinese identify two new susceptibility loci for amyotrophic lateral sclerosis. *Nat. Genet.* 45, 697–700. doi: 10.1038/ng.2627

Desport, J., Preux, P., Truong, T., Vallat, J., Sautereau, D., and Couratier, P. (1999). Nutritional status is a prognostic factor for survival in ALS patients. *Neurology* 53, 1059–1059. doi: 10.1212/wnl.53.5.1059

Diekstra, F. P., Deerlin, V. M., Swieten, J. C., Al-Chalabi, A., Ludolph, A. C., Weishaupt, J. H., et al. (2014). C9orf72 and UNC13A are shared risk loci for amyotrophic lateral sclerosis and frontotemporal dementia: A genome-wide meta-analysis. *Ann. Neurol.* 76, 120–133.

Ellinghaus, D., Jostins, L., Spain, S. L., Cortes, A., Bethune, J., Han, B., et al. (2016). Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* 48, 510–518. doi: 10.1038/ng.3528

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* 50, 621–629. doi: 10.1038/s41588-018-0081-4

Fisher, R. A. (1934). *Statistical Methods for Research Workers: Biological Monographs and Manuals*, 5th Edn. Edinburgh: Oliver and Boyd Ltd.

Fogh, I., Ratti, A., Gellera, C., Lin, K., Tiloca, C., Moskvina, V., et al. (2013). A genome-wide association meta-analysis identifies a novel locus at 17q11. 2 associated with sporadic amyotrophic lateral sclerosis. *Hum. Mol. Genet.* 23, 2220–2231.

Fujita, Y., and Okamoto, K. (2005). Golgi apparatus of the motor neurons in patients with amyotrophic lateral sclerosis and in mice models of amyotrophic lateral sclerosis. *Neuropathology* 25, 388–394. doi: 10.1111/j.1440-1789.2005.00616.x

Garcia-Redondo, A. (2013). Analysis of the C9orf72 gene in patients with amyotrophic lateral sclerosis in Spain and different populations worldwide. *Hum. Mutat.* 34, 79–82.

Gaynor, S. M., Sun, R., Lin, X., and Quackenbush, J. (2019). Identification of differentially expressed gene sets using the Generalized Berk-Jones statistic. *Bioinformatics* 35, 4568–4576. doi: 10.1093/bioinformatics/btz277

Gladman, M., and Zinman, L. (2015). The economic impact of amyotrophic lateral sclerosis: a systematic review. *Exp. Rev. Pharmacoecon. Outcomes Res.* 15, 439–450. doi: 10.1586/14737167.2015.1039941

Graff, M., Scott, R. A., Justice, A. E., Young, K. L., Feitosa, M. F., Barata, L., et al. (2017). Genome-wide physical activity interactions in adiposity — A meta-analysis of 200,452 adults. *PLoS Genet.* 13:e1006528. doi: 10.1371/journal.pgen.1006528

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., et al. (2008). Many sequence variants affecting diversity of adult human height. *Nat. Genet.* 40, 609–615. doi: 10.1038/ng.122

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252. doi: 10.1038/ng.3506

Hao, X., Zeng, P., Zhang, S., and Zhou, X. (2018). Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genet.* 14:e1007186. doi: 10.1371/journal.pgen.1007186

Harrison-Uy, S. J., and Pleasure, S. J. (2012). Wnt signaling and forebrain development. *Cold Spring Harb. Perspect. Biol.* 4:a008094. doi: 10.1101/cshperspect.a008094

He, M., Xu, M., Zhang, B., Liang, J., Chen, P., Lee, J.-Y., et al. (2015). Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* 24, 1791–1800. doi: 10.1093/hmg/ddu583

Heard, N. A., and Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. *Biometrika* 105, 239–246. doi: 10.1093/biomet/asx076

Hu, Q., Xia, Y., Corda, S., Zweier, J. L., and Ziegelstein, R. C. (1998). Hydrogen peroxide decreases pHi in human aortic endothelial cells by inhibiting Na+/H+ exchange. *Circ. Res.* 83, 644–651. doi: 10.1161/01.res.83.6.644

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* 51, 568–576. doi: 10.1038/s41588-019-0345-7

Ille, F., and Sommer, L. (2005). Wnt signaling: multiple functions in neural development. *Cell Mol. Life Sci.* 62, 1100–1108. doi: 10.1007/s00018-005-4552-2

Inestrosa, N. C., and Arenas, E. (2010). Emerging roles of Wnts in the adult nervous system. *Nat. Rev. Neurosci.* 11, 77–86. doi: 10.1038/nrn2755

Infante, C., Ramos-Morales, F., Fedriani, C., Bornens, M., and Rios, R. M. (1999). GMAP-210, A cis-Golgi network-associated protein, is a minus end microtubule-binding protein. *J. Cell Biol.* 145, 83–98. doi: 10.1083/jcb.145.1.83

International Multiple Sclerosis Genetics Consortium (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* 45, 1353–1360. doi: 10.1038/ng.2770

Jawaid, A., Murthy, S. B., Wilson, A. M., Qureshi, S. U., Amro, M. J., Wheaton, M., et al. (2010). A decrease in body mass index is associated with faster progression of motor symptoms and shorter survival in ALS. *Amyotr. Lateral Sclerosis* 11, 542–548. doi: 10.3109/17482968.2010.482592

Justice, A. E., Winkler, T. W., Feitosa, M. F., Graff, M., Fisher, V. A., Young, K., et al. (2017). Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat. Commun.* 8:14977. doi: 10.1038/ncomms14977

Keller, M. F., Ferrucci, L., Singleton, A. B., Tienari, P. J., Laaksovirta, H., Restagno, G., et al. (2014). Genome-wide analysis of the heritability of amyotrophic lateral sclerosis. *JAMA Neurol.* 71, 1123–1134. doi: 10.1001/jamaneurol.2014.1184

Kiernan, M. C., Vucic, S., Cheah, B. C., Turner, M. R., Eisen, A., Hardiman, O., et al. (2011). Amyotrophic lateral sclerosis. *Lancet* 377, 942–955. doi: 10.1016/S0140-6736(10)61156-7

Kost, J. T., and McDermott, M. P. (2002). Combining dependent P-values. *Statist. Prob. Lett.* 60, 183–190. doi: 10.1016/S0167-7152(02)00310-3

Kwee, L. C., Liu, Y., Haynes, C., Gibson, J. R., Stone, A., Schichman, S. A., et al. (2012). A high-density genome-wide association screen of sporadic ALS in US veterans. *PLoS One* 7:e32768. doi: 10.1371/journal.pone.0032768

Laaksovirta, H., Peuralinna, T., Schymick, J. C., Scholz, S. W., Lai, S.-L., Myllykangas, L., et al. (2010). Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *Lancet Neurol.* 9, 978–985. doi: 10.1016/s1474-4422(10)70184-8

Landers, J. E., Melki, J., Meininger, V., Glass, J. D., van den Berg, L. H., van Es, M. A., et al. (2009). Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9004–9009.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838.

Larkindale, J., Yang, W., Hogan, P. F., Simon, C. J., Zhang, Y., Jain, A., et al. (2014). Cost of illness for neuromuscular diseases in the United States. *Muscle and Nerve* 49, 431–438. doi: 10.1002/mus.23942

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., et al. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.* 40, 584–591. doi: 10.1038/ng.125

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604. doi: 10.1126/science.aad9417

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* 104, 410–421. doi: 10.1016/j.ajhg.2019.01.002

Liu, Y., and Xie, J. (2019). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Statist. Assoc.* 115, 393-402. doi: 10.1080/01621459.2018.1554485

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., et al. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* 51, 675–682. doi: 10.1038/s41588-019-0367-1

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.* 9:4079. doi: 10.1038/s41467-018-06302-1

Marin, B., Boumédiene, F., Logroscino, G., Couratier, P., Babron, M.-C., Leutenegger, A. L., et al. (2017). Variation in worldwide incidence of amyotrophic lateral sclerosis: a meta-analysis. *Int. J. Epidemiol.* 46, 57–74. doi: 10.1093/ije/dyw061

McLaughlin, R. L., Kenna, K. P., Vajda, A., Bede, P., Elamin, M., Cronin, S., et al. (2015). A second-generation Irish genome-wide association study for amyotrophic lateral sclerosis. *Neurobiol. Aging* 36, 1221.e7–1221.e13. doi: 10.1016/j.neurobiolaging.2014.08.030

McMahon, A., Malangone, C., Suveges, D., Sollis, E., Cunningham, F., Riat, H. S., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucl. Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Mehta, P., Kaye, W., Raymond, J., Punjani, R., Larson, T., Cohen, J., et al. (2018). Prevalence of amyotrophic lateral sclerosis - United States, 2015. *Morb. Mortal. Weekly Rep.* 67, 1285–1289. doi: 10.15585/mmwr.mm6746a1

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., et al. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6:e1000895. doi: 10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi: 10.1371/journal.pgen.1000888

Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., et al. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* 97, 1268-1283. doi: 10.1016/j.neuron.2018.02.027

O'Reilly, ÉJ., Wang, H., Weisskopf, M. G., Fitzgerald, K. C., Falcone, G., McCullough, M. L., et al. (2013). Premorbid body mass index and risk of amyotrophic lateral sclerosis. *Amyotr. Lateral Sclerosis Front. Degen.* 14, 205–211. doi: 10.3109/21678421.2012.735240

Orlowski, J., and Grinstein, S. (2004). Diversity of the mammalian sodium/proton exchanger SLC9 gene family. *Pflugers Arch.* 447, 549–565. doi: 10.1007/s00424-003-1110-3

Paganoni, S., Deng, J., Jaffa, M., Cudkowicz, M. E., and Wills, A.-M. (2011). Body mass index, not dyslipidemia, is an independent predictor of survival in amyotrophic lateral sclerosis. *Muscle Nerve* 44, 20–24. doi: 10.1002/mus.22114

Pasaniuc, B., and Price, A. L. (2016). Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* 18, 117–127. doi: 10.1038/nrg.2016.142

Peter, R. S., Rosenbohm, A., Dupuis, L., Brehme, T., Kassubek, J., Rothenbacher, D., et al. (2017). Life course body mass index and risk and prognosis of amyotrophic lateral sclerosis: results from the ALS registry Swabia. *Eur. J. Epidemiol.* 32, 901–908. doi: 10.1007/s10654-017-0318-z

Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B., and Knijnenburg, T. A. (2016). Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics* 32, i430–i436. doi: 10.1093/bioinformatics/btw438

Reich-Slotky, R., Andrews, J., Cheng, B., Buchsbaum, R., Levy, D., Kaufmann, P., et al. (2013). Body mass index (BMI) as predictor of ALSFRS-R score decline in ALS patients. *Amyotr. Lateral Sclerosis Front. Degen.* 14, 212–216. doi: 10.3109/21678421.2013.770028

Renton, A. E. (2011). A hexanucleotide repeat expansion in C9orf72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 72, 257–268.

Ryan, M., Heverin, M., McLaughlin, R. L., and Hardiman, O. (2019). Lifetime risk and heritability of amyotrophic lateral sclerosis. *JAMA Neurol.* 76, 1367–1374. doi: 10.1001/jamaneurol.2019.2044

Schymick, J. C., Scholz, S. W., Fung, H.-C., Britton, A., Arepalli, S., Gibbs, J. R., et al. (2007). Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.* 6, 322–328. doi: 10.1016/s1474-4422(07)70037-6

Shatunov, A., Mok, K., Newhouse, S., Weale, M. E., Smith, B., Vance, C., et al. (2010). Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *Lancet Neurol.* 9, 986–994.

Shimizu, T., Nagaoka, U., Nakayama, Y., Kawata, A., Kugimoto, C., Kuroiwa, Y., et al. (2012). Reduction rate of body mass index predicts prognosis for survival in amyotrophic lateral sclerosis: a multicenter study in Japan. *Amyotr. Lateral Sclerosis* 13, 363–366. doi: 10.3109/17482968.2012.678366

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Lockes, A. E., Maegi, R., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 518, 187–196. doi: 10.1038/nature14132

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., et al. (2017). Understanding tissue-specific gene regulation. *Cell Rep.* 21, 1077–1088. doi: 10.1016/j.celrep.2017.10.001

Stuart, P. E., Nair, R. P., Ellinghaus, E., Ding, J., Tejasvi, T., and Gudjonsson, J. E. (2010). Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nat. Genet.* 42, 1000–1004. doi: 10.1038/ng.693

Sun, R., Hui, S., Bader, G. D., Lin, X., and Kraft, P. (2019). Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. *PLoS Genet.* 15:e1007530. doi: 10.1371/journal.pgen.1007530

Sun, R., and Lin, X. (2019). Genetic variant set-based tests using the generalized berk–jones statistic with application to a genome-wide association study of breast cancer. *J. Am. Statist. Assoc.* 115, 1079-1091. doi: 10.1080/01621459.2019.1660170

Tachmazidou, I., Suveges, D., Min, J. L., Ritchie, G. R. S., Steinberg, J., Walter, K., et al. (2017). Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *Am. J. Hum. Genet.* 100, 865–884. doi: 10.1016/j.ajhg.2017.04.014

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

The Alsgen Consortium (2013). Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34. 1. *Neurobiol. Aging* 34:e357.

Van Es, M. A., Van Vught, P. W., Blauw, H. M., Franke, L., Saris, C. G., Andersen, P. M., et al. (2007). ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *Lancet Neurol.* 6, 869–877.

Van Es, M. A., Van Vught, P. W., Blauw, H. M., Franke, L., Saris, C. G., Van Den Bosch, L., et al. (2008). Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nature Genetics* 40, 29–31.

van Es, M. A., Veldink, J. H., Saris, C. G. J., Blauw, H. M., van Vught, P. W. J., Birve, A., et al. (2009). Genome-wide association study identifies 19p13.3 *(UNC*13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat. Genet.* 41, 1083–1087.

van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., et al. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* 48, 1043–1048. doi: 10.1038/ng.3622

Vazquez, M. C. (2008). Incidence and prevalence of amyotrophic lateral sclerosis in Uruguay: a population-based study. *Neuroepidemiology* 30, 105–111. doi: 10.1159/000120023

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* 51, 592–599. doi: 10.1038/s41588-019-0385-z

Wang, J., Thingholm, L. B., Skieceviciene, J., Rausch, P., Kummen, M., Hov, J. R., et al. (2016). Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* 48, 1396–1406. doi: 10.1038/ng.3695

Wang, J., and Wynshaw-Boris, A. (2004). The canonical Wnt pathway in early mammalian embryogenesis and stem cell maintenance/differentiation. *Curr. Opin. Genet. Dev.* 14, 533–539. doi: 10.1016/j.gde.2004.07.013

Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient integrative Multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* 98, 1114–1129. doi: 10.1016/j.ajhg.2016.03.029

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186.

Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., et al. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978. doi: 10.1038/s41588-018-0132-x

Wu, L., Wang, J., Cai, Q., Cavazos, T. B., Emami, N. C., Long, J., et al. (2019). Identification of novel susceptibility loci and genes for prostate cancer risk: a transcriptome-wide association study in over 140,000 european descendants. *Cancer Res.* 79, 3192–3204. doi: 10.1158/0008-5472.can-18-3536

Xie, T., Deng, L., Mei, P., Zhou, Y., Wang, B., Zhang, J., et al. (2014). A genome-wide association study combining pathway analysis for typical sporadic amyotrophic lateral sclerosis in Chinese Han populations. *Neurobiol. Aging* 35, 1778.e9–1778.e23.

Zeng, P., Wang, T., Zheng, J., and Zhou, X. (2019a). Causal association of type 2 diabetes with amyotrophic lateral sclerosis: new evidence from Mendelian randomization using GWAS summary statistics. *BMC Medicine* 17:225. doi: 10.1186/s12916-019-1448-9

Zeng, P., Yu, X., and Xu, H. (2019b). Association between premorbid body mass index and amyotrophic lateral sclerosis: causal inference through genetic approaches. *Front. Neurol.* 10:543. doi: 10.3389/fneur.2019.00543

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., et al. (2015). Statistical analysis for genome-wide association study. *J. Biomed. Res.* 29, 285–297. doi: 10.7555/jbr.29.20140007

Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.* 8:456. doi: 10.1038/s41467-017-00470-2

Zhang, L., Yang, X., Yang, S., and Zhang, J. (2011). The Wnt /beta-catenin signaling pathway in the adult neurogenesis. *Eur. J. Neurosci.* 33, 1–8. doi: 10.1111/j.1460-9568.2010.7483.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

ORIGINAL RESEARCH
published: 03 December 2020
doi: 10.3389/fgene.2020.592461

Improved Detection of Potentially Pleiotropic Genes in Coronary Artery Disease and Chronic Kidney Disease Using GWAS Summary Statistics

Haimiao Chen[1]†, Ting Wang[1]†, Jinna Yang[2], Shuiping Huang[1,3]* and Ping Zeng[1,3]*

[1] Department of Epidemiology and Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou, China, [2] Department of Infectious Diseases, People's Hospital of Zhuji, Shaoxing, China, [3] Center for Medical Statistics and Data Analysis, School of Public Health, Xuzhou Medical University, Xuzhou, China

The coexistence of coronary artery disease (CAD) and chronic kidney disease (CKD) implies overlapped genetic foundation. However, the common genetic determination between the two diseases remains largely unknown. Relying on summary statistics publicly available from large scale genome-wide association studies (n = 184,305 for CAD and n = 567,460 for CKD), we observed significant positive genetic correlation between CAD and CKD (r_g = 0.173, p = 0.024) via the linkage disequilibrium score regression. Next, we implemented gene-based association analysis for each disease through MAGMA (Multi-marker Analysis of GenoMic Annotation) and detected 763 and 827 genes associated with CAD or CKD (FDR $<$ 0.05). Among those 72 genes were shared between the two diseases. Furthermore, by integrating the overlapped genetic information between CAD and CKD, we implemented two pleiotropy-informed informatics approaches including cFDR (conditional false discovery rate) and GPA (Genetic analysis incorporating Pleiotropy and Annotation), and identified 169 and 504 shared genes (FDR $<$ 0.05), of which 121 genes were simultaneously discovered by cFDR and GPA. Importantly, we found 11 potentially new pleiotropic genes related to both CAD and CKD (i.e., *ARHGEF19, RSG1, NDST2, CAMK2G, VCL, LRP10, RBM23, USP10, WNT9B, GOSR2*, and *RPRML*). Five of the newly identified pleiotropic genes were further repeated via an additional dataset CAD available from UK Biobank. Our functional enrichment analysis showed that those pleiotropic genes were enriched in diverse relevant pathway processes including quaternary ammonium group transmembrane transporter, dopamine transport. Overall, this study identifies common genetic architectures overlapped between CAD and CKD and will help to advance understanding of the molecular mechanisms underlying the comorbidity of the two diseases.

Keywords: coronary artery disease, chronic kidney disease, pleiotropy-informed integrative analysis, gene-based association analysis, pleiotropic gene, genome-wide association study

OPEN ACCESS

Edited by:
Sheng Yang,
Nanjing Medical University, China

Reviewed by:
Can Yang,
Hong Kong University of Science and Technology, Hong Kong
Le Shu,
University of California, Los Angeles, United States

Correspondence:
Shuiping Huang
sph@xzhmu.edu.cn
Ping Zeng
zpstat@xzhmu.edu.cn

†These authors share first authorship

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *07 August 2020*
Accepted: *17 November 2020*
Published: *03 December 2020*

Citation:
Chen H, Wang T, Yang J, Huang S and Zeng P (2020) Improved Detection of Potentially Pleiotropic Genes in Coronary Artery Disease and Chronic Kidney Disease Using GWAS Summary Statistics. Front. Genet. 11:592461. doi: 10.3389/fgene.2020.592461

INTRODUCTION

Both coronary artery disease (CAD) and chronic kidney disease (CKD) are the leading causes of death and disability worldwide, representing serious global public health threats (Kessler et al., 2013; Inrig et al., 2014; Ene-Iordache et al., 2016; Levin et al., 2017; Musunuru and Kathiresan, 2019). In practice, it is often observed that CKD patients encounter an increased risk of CAD and CAD is in turn a major cause of death for CKD patients (Tonelli et al., 2012). Pathologically, the endothelial dysfunction is closely related to cardiovascular diseases and plays an important role in all stages of atherosclerosis (Ross, 1999). On the other hand, the role of CAD in CKD is also widely studied; for example, the endothelial dysfunction in the development of CKD was also well documented (Moody et al., 2012). As originally proposed by Lindner et al. (1974), CKD patients with an estimated glomerular filtration rate (eGFR) <60 ml/min per 1.73 m^2 have 2∼16 times higher risk of major adverse cardiovascular events (MACE) compared to those with an eGFR > 60 ml/min per 1.73 m^2 (Go et al., 2004). Moreover, for CKD patients not yet requiring renal replacement therapy, the probability of developing MACE is much higher than reaching end-stage renal disease (ESRD) and requiring renal replacement therapy (Foley et al., 2005).

All those empirical observations suggest that there exist a common susceptible mechanism underlying these two complex diseases. As part of efforts to understand their genetic foundation, in the past few years many large scale genome-wide association studies (GWASs) have been implemented for CAD (Nikpay et al., 2015) and CKD (Wuttke et al., 2019). It is found that a lot of genes and single nucleotide polymorphisms (SNPs) exhibit pleiotropic effects and are associated with both the two diseases (Solovieff et al., 2013; **Supplementary Table 1**). This genetic overlap partly contributes to the co-existence of CAD and CKD. The understanding of common genetic determinants has significant implication for identifying important biomarkers and developing novel therapeutic strategies for joint prediction, prevention, and intervention of CAD and CKD.

However, like many other diseases/traits (Manolio et al., 2009; Eichler et al., 2010; Gusev et al., 2013; Girirajan, 2017; Kim et al., 2017; Young, 2019), CAD- or CKD-associated SNPs identified by GWAS only explain a very small fraction of phenotypic variance of CKD (Wuttke et al., 2019) and CAD (Nikpay et al., 2015), implying that a large number of genetic variants with small to modest effect sizes (but still important) have yet been discovered and that more pleiotropic genes would be found if increasing sample sizes (Wang et al., 2005; Altshuler et al., 2008; Tam et al., 2019). However, the increase of sample sizes is generally not feasible since the recruiting and genotyping of additional participants are time consuming and expensive. Therefore, it is a promising way to leverage genetic computational methods that can efficiently analyze information contained in the existing pool of available GWAS summary statistics for identifying loci with pleiotropic effects.

To achieve this aim, many pleiotropy-informed approaches have been proposed (Andreassen et al., 2013; Chung et al., 2014; Zeng et al., 2018). Those previous studies were focused on individual SNP associations and fine-mapping was further needed to find causal genes once newly novel genetic variants were detected (Hormozdiari et al., 2014, 2015; Wen et al., 2015; Kichaev et al., 2016). In addition, those methods cannot effectively handle the correlation among genetic variants due to linkage disequilibrium (LD) (Zeng et al., 2018). As a result, pruning [e.g., using PLINK (Purcell et al., 2007)] has to be employed to keep less dependent SNPs in their analysis, which inevitably leads to the loss of useful information included in correlated SNPs. Compared with the traditional single SNP analysis which only considers only one SNP each time and often suffers from power reduction (Zeng et al., 2015), the gene-based association study is another popular supplementary analysis, which examines the joint significance of a group of SNPs and has the potential to aggregate weak association signals across multiple genetic variants and is thus more powerful (Zeng et al., 2014). Moreover, gene-based associations are easily to interpret because gene is a more meaningfully biological unit compared with individual genetic variant.

Given the potential pleiotropy between CAD and CKD that was widely implied in previous work (Go et al., 2004; Liu et al., 2012; Ene-Iordache et al., 2016), we hypothesize that shared genes identified by different pleiotropy-informed methods should have a higher probability to be candidate pleiotropic genes. To do so, in the present study we first evaluated the overall genetic correlation between CAD and CKD with summary statistics available from large scale GWASs through cross-trait LDSC (linkage disequilibrium score regression) (Bulik-Sullivan B. et al., 2015). We next conducted a gene-based association analysis using MAGMA (Multi-marker Analysis of GenoMic Annotation) (de Leeuw et al., 2015) to integrate association signals from SNP level into gene level. We thus obtained *P*-value for each protein coding gene. Depending on those gene-level *P*-values, we detected pleiotropic genes with two pleiotropy-informed association methods including cFDR (conditional false discovery rate) (Andreassen et al., 2013; Smeland et al., 2020) and GPA (Genetic analysis incorporating Pleiotropy and Annotation) (Chung et al., 2014). We also attempted to validate our results in another CAD dataset available from the UK Biobank (UKB) cohort. The framework of our data analysis is demonstrated in **Figure 1**.

MATERIALS AND METHODS

GWAS Summary Statistics

We obtained summary statistics (e.g., effect allele, effect size, and *P*-values) for CKD from the latest GWAS of the CKDGen consortium (Wuttke et al., 2019). In this study the creatinine value obtained with a Jaffé assay before 2009 was calibrated by multiplying by 0.95, and glomerular filtration rate (GFR) was estimated with the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation for adults (larger than 18 years age) while using the Schwartz formula for individuals less than 18 years old and was winsorized at 15–200 ml min^{-1} per 1.73 m^2. CKD was defined as an eGFR below 60 ml min^{-1} per 1.73 m^2. After stringent quality control, a total

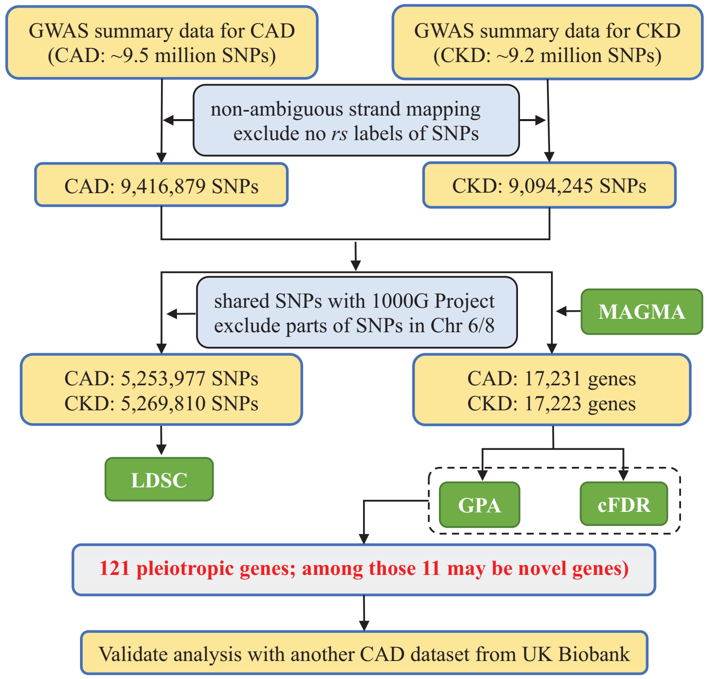
GWAS summary data for CAD
(CAD: ~9.5 million SNPs)
GWAS summary data for CKD
(CKD: ~9.2 million SNPs)
non-ambiguous strand mapping
exclude no rs labels of SNPs
CAD: 9,416,879 SNPs
CKD: 9,094,245 SNPs
shared SNPs with 1000G Project
exclude parts of SNPs in Chr 6/8
MAGMA
CAD: 5,253,977 SNPs
CKD: 5,269,810 SNPs
CAD: 17,231 genes
CKD: 17,223 genes
LDSC
GPA
cFDR
121 pleiotropic genes; among those 11 may be novel genes)
Validate analysis with another CAD dataset from UK Biobank

FIGURE 1 | Flowchart of data preparation and analysis for CKD and CAD in the present study. CAD, coronary artery disease; CKD, chronic kidney disease; MAGMA, Multi-marker Analysis of GenoMic Annotation; LDSC, linkage disequilibrium score regression; GPA, Genetic analysis incorporating Pleiotropy and Annotation; pleiotropy-informed methods, GPA and cFDR; cFDR, conditional false discovery rate; 1000G, 1000 Genomes Project phase III.

of 567,460 (64,164 cases and 502,296 controls; $N_{eff} = 227,584$) individuals of European ancestry and ∼9.6 million SNPs for CKD were left. We yielded summary statistics of CAD from the CARDIoGRAMplusC4D Consortium (Nikpay et al., 2015), which included 184,305 (60,801 cases and 123,504 controls; $N_{eff} = 162,972$) individuals of European ancestry and ∼9.4 million SNPs after quality control.

We further validated our results using another summary statistic of CAD obtained from the UKB cohort[1]. The UKB-CAD dataset included 405,940 individuals of European ancestry (23,888 cases and 382,052 controls; $N_{eff} = 89,929$) and 23,861,747 SNPs after quality control (i.e., INFO scores >0.8, allele count at least 20 and minor allele count less than 20). The association in the UKB-CAD dataset was analyzed through the SAIGE method (Zhou et al., 2018), which implemented the logistic mixed model with a kinship matrix as random effects and age, sex, age × sex, age^2, age^2 × sex as well as the first ten principal components as fixed-effects covariates.

Estimated Overall Genetic Correlation With LDSC

We applied the cross-trait LDSC (Bulik-Sullivan B. et al., 2015) to assess the overall genetic correlation r_g between CKD and CAD using all available SNPs. The software of LDSC (version v1.0.1) was downloaded at https://github.com/bulik/ldsc and our analysis was conducted with default settings. Following prior studies (Bulik-Sullivan B. et al., 2015), we performed stringent quality control procedures during the LDSC analysis: (1) excluded non-biallelic SNPs and those with strand-ambiguous alleles; (2) excluded duplicated SNPs and those having no rs labels; (3) excluded SNPs that were located within two genetic regions including major histocompatibility complex (chr6: 28,500,000–33,500,000) (Bulik-Sullivan B. et al., 2015) and chr8: 7,250,000–12,500,000 (Price et al., 2008); (4) kept SNPs that were included in the 1000 Genomes Project phase III; (5) removed SNPs whose allele did not match that in the 1000 Genomes Project phase III (The 1000 Genomes Project Consortium, 2015).

The LD scores ℓ_j were computed using genotypes of 7,120,251 common SNPs (minor allele frequency >0.01 and the *P*-value of Hardy Weinberg equilibrium test >1E-5) with a 10 Mb window on 503 European individuals in the 1000 Genomes Project phase III (The 1000 Genomes Project Consortium, 2015); and then regressed on the product of *Z*-score statistics of the two diseases

$$E(z_{1j}z_{2j}) = \frac{\sqrt{N_1 N_2 \ell_j}}{M} \times r_g + \frac{\rho N_s}{\sqrt{N_1 N_2}} \qquad (1)$$

where N_1 and N_2 are the sample sizes for CAD and CKD, respectively; N_s is the number of individuals shared by the two GWASs, and ρ is the disease correlation among the N_s overlapping individuals. Theoretically, SNPs with high LD will have higher χ^2 statistics on average than those with low LD provided that the disease has a polygenic genetic foundation (Bulik-Sullivan B. K. et al., 2015). In terms of LSDC shown in

[1] https://pan.ukbb.broadinstitute.org/

(1), the regression slope provides an unbiased estimate for genetic correlation r_g and is in general not influenced by sample overlap (Bulik-Sullivan B. et al., 2015).

Summary Statistics-Based Gene-Level Association With MAGMA

Many gene-based association approaches with only summary statistics have been developed recently; among those MAGMA is a fast and flexible method and widely employed (de Leeuw et al., 2015). During the implementation of MAGMA, we defined the set of SNPs that were located within a given gene in terms of the annotation file provided in VAGIS (Liu et al., 2010). For numerical stability, we only focused on protein coding genes with at least ten SNPs (note that, this threshold was to some extent chosen arbitrarily). The genotypes of 503 European individuals in the 1000 Genomes Project phase III (The 1000 Genomes Project Consortium, 2015) were exploited as reference panel for calculating the LD matrix to incorporate the correlation structure among SNPs. After the implementation of MAGMA, the *P*-value for each gene can be available in the CAD or CKD GWAS. Depending on those *P*-values we attempted to discover significant genes that were related to CAD or CKD as well as potentially pleiotropic genes that were associated with both the two types of disease. To detect newly novel association signals, we ruled out identified genes located within 1 Mb on each side of previously reported CAD- or CKD associated genes or SNPs from the GWAS Catalog[2] as done similarly in other studies (Bis et al., 2020). Of note, doing this was a conservative strategy and might miss potentially important association signals although false discoveries were well controlled.

Pleiotropy-Informed Association Methods With Summary Statistics

To further leverage the pleiotropic information shared between CAD and CKD to identify gene association signals more efficiently, we employed two novel statistical genetic methods in the following. First, we utilized the cFDR method (Andreassen et al., 2013) which extended the unconditional FDR (Benjamini et al., 2001) from an empirical Bayes perspective. The cFDR measures the probability of the association of the principal disease conditioned on the strength of association with the conditional disease (Andreassen et al., 2013)

$$\mathrm{cFDR}(p_i || P_i \leq p_i,\ P_j \leq p_j) \tag{2}$$

where p_i and p_j are the observed *P*-values of a particular gene of the principal and conditional diseases, respectively; $H_0^{(i)}$ denotes the null hypothesis that there does not exist association between the gene and the principal disease.

Besides cFDR, we also carried out the GPA analysis (Chung et al., 2014), which was constructed as

$$\pi_{00} = \mathrm{Prob}(Z_{j00} = 1) : (P_{j1}|Z_{j00} = 1) \sim$$

$$U[0,\ 1],\ (P_{j2}|Z_{j00} = 1) \sim U[0,\ 1]$$

$$\pi_{10} = \mathrm{Prob}(Z_{j10} = 1) : (P_{j1}|Z_{j10} = 1) \sim$$

$$Beta(\alpha_1,\ 1),\ (P_{j2}|Z_{j10} = 1) \sim U[0,\ 1]$$

$$\pi_{01} = \mathrm{Prob}(Z_{j01} = 1) : (P_{j1}|Z_{j01} = 1) \sim$$

$$U[0,\ 1],\ (P_{j2}|Z_{j01} = 1) \sim Beta(\alpha_2,\ 1)$$

$$\pi_{11} = \mathrm{Prob}(Z_{j11} = 1) : (P_{j1}|Z_{j11} = 1) \sim$$

$$Beta(\alpha_1,\ 1),\ (P_{j2}|Z_{j11} = 1) \sim Beta(\alpha_2,\ 1) \tag{3}$$

where the latent variables $Z_j = (Z_{j00}, Z_{j10}, Z_{j01}, Z_{j11})$ indicates the association between the *j*-th gene and the two diseases: $Z_{j00} = 1$ denotes the *j*-th gene is associated with neither of them (with probability π_{00}), $Z_{j10} = 1$ denotes the *j*-th gene is only associated with the first one (with probability π_{10}), $Z_{j01} = 1$ denotes the *j*-th gene is only associated with the second one (with probability π_{01}), and $Z_{j11} = 1$ denotes the *j*-th gene is associated with both the diseases (with probability π_{11}), indicating the extent of common biological pathways to which the two diseases may share (Chung et al., 2014). In addition, α_1 and α_2 ($0 < \alpha_k < 1$, $k = 1, 2$) are unknown shape parameters of the Beta distribution.

Functional Analysis

To explore functional features of newly discovered pleiotropic genes, we performed functional enrichment analysis [e.g., Gene Ontology (GO) and KEGG pathway analysis] with DAVID 6.8[3] (Huang da et al., 2009). Enrichment analysis allows us to validate our findings by determining functional annotations for those genes with pleiotropic effects. We also conducted the protein–protein interaction analysis to detect interaction and association in terms of the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING 11.0 at https://string-db.org/) database (Szklarczyk et al., 2019). We implemented the signaling pathways of these significant genes by Cytoscape software and visualized them by CluePedia (Bindea et al., 2009).

RESULTS

Estimated Overall Genetic Correlation Between CAD and CKD

After quality control, a total of 5,253,977 and 5,269,810 genetic variants are reserved for CAD or CKD, respectively. The genome-wide SNP-based heritability is estimated to be 4.69% (SE = 0.35%) for CAD and 0.53% (SE = 0.12%) for CKD with LDSC. The genomic inflation factor (i.e., the ratio of the observed median χ^2 statistic to the expected median) is 1.015 for CAD and 1.143 for CKD, which, together the LDSC intercept [i.e., 0.903 (SE = 0.005) for CAD and 1.134 (SE = 0.007) for CKD], suggests that the weak inflation of the χ^2 statistic of CKD is primarily due to polygenicity rather than population stratification or cryptic

[2] www.ebi.ac.uk

[3] https://david.ncifcrf.gov/

relatedness. In terms of those results, the adjustment of genomic control is also not necessary.

Next, based on all overlapped genetic variants (i.e., 5,117,020 SNPs), using LDSC we observe there exists a positive genetic correlation between the two types of diseases [$\hat{r}_g = 0.173$, 95% confidence interval (CI) 0.023 ∼ 0.332, $P = 0.024$], providing empirical evidence that the two diseases share common genetic components. We further quantify genetic correlation between CAD and CKD separately in six functional categories (Gusev et al., 2014), including coding, UTR (untranslated region), promoter, DHS (DNaseI hypersensitivity sites), intronic and intergenic. It is found that all the estimates of r_g in those categories are positive, again supporting the statement that CAD and CKD have overlapped genetic foundation. In particular, there exists a significantly positive genetic correlation in the regions of DHS ($\hat{r}_g = 0.197$, 95% CI 0.074 ∼ 0.319, P = 1.60E-3) and intergenic ($\hat{r}_g = 0.264$, 95% CI 0.153 ∼ 0.375, P = 3.03E-6) (**Supplementary Table S2** and **Figure 2**).

Overall, through genetic correlation analysis we reveal that CAD and CKD are genetically similar and share moderate overlap in genetic etiology, especially at some certain regions. Therefore, it is worthy of additional investigation into shared genetic mechanisms through pleiotropy-informed statistical tools.

Associated Genes Identified With MAGMA, cFDR, and GPA

In our gene-based association analysis, we assign a set of genetic variants to predefined genes and obtain a total of 17,231 and 17,223 protein coding genes for CAD or CKD, respectively. Using MAGMA, we identify 763 CAD-associated genes and 827 CKD-associated genes (FDR < 0.05) (**Supplementary Tables 3, 4** and **Supplementary Figure 1**). Importantly, 25.8% (=197/763) CAD-associated genes (e.g., *ACER2*, *ACSS2*, *ARHGEF19,* and *BBS10*) and 60.7% (=503/827) CKD-associated genes (e.g., *BAG6*, *BAK1*, *BTNL2,* and *C4BPB*) are likely novel genes because those genes are not nearby (within 1 Mb upstream and downstream) any previous GWAS index SNPs or associated genes in terms of the GWAS catalog (McMahon et al., 2019).

In our cFDR analysis the Q–Q plot of CAD conditional on the nominal P-value of CKD illustrates the existence of enrichment at different significance thresholds of CKD (**Supplementary Figure 2A**). The presence of leftward shift suggests that the proportion of true associations for a given CKD P-value would increase when the analysis is limited to include more significant SNPs. On the other hand, in terms of the Q–Q plot of CKD conditional on the nominal P-value of CAD (**Supplementary Figure 2B**), we observe a more pronounced separation in different curves, implying that there exists a stronger enrichment for CKD given CAD than that for CAD given CKD. We further formally analyze the two diseases jointly using cFDR and show the results in **Supplementary Tables 5, 6** and **Supplementary Figure 3**. Briefly, with cFDR we identify 875 CAD-associated genes and 1,062 CKD-associated genes (cFDR < 0.05). Among those genes, 243 CAD-associated and 639 CKD-associated genes are possibly novel (**Supplementary Tables 5**, **6**). More interesting, all CAD-associated genes identified by MAGMA are replicated and 111 additional genes are discovered (**Supplementary Figure 4**); and all CKD-associated genes identified by MAGMA are also verified and 234 more genes are newly discovered (**Supplementary Figure 5**).

We next employ GPA to implement another integrative analysis for the two diseases. In terms of the GPA result we discover 504 and 1395 significant genes that are related to CAD or CKD (**Supplementary Tables 7, S8** and **Supplementary Figure 6**). Among those, 17.3% (=87/504) novel CAD-associated genes (e.g., *ACVR2A*, *AP3M1*, *ARHGEF19,* and *BACH1*) and 61.2% (=854/1395) CKD-associated genes (e.g., *ABCA4*, *ABCC2*, *ABCF3,* and *ACOX1*) may be newly novel genes because they are not nearby (within 1 Mb upstream and downstream) any previous GWAS index SNPs or associated genes in terms of the GWAS catalog (McMahon et al., 2019). Furthermore, we find 504 CAD-associated and 770 CKD-associated genes that are identified simultaneously by GPA and MAGMA (**Supplementary Figures 7**, **8**).

Identified Pleiotropic Gene With Both cFDR and GPA

According to the result of MAGMA, 72 genes are related to both CAD and CKD (**Supplementary Table 9** and **Figure 3A**). Based on the two integrative analyses, 169 genes are shared between CAD and CKD when using cFDR (**Supplementary Table 10** and **Figure 3B**) and 504 genes are shared between CAD and CKD when using GPA (**Supplementary Table 11** and **Figure 3C**). In addition, through GPA we observe that a substantial fraction of genes that are simultaneously related to CAD and CKD, with π_{11} estimated to be 8.2% (SE = 0.1%), offering additional statistical evidence supporting the existence of pleiotropy between CAD and CKD [the statistic of the likelihood ratio test is 225.6 and P = 5.35E-51 (Chung et al., 2014)].

Due to the difference of power in identifying pleiotropic genes via cFDR or GPA, we expect that a gene would be more likely to have pleiotropic effect if it is discovered by cFDR and GPA simultaneously. Relying on this principle we define a set of 121 genes that are associated with CAD and CKD and are jointly detected by cFDR and GPA to be pleiotropic genes (**Supplementary Table 12** and **Figure 3D**), among which five (i.e., *IGF2R*, *LPA*, *BCAS3*, *SLC22A2,* and *ATXN2*) were identified in previous studies (**Supplementary Table 1**). Furthermore, after ruling out genes located within 1 Mb on each side of previously reported genes or SNPs, we ultimately discover 11 newly novel pleiotropic genes associated with both CAD and CKD (i.e., *RHGEF19*, *RSG1*, *NDST2*, *CAMK2G*, *VCL*, *LRP10*, *RBM23*, *USP10*, *WNT9B*, *GOSR2,* and *RPRML*) (**Table 1** and **Supplementary Figures 9–13**).

Validation the Results in a Latest GWAS From the UK Biobank

We further validate the main results using the UKB-CAD summary statistics and show the results in **Supplementary Tables 13–16**. The genome-wide SNP-based heritability is estimated to be 2.42% (SE = 0.20%) for UKB-CAD with LDSC.

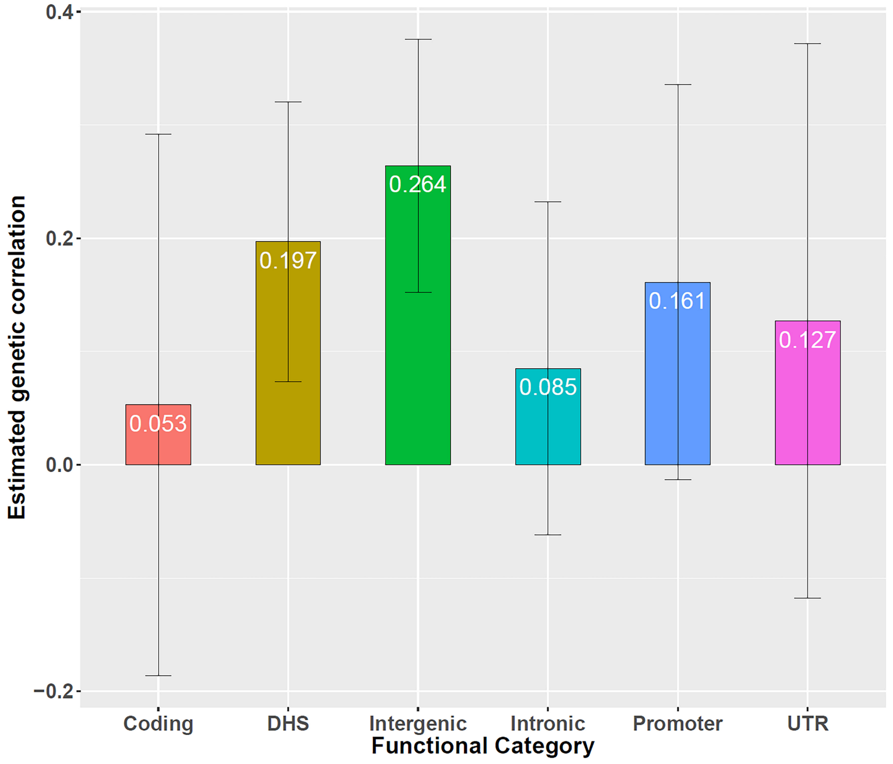

FIGURE 2 | Genetic correlation between CAD and CKD in six functional categories, including coding, UTR, promoter, DHS, intronic, and intergenic. Error bars show 1.96 × SE. Besides DHS and intergenic, the genetic correlation is $\hat{r}_g = 0.053$ (SE = 0.122, *P* = 6.64E-1) for coding, $\hat{r}_g = 0.127$ (SE = 0.125, *P* = 3.10E-1) for UTR, $\hat{r}_g = 0.161$ (SE = 0.089, *P* = 7.00E-2) for promoter, $\hat{r}_g = 0.089$ (SE = 0.075, *P* = 2.55E-1) for intronic.

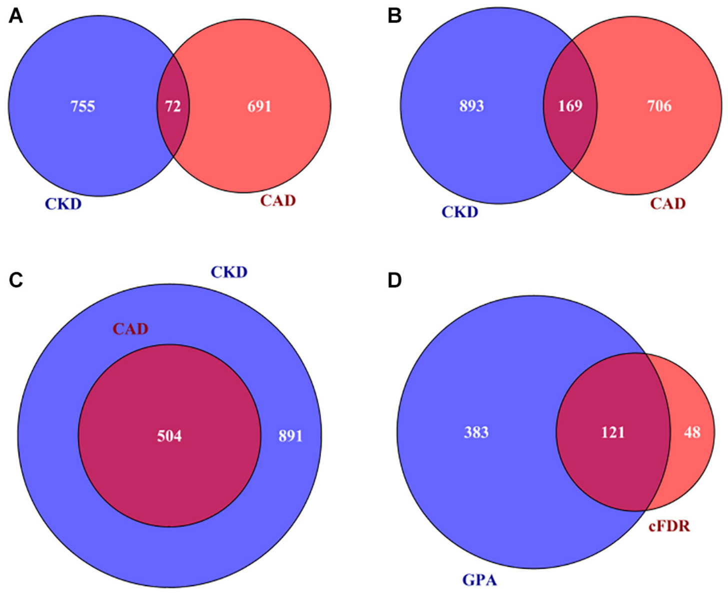

FIGURE 3 | (A) A total of 72 associated genes shared by CAD and CKD using MAGMA; **(B)** 169 associated genes shared by CAD and CKD using cFDR; **(C)** a total of 504 genes shared by CAD and CKD using GPA; **(D)** a total of 121 pleiotropic genes of CAD and CKD simultaneously discovered by cFDR and GPA. CAD, coronary artery disease; CKD, chronic kidney disease; MAGMA, Multi-marker Analysis of GenoMic Annotation; GPA, Genetic analysis incorporating Pleiotropy and Annotation; cFDR, conditional false discovery rate.

TABLE 1 | Pleiotropic genes associated with CAD and CKD identified by cFDR and GPA jointly.

Gene	CHR	Position	cFDR		GPA	
			CAD	CKD	CAD	CKD
ARHGEF19	1	16,424,598–16,639,104	2.55E-03	4.16E-02	1.19E-02	1.76E-03
RSG1	1	16,458,181–16,663,659	2.83E-03	3.38E-02	1.28E-02	1.47E-03
NDST2	10	75,461,668–75,671,589	1.09E-02	4.13E-03	4.69E-02	8.63E-04
CAMK2G	10	75,472,258–75,734,349	5.06E-03	3.02E-03	2.51E-02	4.13E-04
VCL	10	75,657,871–75,979,914	6.07E-03	1.52E-02	2.75E-02	1.58E-03
LRP10	14	23,240,959–23,447,291	8.22E-03	2.21E-02	3.35E-02	2.42E-03
RBM23	14	23,269,853–23,488,396	7.62E-03	2.80E-02	3.13E-02	2.78E-03
USP10	16	84,633,554–84,913,527	6.48E-03	2.63E-04	4.08E-02	1.01E-04
WNT9B	17	44,828,967–45,054,437	4.20E-04	3.37E-02	1.84E-03	2.47E-04
GOSR2	17	44,900,485–45,118,733	8.13E-04	3.22E-02	4.08E-03	5.39E-04
RPRML	17	44,955,521–45,156,614	1.90E-03	2.16E-02	8.30E-03	6.50E-04

We also do not observe a substantial inflation in the UKB-CAD summary statistics [the estimated λ = 1.178 with the intercept = 1.057 (SE = 0.005)].

According to the result of MAGMA, 184 genes are related to both UKB-CAD and CKD (**Supplementary Table 13** and **Supplementary Figure 14**). Based on the two pleiotropy-informed integrative analyses, 373 genes are shared between UKB-CAD and CKD using cFDR (**Supplementary Table 14** and **Supplementary Figure 15**) and 371 genes are shared between UKB-CAD and CKD using GPA (**Supplementary Table 15** and **Supplementary Figure 16**). All the 11 pleiotropic genes described above are also analyzed here and five (i.e., *RSG1*, *LRP10*, *RBM23*, *WNT9B,* and *GOSR2*) are replicated (**Supplementary Table 16**).

Functional Analyses for Pleiotropic Genes

We now undertake functional analyses for the 121 pleiotropic genes. Among these, most are located within chr 17 (20.7% = 25/121), followed by chr 1 (15.7% = 19/121) and chr 11 (12.4% = 15/121) (**Supplementary Figure 17**). In terms of the DAVID analysis, these genes are enriched in 34 GO terms (**Supplementary Table 17**). The top five candidate pathways include "dopamine transmembrane transporter activity" (P = 2.28E-04), "quaternary ammonium group transport" (P = 3.54E-04), "quaternary ammonium group transmembrane transporter activity" (P = 3.79E-04), "dopamine transport" (P = 7.38E-04), and "organic cation transmembrane transporter activity" (P = 7.89E-04). There pathways offer part of evidence supporting common genetic foundations between CAD and CKD. For instance, it has been shown that CKD patients had higher levels for some quaternary ammonium salts (e.g., choline) (Rennick et al., 1976), which were also risk factors for CAD (Guo et al., 2020). In our PPI analysis (**Supplementary Figure 18**), strong interactions are found among pleiotropic genes, such as *NDST2*, *CAMK2G*, *RASGRF1*, *IGF2R*, *SORT1,* and *TRIB1*. These genes were reported to be associated with organic cation transmembrane transporter, such as organic anion transporters oat1 and oat3, and organic cation transporters oct1 and oct2, which was also altered with chronic kidney failure in rats (Komazawa et al., 2013).

DISCUSSION

It has been widely observed that CAD and CKD share common pathological and clinical feature (Go et al., 2004; Liu et al., 2012; Tonelli et al., 2012; Ene-Iordache et al., 2016). However, the underlying genetic overlap between the two diseases remains unclear and a large proportion of genes related to CAD and CKD are yet discovered (Manolio et al., 2009). Large-scale GWASs undertaken for CAD and CKD offer an unprecedented opportunity to answer this question. In the present study a positive genetic correlation was found between CAD and CKD, implying genetic variants that were associated with the risk of CKD would be also related to the risk of CAD. This finding also partly explained the observed comorbidity of the two diseases (Go et al., 2004; Ene-Iordache et al., 2016).

Using existing well-established statistical approaches, we ultimately identified 11 novel pleiotropic genes shared by CAD and CKD, including *ARHGEF19*, *RSG1*, *NDST2*, *CAMK2G*, *VCL*, *LRP10*, *RBM23*, *USP10*, *WNT9B*, *GOSR2*, and *RPRML*, some of which were previously reported to play important roles in the pathogenesis of CAD or CKD (Agosti, 2002; Sivapalaratnam et al., 2012; Zanders, 2015). Furthermore, we also validated our main finding in an independent UKB-CAD dataset and replicated five genes.

Specifically, prior studies showed that *ARHGEF19* (Klarin et al., 2018) and *LRP10* (Sugiyama et al., 2000) were associated with total cholesterol and low-density lipoprotein (LDL) cholesterol, which were in turn related to CAD (Nissen et al., 2005) and CKD (Baigent et al., 2011). *RSG1* is involved in targeted membrane trafficking, and further involved in cilium biogenesis by regulating the transportation of cargo proteins to the basal body and apical tips of cilia with its protein (Agbu et al., 2018). Mice and humans with abnormal primary cilia can exhibit defects in cardiac morphogenesis, and also can cause kidney disease (Agbu et al., 2018).

NDST2 encodes a member of the N-deacetylase/N-sulfotransferase subfamily, which has dual functions (N-deacetylation and N-sulfation) in processing heparin polymers (Humphries et al., 1998). Inactivation of *NDST2* may impact the atherosclerosis by altering the structure of monocytes/macrophages heparan sulfate (HS) (Gordts et al., 2014), while also alter the glomerular HS to impact the primary kidney diseases (Goode et al., 1995). *CAMK2G* belongs to the Ca^{2+}/calmodulin-dependent protein kinase subfamily (Moyers et al., 1997). Vascular calcification correlates with the vessel stiffening and hypertension, and further increases the risk of atherosclerosis and myocardial infarction. It also exhibits a hugely elevated risk of cardiovascular mortality in CKD patients (Shanahan Catherine et al., 2011).

Vinculin (*VCL*) is a membrane-cytoskeletal protein, which associated with the linkage of integrin adhesion molecules to the actin cytoskeleton (Burridge and Feramisco, 1980), and the cell–cell and cell-matrix junctions, where it is thought to function in anchoring F-actin to the membrane (Geiger, 1979). Endothelial dysfunction caused by F-actin cytoskeleton disorder is a well-recognized instigator of cardiovascular diseases and CKD (Ding et al., 2016). *USP10* encodes a member of the ubiquitin-specific protease family of cysteine proteases (Wang et al., 2015; Lim et al., 2019). Inactivation of *USP10* can diminish Notch-induced target gene expression in endothelial cells. Importantly, tight quantitative and temporal control of Notch activity is essential for vascular development (Wang et al., 2015; Lim et al., 2019).

WNT9B, encodes the secreted signaling proteins (Garriock et al., 2007), is significantly associated with systolic blood pressure (Hoffmann et al., 2017), which is further related to the risk of CAD (Turner et al., 1998) and CKD (Jafar et al., 2003). *GOSR2* encodes a trafficking membrane protein which transports proteins among the medial- and trans-Golgi compartments (Bui et al., 1999). Due to its chromosomal location and trafficking function, *GOSR2* may be involved in familial essential hypertension (Boissé Lomax et al., 2013), and also was reported to be relevant to systolic blood pressure (Ehret et al., 2011) and CAD (van der Harst and Verweij, 2018).

The major strength of our work is that multiple pleiotropy-informed methods were implemented to detect pleiotropic genes by combining existing GWASs summary results without requiring individual-level datasets. Unlike previous studies (Andreassen et al., 2013; Chung et al., 2014; Zeng et al., 2018), we perform MAGMA methods to enrich a group of SNPs which may be likely associated with CAD or CKD but cannot reach genome-wide significance because of modest effects if using single marker analysis. Moreover, to minimize possible false discovery, we only reported pleiotropic genes that were simultaneously discovered by GPA and cFDR and thus were more likely to be related to both CAD and CKD. Therefore, our findings are robust.

Nevertheless, there are some limitations needed to state. First, we cannot replicate all these genes via *in vivo* and *in vitro* experiments. Second, the individuals involved in our study are of European ancestry, it is not clear whether the finding can be generalized to other populations because of ethnic diversity in genetics. Third, although empirical evidence shown above indicates that the newly identified pleiotropic genes may underlie certain aspects of the pathogenesis of CAD and CKD in a direct or indirect way, the causally biological mechanisms of those genes are still largely unclear; therefore, further studies are needed to completely delineate their functions on CAD and CKD.

CONCLUSION

This study identifies common genetic architectures overlapped between CAD and CKD and will help to advance understanding of the molecular mechanisms underlying the comorbidity of the two diseases.

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

PZ and SH conceived the idea for the study. PZ, TW, and HC obtained the data. PZ and HC performed the data analyses and wrote the manuscript with the participation of all authors. PZ, JY, TW, and HC interpreted the results of the data analyses. All authors contributed to the article and approved the submitted version.

FUNDING

The research of PZ was supported in part by the Youth Foundation of Humanity and Social Science funded by Ministry of Education of China (18YJC910002), the Natural Science Foundation of Jiangsu Province of China (BK20181472), the China Postdoctoral Science Foundation (2018M630607 and 2019T120465), the QingLan Research Project of Jiangsu Province for Outstanding Young Teachers, the Six-Talent Peaks Project in Jiangsu Province of China (WSN-087), the Training Project for Youth Teams of Science and Technology Innovation at Xuzhou Medical University (TD202008), the Postdoctoral Science Foundation of Xuzhou Medical University, the National Natural Science Foundation of China (81402765), Postgraduate Research & Practice Innovation Program of Jiangsu (KYCX20_2501), and the Statistical Science Research Project from National Bureau of Statistics of China (2014LY112). The research of SH was supported in part by the Social Development Project of Xuzhou City (KC19017). The research of TW was supported in part by the Social Development Project of Xuzhou City (KC20062).

ACKNOWLEDGMENTS

We thank the CKDGen Consortium and the CARDIoGRAMplusC4D Consortium who made the summary data of CKD and CAD publicly available. The genetic and phenotypic UK Biobank data were available at https://pan.ukbb.broadinstitute.org/. We are also grateful to all the investigators and participants contributed to those studies. The data analyses in the present study were supported by the high-performance computing cluster at Xuzhou Medical University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.592461/full#supplementary-material

REFERENCES

Agbu, S. O., Liang, Y., Liu, A., and Anderson, K. V. (2018). The small GTPase RSG1 controls a final step in primary cilia initiation. *J. Cell Biol.* 217, 413–427. doi: 10.1083/jcb.201604048

Agosti, J. (2002). *Biotherapeutic Approaches to Asthma*. Boca Raton, FL: CRC Press.

Altshuler, D., Daly, M., and Lander, E. (2008). Genetic mapping in human disease. *Science* 322, 881–888. doi: 10.1126/science.1156409

Andreassen, O. A., Djurovic, S., Thompson, W. K., Schork, A. J., Kendler, K. S., O'Donovan, M. C., et al. (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.* 92, 197–209. doi: 10.1016/j.ajhg.2013.01.001

Baigent, C., Landray, M. J., Reith, C., Emberson, J., Wheeler, D. C., Tomson, C., et al. (2011). The effects of lowering LDL cholesterol with simvastatin plus ezetimibe in patients with chronic kidney disease (Study of Heart and Renal Protection): a randomised placebo-controlled trial. *Lancet* 377, 2181–2192. doi: 10.1016/S0140-6736(11)60739-3

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* 125, 279–284. doi: 10.1016/s0166-4328(01)00297-2

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Bis, J. C., Jian, X., Kunkle, B. W., Chen, Y., Hamilton-Nelson, K. L., Bush, W. S., et al. (2020). Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatry* 25, 1859–1875. doi: 10.1038/s41380-018-0112-7

Boissé Lomax, L., Bayly, M. A., Hjalgrim, H., Møller, R. S., Vlaar, A. M., Aaberg, K. M., et al. (2013). 'North Sea' progressive myoclonus epilepsy: phenotype of subjects with GOSR2 mutation. *Brain* 136(Pt 4), 1146–1154. doi: 10.1093/brain/awt021

Bui, T. D., Levy, E. R., Subramaniam, V. N., Lowe, S. L., and Hong, W. (1999). cDNA characterization and chromosomal mapping of human golgi SNARE GS27 and GS28 to chromosome 17. *Genomics* 57, 285–288. doi: 10.1006/geno.1998.5649

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47, 1236–1241. doi: 10.1038/ng.3406

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47:291. doi: 10.1038/ng.3211

Burridge, K., and Feramisco, J. R. (1980). Microinjection and localization of a 130K protein in living fibroblasts: a relationship to actin and fibronectin. *Cell* 19, 587–595. doi: 10.1016/s0092-8674(80)80035-3

Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet.* 10:e1004787. doi: 10.1371/journal.pgen.1004787

de Leeuw, C. A., Mooij, J. M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* 11:e1004219. doi: 10.1371/journal.pcbi.1004219

Ding, N., Liu, B., Song, J., Bao, S., Zhen, J., Lv, Z., et al. (2016). Leptin promotes endothelial dysfunction in chronic kidney disease through AKT/GSK3β and β-catenin signals. *Biochem. Biophys. Res. Commun.* 480, 544–551. doi: 10.1016/j.bbrc.2016.10.079

Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478, 103–109. doi: 10.1038/nature10405

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* 11, 446–450. doi: 10.1038/nrg2809

Ene-Iordache, B., Perico, N., Bikbov, B., Carminati, S., Remuzzi, A., Perna, A., et al. (2016). Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study. *Lancet Glob. Health* 4, e307–e319. doi: 10.1016/s2214-109x(16)00071-71

Foley, R. N., Murray, A. M., Li, S., Herzog, C. A., McBean, A. M., Eggers, P. W., et al. (2005). Chronic kidney disease and the risk for cardiovascular disease, renal replacement, and death in the United States Medicare population, 1998 to 1999. *J. Am. Soc. Nephrol.* 16, 489–495. doi: 10.1681/asn.2004030203

Garriock, R. J., Warkman, A. S., Meadows, S. M., D'Agostino, S., and Krieg, P. A. (2007). Census of vertebrate Wnt genes: isolation and developmental expression of *Xenopus* Wnt2, Wnt3, Wnt9a, Wnt9b, Wnt10a, and Wnt16. *Dev. Dyn.* 236, 1249–1258. doi: 10.1002/dvdy.21156

Geiger, B. (1979). A 130K protein from chicken gizzard: its localization at the termini of microfilament bundles in cultured chicken cells. *Cell* 18, 193–205. doi: 10.1016/0092-8674(79)90368-4

Girirajan, S. (2017). Missing heritability and where to find it. *Genome Biol.* 18:89. doi: 10.1186/s13059-017-1227-x

Go, A. S., Chertow, G. M., Fan, D., McCulloch, C. E., and Hsu, C. Y. (2004). Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *N. Engl. J. Med.* 351, 1296–1305. doi: 10.1056/NEJMoa041031

Goode, N. P., Shires, M., Crellin, D. M., Aparicio, S. R., and Davison, A. M. (1995). Alterations of glomerular basement membrane charge and structure in diabetic nephropathy. *Diabetologia* 38, 1455–1465. doi: 10.1007/bf00400607

Gordts, P. L., Foley, S. M., Erin, M., Lawrence, R., Sinha, R., Lameda-Diaz, C., et al. (2014). Reducing macrophage proteoglycan sulfation increases atherosclerosis and obesity through enhanced Type I interferon signaling. *Cell Metab.* 20, 813–826. doi: 10.1016/j.cmet.2014.09.016

Guo, F., Zhou, J., Li, Z., Yu, Z., and Ouyang, D. (2020). The association between trimethylamine N-Oxide and its predecessors choline, L-carnitine, and betaine with coronary artery disease and artery stenosis. *Cardiol. Res. Pract.* 2020:5854919. doi: 10.1155/2020/5854919

Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B. J., Diogo, D., Stahl, E. A., et al. (2013). Quantifying missing heritability at known GWAS loci. *PLoS Genet.* 9:e1003993. doi: 10.1371/journal.pone.1003993

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* 95, 535–552. doi: 10.1016/j.ajhg.2014.10.004

Hoffmann, T. J., Ehret, G. B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P.-Y., et al. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* 49, 54–64. doi: 10.1038/ng.3715

Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B., and Eskin, E. (2015). Identification of causal genes for complex traits. *Bioinformatics* 31, 206–213. doi: 10.1093/bioinformatics/btv240

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–U484. doi: 10.1534/genetics.114.167908

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Humphries, D. E., Lanciotti, J., and Karlinsky, J. B. (1998). cDNA cloning, genomic organization and chromosomal localization of human heparan glucosaminyl N-deacetylase/N-sulphotransferase-2. *Biochem. J.* 332(Pt 2), 303–307. doi: 10.1042/bj3320303

Inrig, J. K., Califf, R. M., Tasneem, A., Vegunta, R. K., Molina, C., Stanifer, J. W., et al. (2014). The landscape of clinical trials in nephrology: a systematic review of Clinicaltrials.gov. *Am. J. Kidney Dis.* 63, 771–780. doi: 10.1053/j.ajkd.2013.10.043

Jafar, T. H., Stark, P. C., Schmid, C. H., Landa, M., Maschio, G., de Jong, P. E., et al. (2003). Progression of chronic kidney disease: the role of blood pressure control, Proteinuria, and Angiotensin-converting enzyme inhibition: a patient-level meta-analysis. *Ann. Intern. Med.* 139, 244–252. doi: 10.7326/0003-4819-139-4-200308190-00006

Kessler, T., Erdmann, J., and Schunkert, H. (2013). Genetics of coronary artery disease and myocardial infarction–2013. *Curr. Cardiol. Rep.* 15:368. doi: 10.1007/s11886-013-0368-0

Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstroem, S., Kraft, P., et al. (2016). Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* 2020:btw615.

Kim, H., Grueneberg, A., Vazquez, A. I, Hsu, S., and de los Campos, G. (2017). Will big data close the missing heritability gap? *Genetics* 207, 1135–1145. doi: 10.1534/genetics.117.300271

Klarin, D., Damrauer, S. M., Cho, K., Sun, Y. V., Teslovich, T. M., Honerlaw, J., et al. (2018). Genetics of blood lipids among ∼300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* 50, 1514–1523. doi: 10.1038/s41588-018-0222-9

Komazawa, H., Yamaguchi, H., Hidaka, K., Ogura, J., Kobayashi, M., and Iseki, K. (2013). Renal uptake of substrates for organic anion transporters Oat1 and Oat3 and organic cation transporters Oct1 and Oct2 is altered in rats with adenine-induced chronic renal failure. *J. Pharm. Sci.* 102, 1086–1094. doi: 10.1002/jps.23433

Levin, A., Tonelli, M., Bonventre, J., Coresh, J., Donner, J. A., Fogo, A. B., et al. (2017). Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. *Lancet* 390, 1888–1917. doi: 10.1016/s0140-6736(17)30788-2

Lim, R., Sugino, T., Nolte, H., Andrade, J., Zimmermann, B., Shi, C., et al. (2019). Deubiquitinase USP10 regulates Notch signaling in the endothelium. *Science* 364, 188–193. doi: 10.1126/science.aat0778

Lindner, A., Charra, B., Sherrard, D. J., and Scribner, B. H. (1974). Accelerated atherosclerosis in prolonged maintenance hemodialysis. *N. Engl. J. Med.* 290, 697–701.

Liu, H., Yan, L., Ma, G. S., Zhang, L. P., Gao, M., Wang, Y. L., et al. (2012). Association of chronic kidney disease and coronary artery disease in 1,010 consecutive patients undergoing coronary angiography. *J. Nephrol.* 25, 219–224. doi: 10.5301/jn.2011.8478

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* 87, 139–145. doi: 10.1016/j.ajhg.2010.06.009

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

McMahon, A., Malangone, C., Suveges, D., Sollis, E., Cunningham, F., Riat, H. S., 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

Moody, W. E., Edwards, N. C., Madhani, M., Chue, C. D., Steeds, R. P., Ferro, C. J., et al. (2012). Endothelial dysfunction and cardiovascular disease in early-stage chronic kidney disease: cause or association? *Atherosclerosis* 223, 86–94. doi: 10.1016/j.atherosclerosis.2012.01.043

Moyers, J. S., Bilan, P. J., Zhu, J., and Kahn, C. R. (1997). Rad and Rad-related GTPases interact with calmodulin and calmodulin-dependent protein kinase II. *J. Biol. Chem.* 272, 11832–11839. doi: 10.1074/jbc.272.18.11832

Musunuru, K., and Kathiresan, S. (2019). Genetics of common, complex coronary artery disease. *Cell* 177, 132–145. doi: 10.1016/j.cell.2019.02.015

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47, 1121–1130. doi: 10.1038/ng.3396

Nissen, S. E., Tuzcu, E. M., Schoenhagen, P., Crowe, T., Sasiela, W. J., Tsai, J., et al. (2005). Statin therapy, LDL cholesterol, C-reactive protein, and coronary artery disease. *New Engl. J. Med.* 352, 29–38. doi: 10.1056/NEJMoa042000

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., et al. (2008). Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* 83, 132–135. doi: 10.1016/j.ajhg.2008.06.005

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Rennick, B., Acara, M., Hysert, P., and Mookerjee, B. (1976). Choline loss during hemodialysis: homeostatic control of plasma choline concentrations. *Kidney Int.* 10, 329–335. doi: 10.1038/ki.1976.116

Ross, R. (1999). Atherosclerosis–an inflammatory disease. *N .Engl. J. Med.* 340, 115–126. doi: 10.1056/nejm199901143400207

Shanahan Catherine, M., Crouthamel Matthew, H., Kapustin, A., Giachelli Cecilia, M., and Towler Dwight, A. (2011). Arterial calcification in chronic kidney disease: key roles for calcium and phosphate. *Circ. Res.* 109, 697–711. doi: 10.1161/CIRCRESAHA.110.234914

Sivapalaratnam, S., Basart, H., Watkins, N. A., Maiwald, S., Rendon, A., Krishnan, U., et al. (2012). Monocyte gene expression signature of patients with early onset coronary artery disease. *PLoS One* 7:e32166. doi: 10.1371/journal.pone.0032166

Smeland, O. B., Frei, O., Shadrin, A., O'Connell, K., Fan, C.-C., Bahrami, S., et al. (2020). Discovery of shared genomic loci using the conditional false discovery rate approach. *Hum. Genet.* 139, 85–94. doi: 10.1007/s00439-019-02060-2

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 14, 483–495. doi: 10.1038/nrg3461

Sugiyama, T., Kumagai, H., Morikawa, Y., Wada, Y., Sugiyama, A., Yasuda, K., et al. (2000). A novel low-density lipoprotein receptor-related protein mediating cellular uptake of apolipoprotein E-enriched beta-VLDL in vitro. *Biochemistry* 39, 15817–15825. doi: 10.1021/bi001583s

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484. doi: 10.1038/s41576-019-0127-1

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Tonelli, M., Muntner, P., Lloyd, A., Manns, B. J., Klarenbach, S., Pannu, N., et al. (2012). Risk of coronary events in people with chronic kidney disease compared with those with diabetes: a population-level cohort study. *Lancet* 380, 807–814. doi: 10.1016/s0140-6736(12)60572-8

Turner, R. C., Millns, H., Neil, H. A. W., Stratton, I. M., Manley, S. E., Matthews, D. R., et al. (1998). Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United Kingdom prospective diabetes study (UKPDS: 23). *BMJ* 316:823. doi: 10.1136/bmj.316.7134.823

van der Harst, P., and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* 122, 433–443. doi: 10.1161/circresaha.117.312086

Wang, W., Huang, X., Xin, H. B., Fu, M., Xue, A., and Wu, Z. H. (2015). TRAF family member-associated NF-κB activator (TANK) inhibits genotoxic nuclear factor κB activation by facilitating deubiquitinase USP10-dependent Deubiquitination of TRAF6 ligase. *J. Biol. Chem.* 290, 13372–13385. doi: 10.1074/jbc.M115.643767

Wang, W. Y., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* 6, 109–118. doi: 10.1038/nrg1522

Wen, X., Luca, F., and Pique-Regi, R. (2015). Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* 11:e1005176. doi: 10.1371/journal.pgen.1005176

Wuttke, M., Li, Y., Li, M., Sieber, K. B., Feitosa, M. F., Gorski, M., et al. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* 51, 957–972. doi: 10.1038/s41588-019-0407-x

Young, A. I. (2019). Solving the missing heritability problem. *PLoS Genet.* 15:e1008222. doi: 10.1371/journal.pgen.1008222

Zanders, E. D. (2015). *Human Drug Targets: A Compendium for Pharmaceutical Discovery*. Hoboken, NJ: Wiley.

Zeng, P., Hao, X., and Zhou, X. (2018). Pleiotropic mapping and annotation selection in genome-wide association studies with penalized Gaussian mixture models. *Bioinformatics* 34, 2797–2807. doi: 10.1093/bioinformatics/bty204

Zeng, P., Zhao, Y., Liu, J., Liu, L., Zhang, L., Wang, T., et al. (2014). Likelihood ratio tests in rare variant detection for continuous phenotypes. *Ann. Hum. Genet.* 78, 320–332. doi: 10.1111/ahg.12071

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., et al. (2015). Statistical analysis for genome-wide association study. *J. Biomed. Res.* 29, 285–297. doi: 10.7555/jbr.29.20140007

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* 50, 1335–1341. doi: 10.1038/s41588-018-0184-y

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 December 2020
doi: 10.3389/fgene.2020.604324

Integrative Analysis of Transcriptome-Wide Association Study and mRNA Expression Profiles Identifies Candidate Genes Associated With Idiopathic Pulmonary Fibrosis

Weiming Gong[1†], *Ping Guo*[1†], *Lu Liu*[1], *Qingbo Guan*[2,3,4*] *and Zhongshang Yuan*[1*]

[1] *Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China,* [2] *Department of Endocrinology, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China,* [3] *Shandong Clinical Medical Center of Endocrinology and Metabolism, Jinan, China,* [4] *Shandong Institute of Endocrine and Metabolic Diseases, Jinan, China*

OPEN ACCESS

Edited by:
Sheng Yang,
Nanjing Medical University, China

Reviewed by:
Xiaohui Niu,
Huazhong Agricultural University, China
Haitao Yang,
Hebei Medical University, China

****Correspondence:***
Qingbo Guan
doctorguanqingbo@163.com
Zhongshang Yuan
yuanzhongshang@sdu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 September 2020*
Accepted: *17 November 2020*
Published: *10 December 2020*

Citation:
Gong W, Guo P, Liu L, Guan Q and Yuan Z (2020) Integrative Analysis of Transcriptome-Wide Association Study and mRNA Expression Profiles Identifies Candidate Genes Associated With Idiopathic Pulmonary Fibrosis. Front. Genet. 11:604324. doi: 10.3389/fgene.2020.604324

Idiopathic pulmonary fibrosis (IPF) is a type of scarring lung disease characterized by a chronic, progressive, and irreversible decline in lung function. The genetic basis of IPF remains elusive. A transcriptome-wide association study (TWAS) of IPF was performed by FUSION using gene expression weights of three tissues combined with a large-scale genome-wide association study (GWAS) dataset, totally involving 2,668 IPF cases and 8,591 controls. Significant genes identified by TWAS were then subjected to gene ontology (GO) and pathway enrichment analysis. The overlapped GO terms and pathways between enrichment analysis of TWAS significant genes and differentially expressed genes (DEGs) from the genome-wide mRNA expression profiling of IPF were also identified. For TWAS significant genes, protein–protein interaction (PPI) network and clustering modules analyses were further conducted using STRING and Cytoscape. Overall, TWAS identified a group of candidate genes for IPF under the Bonferroni corrected P value threshold ($0.05/14929 = 3.35 \times 10^{-6}$), such as *DSP* ($P_{TWAS} = 1.35 \times 10^{-29}$ for lung tissue), *MUC5B* ($P_{TWAS} = 1.09 \times 10^{-28}$ for lung tissue), and *TOLLIP* ($P_{TWAS} = 1.41 \times 10^{-15}$ for whole blood). Pathway enrichment analysis identified multiple candidate pathways, such as herpes simplex infection (P value = 7.93×10^{-5}) and antigen processing and presentation (P value = 6.55×10^{-5}). 38 common GO terms and 8 KEGG pathways shared by enrichment analysis of TWAS significant genes and DEGs were identified. In the PPI network, 14 genes (*DYNLL1*, *DYNC1LI1*, *DYNLL2*, *HLA-DRB5*, *HLA-DPB1*, *HLA-DQB2*, *HLA-DQA2*, *HLA-DQB1*, *HLA-DRB1*, *POLR2L*, *CENPP*, *CENPK*, *NUP133*, and *NUP107*) were simultaneously detected by hub gene and module analysis. In conclusion, through integrative analysis of TWAS and mRNA expression profiles, we identified multiple novel candidate genes, GO terms and pathways for IPF, which contributes to the understanding of the genetic mechanism of IPF.

Keywords: idiopathic pulmonary fibrosis, transcriptome-wide association study, gene expression profiling, pathway enrichment, protein–protein interaction network

INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a chronic interstitial lung disease characterized by the formation of scar tissue and a progressive, and irreversible decline in lung function (Raghu et al., 2011; Lederer and Martinez, 2018), with a median survival time from diagnosis of 2–4 years (Ley et al., 2011). The incidence of IPF is increasing worldwide and has been estimated to be 3–9 cases per 100,000 people per year in Europe and North America, and fewer than four cases per 100,000 people per year in East Asia and South America (Hutchinson et al., 2015). IPF has been confirmed to be related to varieties of environmental and genetic factors. Potential risk factors for IPF include aging, male sex, smoking, certain occupational exposures (Baumgartner et al., 2000), gastroesophageal reflux (Tobin et al., 1998; Bedard Methot et al., 2019), herpesvirus infection (Tang et al., 2003), air pollution (Sack et al., 2017), and obstructive sleep apnea (Kim et al., 2017). Genome-wide association studies (GWAS) on IPF (Mushiroda et al., 2008; Fingerlin et al., 2013, 2016; Noth et al., 2013; Allen et al., 2017, 2020) have identified common genetic variants related to IPF, highlighting the significance of several IPF susceptibility factors, such as telomere maintenance, host defense, cell-cell adhesion. Rare genetic variants regarding surfactant dysfunction and telomere biology have also been identified in studies of familial pulmonary fibrosis (Nogee et al., 2001; Armanios et al., 2007; Cogan et al., 2015; Stuart et al., 2015).

Genome-wide association study have significantly succeeded in identifying IPF-related susceptibility genetic loci. However, a great number of genetic variations identified reside in non-coding regions, which are generally difficult to characterize biologically. Indeed, one common sense for GWAS is that most disease-associated genetic variants are located in non-coding regions, resulting in the hypothesis that the underlying biological mechanism of disease may be closely related to gene expression regulation. Furthermore, several expression quantitative trait loci (eQTLs) studies have illustrated that the information on expression regulation may play a pivotal role in disease development (Albert and Kruglyak, 2015). Transcriptome-wide association study (TWAS) is widely utilized in integrating GWAS with eQTL studies for investigating the causal genes associated with complex traits or diseases (Gamazon et al., 2015; Gusev et al., 2016; Yuan et al., 2020). Therefore, TWAS analysis may help us to identify novel genes associated with IPF. On the other hand, the genome-wide mRNA expression profiling of IPF provides the opportunity to identify differentially expressed genes (DEGs). Furthermore, omics integrative analysis can combine different types of omics data and provides more comprehensive insights than that offered by any single type of omics data (Liu et al., 2013). These integrative analyses are implemented and expected to rebuild meaningful biological networks by integrating information from different types of data, thus have the potential to provide a more novel and reliable understanding with respect to the underlying biological mechanisms. Statistically, complementary information can be better captured and exploited by such data integration analyses (Yu and Zeng, 2018). Indeed, in high-throughput genomic studies, one common sense is that the analysis from single dataset often lack of reproducibility and integrative analysis can efficiently investigate and make full use of multiple datasets in a cost-efficient manner to enhance reproducibility (Yu and Zeng, 2018). This motivated us to perform a comprehensive integrative analysis of TWAS and mRNA expression profile of IPF, which may provide the better understanding of the molecular mechanisms of IPF.

In the present study, we leveraged expression imputation from a large-scale IPF GWAS dataset to perform a TWAS analysis in peripheral blood, whole blood and lung tissue. The TWAS significant genes and DEGs identified by mRNA expression profiling of IPF were then subjected to gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. Through this analysis, the common GO terms and KEGG pathways were identified. Furthermore, for significant genes identified by TWAS for IPF, STRING and Cytoscape software were applied to implement protein–protein interaction (PPI) network and clustering modules analyses. Our results may provide novel insights into the understanding of the molecular mechanisms underlying the development of IPF. The detailed procedure of integrative analysis was displayed in **Figure 1**.

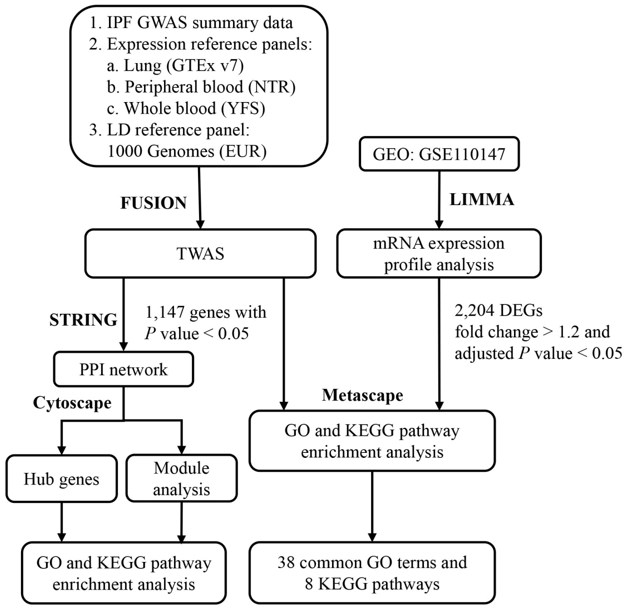
1. IPF GWAS summary data
2. Expression reference panels:
a. Lung (GTEx v7)
b. Peripheral blood (NTR)
c. Whole blood (YFS)
3. LD reference panel:
1000 Genomes (EUR)
GEO: GSE110147
FUSION
LIMMA
TWAS
mRNA expression profile analysis
STRING
1,147 genes with P value < 0.05
2,204 DEGs
fold change > 1.2 and adjusted P value < 0.05
PPI network
Cytoscape
Metascape
GO and KEGG pathway enrichment analysis
Hub genes
Module analysis
GO and KEGG pathway enrichment analysis
38 common GO terms and 8 KEGG pathways

FIGURE 1 | The flowchart illustrates the procedure of integrative analysis of TWAS and mRNA expression profile of IPF. Software for the integrative analysis were shown in bold. IPF, idiopathic pulmonary fibrosis; GWAS, genome-wide association study; GTEx, Genotype-Tissue Expression Project; NTR, Netherlands Twin Registry study; YFS, Young Finns Study; EUR, European; LD, linkage disequilibrium; TWAS, transcriptome-wide association study; GEO, Gene Expression Omnibus database; DEGs, differentially expressed genes; STRING, Search Tool for the Retrieval of Interacting Genes; PPI, protein–protein interaction; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

MATERIALS AND METHODS

GWAS of IPF

The current and largest-scale GWAS summary data of IPF were used (Allen et al., 2020). Briefly, it included 2,668 IPF cases and 8,591 controls of European ancestry from a meta-analysis of three case-control studies and restricted to unrelated individuals of European ancestry. Genotype data were imputed using the Haplotype Reference Consortium r1.1 panel. Stringent quality control was performed for the genotyped data. In each separate study, a genome-wide analysis of IPF susceptibility was conducted using SNPTEST v2.5.2, adjusting for the first 10 principal components to account for fine-scale population structure. Only biallelic autosomal variants that had a minor allele count ≥10, were in Hardy-Weinberg Equilibrium ($P > 1 \times 10^{-6}$), and well-imputed (imputation quality $R^2 > 0.5$) in at least two studies were included. Detailed description related to study participants, genotyping, imputation, association analysis, and quality control can be found in the previous IPF GWAS study (Allen et al., 2020).

TWAS of IPF

FUSION software was applied here for tissue-related TWAS analysis (Gusev et al., 2016). Briefly, FUSION leveraged a set of reference individuals to measure both gene expression and SNPs, and then to impute the *cis* genetic component of expression into a much larger set of individuals using their SNP genotype data. The imputed expression data can be viewed as a linear model of genotypes with weights based on the correlation between SNPs and gene expression in the reference data while accounting for linkage disequilibrium among SNPs. FUSION uses pre-computed gene expression weights together with disease GWAS summary statistics to evaluate the association between the expression levels of genes and target diseases (Gusev et al., 2016). The genetic values of expression were computed as one probe set at a time using SNP genotyping data located 500 kb on both sides of the gene boundary. The pre-computed expression reference weights of different tissues were downloaded from the FUSION websites[1]. For IPF TWAS, we used three expression reference panels, including lung, peripheral blood and whole blood, and a TWAS *P* value was obtained for each gene. Gene expression weights of lung were driven from the Genotype-Tissue Expression Project (GTEx v7; $n = 383$) (GTEx Consortium et al., 2017). Gene expression weights of peripheral blood and whole blood reference panels were driven from the Netherlands Twin Registry study (NTR) ($n = 1,247$) (Boomsma et al., 2006; Wright et al., 2014) and Young Finns Study (YFS) ($n = 1,264$) (Raitakari et al., 2008), respectively.

Gene Expression Profile Associated With IPF

The IPF gene expression profile data of lung tissue were obtained from the Gene Expression Omnibus database (access number: GSE110147) (Cecchini et al., 2018). Briefly, fresh frozen lung samples were obtained from the organs of 22 patients with IPF; normal lung tissue ($n = 11$) was obtained from the tissue flanking lung cancer resections. RNA was extracted and hybridized on Affymetrix microarrays. Individual-level gene expression data were included in the mRNA expression profile analysis implemented by LIMMA package (Ritchie et al., 2015). The DEGs between IPF patients and controls were identified at fold change >1.2 and adjusted *P* value < 0.05. Detailed description of sample characteristics, experimental design, statistical analysis, and quality control can be found in the previous study (Cecchini et al., 2018).

Gene Set Enrichment Analysis

The IPF-related genes identified by TWAS and mRNA expression profiling were, respectively, subjected to GO and KEGG pathway enrichment analysis implemented by Metascape (Zhou et al., 2019)[2]. Note that in the enrichment analysis for IPF-related genes identified by TWAS, we included all the genes with a TWAS *P* value less than 0.05, rather than those under the Bonferroni corrected *P* value threshold ($0.05/14929 = 3.35 \times 10^{-6}$), to increase the ability to identify more biological processes relevant to IPF and to make the results more stable by including more input genes (Reimand et al., 2019). A *P* value was calculated by Metascape for each GO term and pathway. Terms with a *P* value < 0.01, a minimum count of 3, and an enrichment factor >1.5 were collected and grouped into clusters based on their membership similarities. Kappa scores were used as the similarity metric when performing hierarchical clustering on the enriched terms, and sub-trees with a similarity of >0.3 were considered a cluster. The most statistically significant term within a cluster was chosen to represent the cluster. Finally, the Metascape analysis of TWAS was compared with that of mRNA expression profiles of lung tissue to identify the common GO terms and pathways shared by enrichment analysis for IPF-related genes from TWAS and for DEGs from mRNA expression profiling of IPF. Note that the common GO terms were obtained by overlapping the original GO enrichment results before grouping into clusters.

Protein–Protein Interaction Network, Hub Genes, and Module Analysis

The PPI network of TWAS significant genes was constructed by the online Search Tool for the Retrieval of Interacting Genes (Szklarczyk et al., 2019) (STRING; 2017 release) database to evaluate the interactive relationships among the genes. Interactions with a combined score >0.9 were defined as statistically significant. Cytoscape software (Shannon et al., 2003) (version 3.5.1) was applied to visualize the integrated regulatory networks. The cytoHubba plugin and Molecular Complex Detection (MCODE) plugin in Cytoscape were used to identify hub genes and screen modules of the PPI network. All parameters of the plugin were set at their default values. Again, GO and KEGG enrichment of hub genes and genes in modules were also analyzed by Metascape.

[1] http://gusevlab.org/projects/fusion/

[2] http://metascape.org

RESULTS

TWAS Analysis Results

Totally, 14,929 genes were analyzed by TWAS in this study. Overall, TWAS identified 29 genes under the Bonferroni corrected P value threshold ($0.05/14929 = 3.35 \times 10^{-6}$) and 1,147 genes with P value < 0.05 (**Supplementary Table S1**), such as *DSP* ($P_{TWAS} = 1.35 \times 10^{-29}$ for lung tissue), *MUC5B* ($P_{TWAS} = 1.09 \times 10^{-28}$ for lung tissue), *TOLLIP* ($P_{TWAS} = 1.41 \times 10^{-15}$ for whole blood), *MAPT* ($P_{TWAS} = 9.60 \times 10^{-15}$ for lung tissue), and *DEPTOR* ($P_{TWAS} = 8.58 \times 10^{-9}$ for lung tissue). The top 30 genes identified by TWAS are summarized in **Table 1**.

Gene Set Enrichment Analysis

A total of 1,147 genes with a TWAS P value < 0.05 were included in the GO enrichment analysis. Metascape detected 76 GO terms under P value < 0.01 (**Figure 2** and **Supplementary Table S2**), such as antigen processing and presentation of peptide or polysaccharide antigen via major histocompatibility complex (MHC) class II ($P_{TWAS} = 1.43 \times 10^{-8}$), vacuolar part ($P_{TWAS} = 5.98 \times 10^{-6}$), ncRNA metabolic process ($P_{TWAS} = 1.61 \times 10^{-5}$), snRNA transcription by RNA polymerase II ($P_{TWAS} = 3.03 \times 10^{-5}$), and SWI/SNF complex ($P_{TWAS} = 5.81 \times 10^{-5}$). For KEGG pathway enrichment analysis of the genes identified by TWAS, Metascape detected 30 candidate pathways for IPF under P value < 0.01 (**Figure 3** and **Supplementary Table S3**), such as *Staphylococcus aureus* infection ($P_{TWAS} = 3.69 \times 10^{-7}$), allograft rejection ($P_{TWAS} = 4.45 \times 10^{-6}$), asthma ($P_{TWAS} = 7.65 \times 10^{-6}$), type I diabetes mellitus ($P_{TWAS} = 1.32 \times 10^{-5}$), inflammatory bowel disease (IBD) ($P_{TWAS} = 7.37 \times 10^{-5}$), and herpes simplex infection ($P_{TWAS} = 7.93 \times 10^{-5}$).

TABLE 1 | Top 30 genes identified by TWAS for IPF.

Gene	Chromosome	P_{TWAS}	Tissue
DSP	6	1.35×10^{-29}	Lung
MUC5B	11	1.09×10^{-28}	Lung
TOLLIP	11	1.41×10^{-15}	Whole blood
DND1P1	17	8.12×10^{-15}	Lung
CRHR1-IT1	17	9.20×10^{-15}	Lung
MAPT	17	9.60×10^{-15}	Lung
RP11-259G18.2	17	1.04×10^{-14}	Lung
RP11-707O23.5	17	1.19×10^{-14}	Lung
RP11-259G18.3	17	1.22×10^{-14}	Lung
LRRC37A4P	17	1.59×10^{-14}	Lung
KANSL1-AS1	17	2.75×10^{-14}	Lung
RP11-259G18.1	17	3.81×10^{-14}	Lung
KIAA1267	17	2.19×10^{-13}	Peripheral blood
LRRC37A2	17	2.91×10^{-13}	Lung
WNT3	17	1.21×10^{-12}	Lung
DND1	17	2.17×10^{-12}	Peripheral blood
PLEKHM1	17	5.00×10^{-11}	Peripheral blood
FAM215B	17	7.66×10^{-11}	Lung
PLEKHM1	17	3.08×10^{-10}	Lung
RP11-158M2.5	15	1.41×10^{-9}	Lung
FAM13A	4	2.12×10^{-9}	Lung
LRRC37A	17	6.40×10^{-9}	Lung
DEPTOR	8	8.58×10^{-9}	Lung
RP11-760H22.2	8	1.06×10^{-8}	Lung
BAHD1	15	1.16×10^{-8}	Lung
BRSK2	11	1.49×10^{-8}	Lung
RP11-798G7.5	17	2.97×10^{-8}	Lung
GCHFR	15	1.94×10^{-7}	Whole blood
RP11-64K12.8	15	2.37×10^{-7}	Lung
ZNF514	2	4.41×10^{-6}	Whole blood

A total of 2,204 DEGs were identified by mRNA expression profiling analysis of IPF (**Supplementary Table S4**), and were then conducted GO and KEGG pathway enrichment analysis (**Supplementary Tables S5**, **S6**). 38 common GO terms were identified by enrichment analysis of IPF-related genes identified by TWAS and DEGs (**Supplementary Table S7**), including antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (GO:0002504, $P_{TWAS} = 1.43 \times 10^{-8}$, $P_{mRNA} = 8.87 \times 10^{-3}$), lytic vacuole (GO:0000323, $P_{TWAS} = 6.25 \times 10^{-6}$, $P_{mRNA} = 5.91 \times 10^{-3}$), ncRNA metabolic process (GO:0034660, $P_{TWAS} = 1.61 \times 10^{-5}$, $P_{mRNA} = 4.43 \times 10^{-7}$), microtubule organizing center (GO:0005815, $P_{TWAS} = 1.24 \times 10^{-4}$, $P_{mRNA} = 2.46 \times 10^{-23}$), and autophagy (GO:0002504, $P_{TWAS} = 1.10 \times 10^{-3}$, $P_{mRNA} = 2.11 \times 10^{-4}$). **Table 2** summarizes the top 20 common GO terms detected by enrichment analysis of IPF-related genes identified by TWAS and DEGs. We also detected 8 common KEGG pathways (**Table 3**), such as staphylococcus aureus infection ($P_{TWAS} = 3.69 \times 10^{-7}$, $P_{mRNA} = 9.17 \times 10^{-3}$), herpes simplex infection ($P_{TWAS} = 7.93 \times 10^{-5}$, $P_{mRNA} = 8.76 \times 10^{-4}$), HTLV-I infection ($P_{TWAS} = 8.84 \times 10^{-5}$, $P_{mRNA} = 5.92 \times 10^{-4}$), phagosome ($P_{TWAS} = 8.99 \times 10^{-5}$, $P_{mRNA} = 1.93 \times 10^{-4}$), and systemic lupus erythematosus ($P_{TWAS} = 2.25 \times 10^{-3}$, $P_{mRNA} = 3.12 \times 10^{-3}$).

PPI Network, Hub Gene and Module Analysis

To evaluate the association of IPF-related genes identified by TWAS, a PPI network was constructed by STRING and visualized by Cytoscape, containing 329 nodes and 893 edges (**Supplementary Figure S1**). The top 20 hub genes were identified by cytoHubba plugin that uses 12 different algorithms (**Supplementary Table S8**). Then, from the genes that can be detected by more than five algorithms, 16 hub genes with the highest degree of connectivity were selected to build the hub gene PPI network (**Figure 4A**). The enrichment analysis showed that the IPF-related processes hub genes were enriched in antigen processing and presentation of exogenous peptide antigen via MHC class II, chromosome, centromeric region, and cytoplasmic dynein complex.

In addition, module analysis conducted by MCODE plugin in Cytoscape identified several modules in the PPI network. Then, the top four significant modules were selected for subsequent analysis. A significant module, which gained the highest MCODE

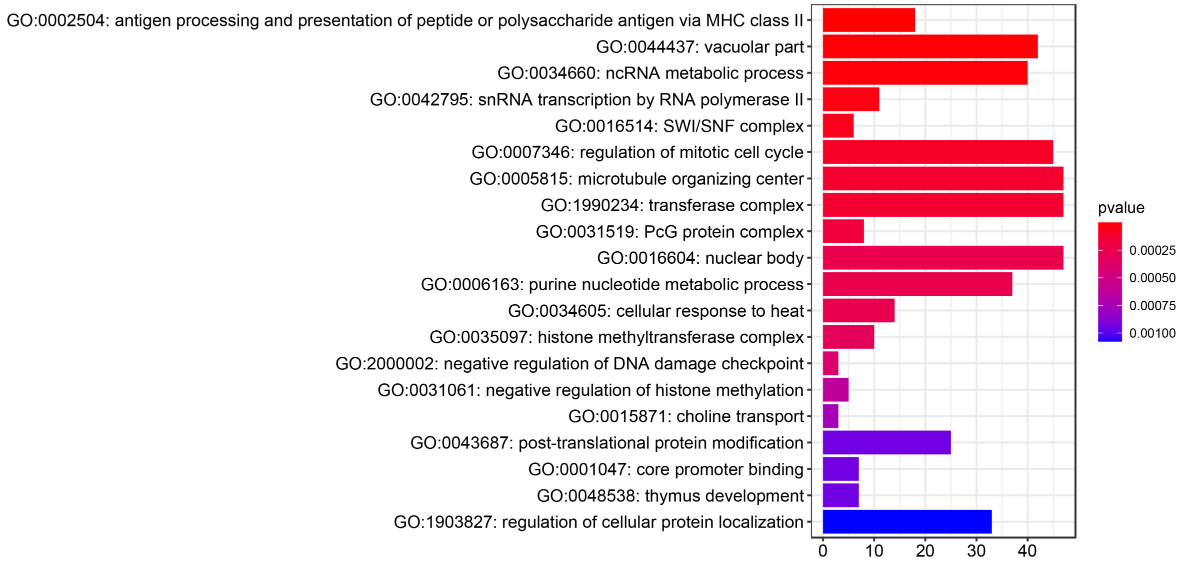

FIGURE 2 | The top 20 gene ontology terms identified by enrichment analysis for IPF-related genes from TWAS.

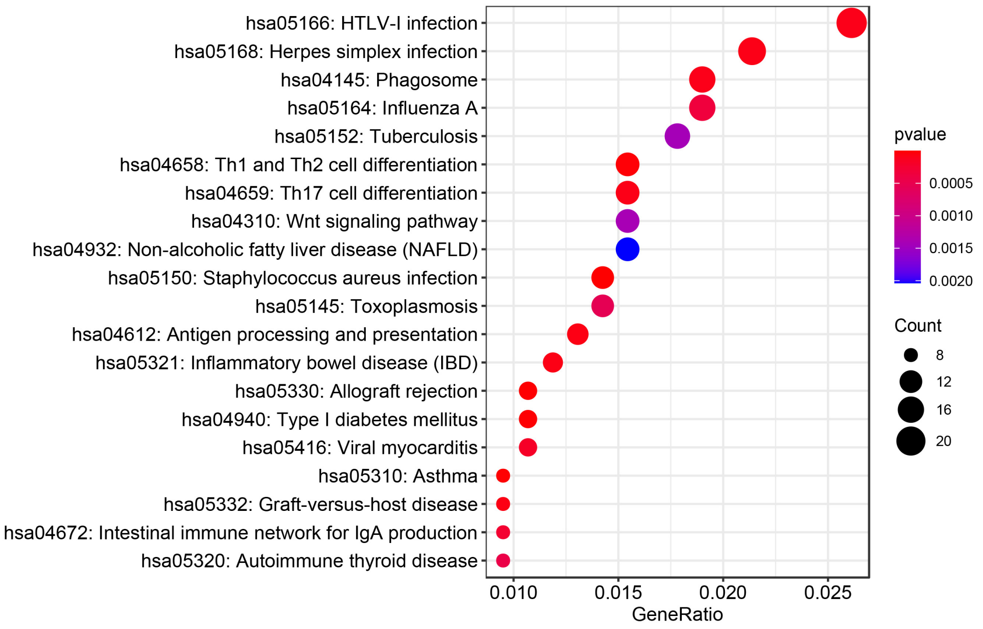

FIGURE 3 | The top 20 KEGG pathways identified by enrichment analysis for IPF-related genes from TWAS.

score, contained 24 nodes and 150 edges. Subsequent functional enrichment analysis indicated that the genes in this module were primarily enriched in antigen processing and presentation of exogenous peptide antigen via MHC class II, chromosome, centromeric region, cytoplasmic dynein complex, and type I interferon signaling pathway (**Figure 4B**). Fourteen genes (*DYNLL1*, *DYNC1LI1*, *DYNLL2*, *HLA-DRB5*, *HLA-DPB1*, *HLA-DQB2*, *HLA-DQA2*, *HLA-DQB1*, *HLA-DRB1*, *POLR2L*, *CENPP*, *CENPK*, *NUP133*, and *NUP107*) were simultaneously detected by both hub gene and module analysis.

DISCUSSION

Although the biological basis of IPF has been investigated in the past years, the cellular and molecular mechanisms of IPF are very complicated and remain unclear. In the present study, we performed the first large-scale integrative analysis of TWAS and mRNA expression profiles for IPF, which successfully detected some plausible genes as well as pathways, and can potentially provide novel insights to better understand the molecular mechanisms underlying the development of IPF.

TABLE 2 | Top 20 overlapped gene ontology terms identified by enrichment analysis for IPF-related genes from TWAS and for differentially expressed genes from mRNA expression profiling of IPF.

ID	Category	Description	P_{TWAS}	P_{mRNA}
GO:0002504	BP	Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	1.43×10^{-8}	8.87×10^{-3}
GO:0000323	CC	Lytic vacuole	6.25×10^{-6}	5.91×10^{-3}
GO:0034660	BP	ncRNA metabolic process	1.61×10^{-5}	4.43×10^{-7}
GO:0042795	BP	snRNA transcription by RNA polymerase II	3.03×10^{-5}	8.88×10^{-3}
GO:0006338	BP	Chromatin remodeling	2.47×10^{-4}	1.37×10^{-5}
GO:0007346	BP	Regulation of mitotic cell cycle	9.76×10^{-5}	1.43×10^{-12}
GO:0005815	CC	Microtubule organizing center	1.24×10^{-4}	2.46×10^{-23}
GO:1990234	CC	Transferase complex	1.32×10^{-4}	9.74×10^{-6}
GO:0016604	CC	Nuclear body	2.50×10^{-4}	2.37×10^{-16}
GO:0006163	BP	Purine nucleotide metabolic process	2.56×10^{-4}	2.17×10^{-6}
GO:0034605	BP	Cellular response to heat	2.68×10^{-4}	4.12×10^{-3}
GO:0044417	BP	Translocation of molecules into host	2.95×10^{-3}	5.04×10^{-4}
GO:0043687	BP	Post-translational protein modification	9.48×10^{-4}	8.44×10^{-6}
GO:1903827	BP	Regulation of cellular protein localization	1.08×10^{-3}	2.58×10^{-6}
GO:0006914	BP	Autophagy	1.10×10^{-3}	2.11×10^{-4}
GO:0098687	CC	Chromosomal region	1.31×10^{-3}	7.26×10^{-15}
GO:0072594	BP	Establishment of protein localization to organelle	1.55×10^{-3}	6.00×10^{-5}
GO:0008134	MF	Transcription factor binding	1.55×10^{-3}	4.44×10^{-4}
GO:0032984	BP	Protein-containing complex disassembly	3.84×10^{-3}	7.07×10^{-3}
GO:0000139	CC	Golgi membrane	1.87×10^{-3}	7.40×10^{-6}

Biological Processes (BP); Cellular Components (CC); Molecular Functions (MF).

TABLE 3 | Overlapped KEGG pathways identified by enrichment analysis for IPF-related genes from TWAS and for differentially expressed genes from mRNA expression profiling of IPF.

ID	Description	P_{TWAS}	P_{mRNA}
hsa05150	Staphylococcus aureus infection	3.69×10^{-7}	9.17×10^{-3}
hsa05168	Herpes simplex infection	7.93×10^{-5}	8.76×10^{-4}
hsa05166	HTLV-I infection	8.84×10^{-5}	5.92×10^{-4}
hsa04145	Phagosome	8.99×10^{-5}	1.93×10^{-4}
hsa05164	Influenza A	3.46×10^{-4}	1.32×10^{-3}
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	2.04×10^{-3}	2.64×10^{-4}
hsa05322	Systemic lupus erythematosus	2.25×10^{-3}	3.12×10^{-3}
hsa04620	Toll-like receptor signaling pathway	9.97×10^{-3}	9.76×10^{-3}

Transcriptome-wide association study detected several significant IPF-related genes previously reported in GWAS, such as *DSP* and *MUC5B*. *DSP* encodes desmoplakin, which is an important component of the desmosome structure. Desmoplakin is involved in the mechanical linkage of cells, stabilization of tissue structure and the process of cell migration, proliferation, and differentiation. Accordingly, *DSP* is essential to cell-cell adhesion and epithelial barrier function (Vasioukhin et al., 2001). Previous evidence has suggested that *DSP* expression is higher in IPF lung than in the lungs of healthy control subjects and the intron 5 variant rs2076295 was found to be associated with decreased *DSP* expression, indicating that differential *DSP* expression plays an important role in IPF etiology (Mathai et al., 2016). *MUC5B* encodes for mucin 5B, which is produced by airway epithelial cells and is a major gel-forming mucin in the mucus. Mucin 5B is involved in the production of airway mucous and may have a significant role in mucociliary clearance and airway defense (Roy et al., 2014; Evans et al., 2016). Increased *MUC5B* expression might impair mucosal defense of host and thus lead to the reduction of lung clearance of inhaled particles, dissolved chemicals, and microorganisms (Seibold et al., 2011). The *MUC5B* promotor variant rs35705950 is a common variant that accounts for a large proportion of risk for the development of familial interstitial pneumonia and IPF (Seibold et al., 2011; Todd et al., 2015; Evans et al., 2016; Zhang et al., 2019). Interestingly, a retrospective study has demonstrated improved survival of patients with this promoter variant compared with those without this variant, indicating that this variant might be a potential prognostic indicator (Peljto et al., 2013). These paradoxical findings imply that further investigation is required to clarify the biological mechanism by which this promoter variant promotes the development of IPF.

Besides, attention should be paid to genes simultaneously detected by hub gene and module analysis, such as *DYNC1LI1*,

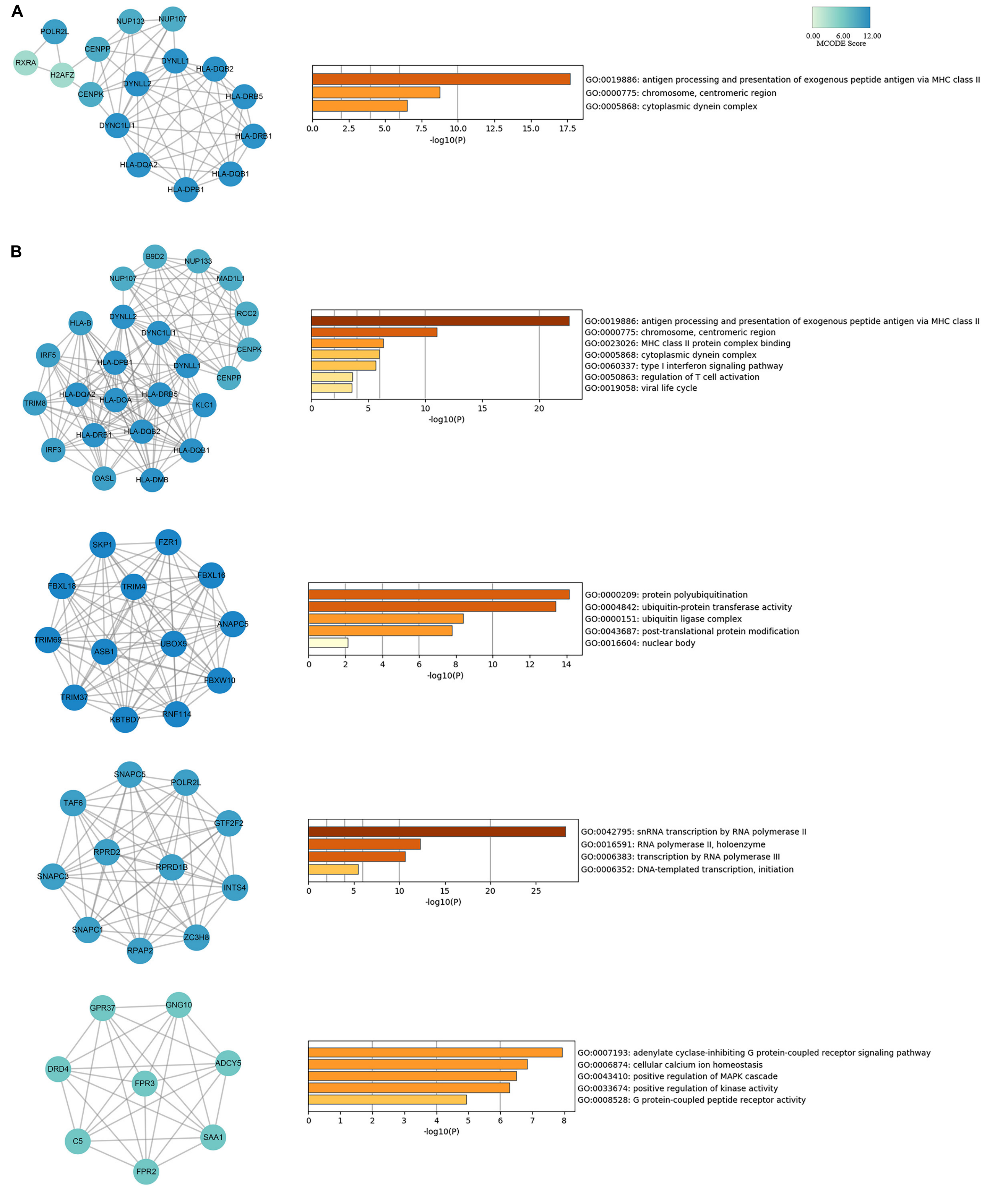

FIGURE 4 | PPI network for hub genes and modules analyses of IPF-related genes identified by TWAS. **(A)** The network of 16 hub genes with a higher degree of connectivity and enrichment analysis of these genes. **(B)** Genes of top four modules were subjected to GO and KEGG enrichment analysis by Metascape.

DYNLL1, and *DYNLL2*, which are likely to be related with IPF but not reported earlier. *DYNLL1* and *DYNLL2* are related to cell cycle spindle assembly and chromosome separation and may be involved in the change or maintenance of the spatial distribution of cytoskeletal structures (Dunsch et al., 2012). *DYNC1LI1* is related to microtubule motor activity and may play a role in binding dynein to membranous organelles. These three genes belong to the cytoplasmic dynein subunit gene. Cytoplasmic dynein acts as a motor for the intracellular retrograde motility of vesicles and organelles along microtubules (Dunsch et al., 2012). Besides, human airway epithelium is characterized by the presence of ciliated cells bearing motile cilia, and specialized cell surface projections containing axonemes consisted of microtubules and dynein arms, providing ATP-driven motility (Tilley et al., 2015). In the airways, cilia function together with airway mucus, plays an important role in mediating mucociliary clearance and eliminating the inhaled particles and pathogens (Evans et al., 2016). Cilia dysfunction and clearance impairment would result in chronic airway inflammation and infection, bronchiectasis, and distal lung remodeling. Besides, the hub gene PPI network showed that the genes involved are mainly enriched in cytoplasmic dynein complex and antigen processing and presentation, suggesting the critical role of cilia function and immune response in the development of IPF.

Gene ontology and KEGG pathway enrichment analysis detected several candidate biological pathways for IPF, mainly involved in immune inflammation response and infection. For instance, antigen processing and presentation of peptide or polysaccharide antigen via MHC class II. Human leukocyte antigen (HLA), encoded by the human MHC gene complex, plays a critical role in the antigen presentation of peptides and the regulation of immune response (Bodis et al., 2018). A GWAS analysis has identified two risk alleles in the HLA region (DRB1*15:01 and DQB1*06:02) that were related to IPF (Fingerlin et al., 2016). Besides, several studies have suggested the role of HLA region in the development of IPF (Falfan-Valencia et al., 2005; Aquino-Galvez et al., 2009; Xue et al., 2011; Zhang et al., 2012, 2015). The association between HLA and IPF may suggest the potential etiologic role of autoimmunity in IPF. Recently, a nationwide retrospective cohort study in Korea involving 38,921 IBD patients and 116,763 patients without IBD suggested that patients with IBD, especially Crohn's disease, have an increasing risk for the development of IPF (Kim et al., 2020). In addition, a 1:1 retrospective case-control study (196 IPF cases and 196 controls) has indicated that hypothyroidism, an immune-mediated process, was common among IPF patients and was found to be associated with decreased survival time as an independent predictor of mortality in IPF patients (Oldham et al., 2015). Besides, diabetes has been reported to be a risk factor of IPF (Enomoto et al., 2003; Gribbin et al., 2009). Type 1 diabetes, also known as insulin-dependent diabetes, is an organ-specific autoimmune disease. Gene variants in the HLA region have been found to be related to the susceptibility of type 1 diabetes mellitus (Nejentsev et al., 2007). Naturally, these autoimmune diseases may share genetic basis contributed by genetic variations in HLA region. Since, observational associations are prone to reverse causality and confounding, further investigation is warranted to characterize the pathophysiologic link between this genetic variation and disease.

Another primary candidate biological pathway was shown to be infections (both viral and bacterial), which is also closely related to the antigen stimulation and immune response, such as herpes simplex and Staphylococcus aureus infection. Previous studies have demonstrated that virus may be involved in disease initiation. And the presence of herpes viral DNA and epithelial cell stress in the lungs of asymptomatic relatives are at risk for the development of familial IPF (Moore and Moore, 2015). A recent meta-analysis of 20 case-control studies with 1,287 participants (634 IPF cases and 653 controls) has reported that the existence of persistent or chronic viral infections significantly associate with the increasing risk of the development of IPF, but not with the aggravation of IPF (Sheng et al., 2020). Previous studies in mice models have reported that viral infection could promote the formation of lung fibrosis (Mora et al., 2006, 2007; Qiao et al., 2009). Especially, animal experiments have been applied to provide evidence of the pathogenesis of lung fibrosis regulated by gamma herpesvirus (Moore and Moore, 2015). These animal experiments have also indicated that previous infections seem to make lung epithelial cells reprogrammed during the incubation period, producing profibrotic factors, leading to the enhanced susceptibility to subsequent fibrosis damage in lung. Nevertheless, infections in susceptible hosts or the exacerbation of existing fibrosis involve active viral replication and are affected by antiviral therapy (Moore and Moore, 2015). In addition, activated leukocyte signals in IPF patients provide further support for infectious processes driving the progression of IPF. Studies have also reported that bacterial infections play a role in the progression and prognosis of IPF. A microbiome analysis of IPF bronchoscopic alveolar lavage (BAL) samples suggested that the increase in relative abundance of two operational taxonomic units (*Streptococcus* OTU1345 and *Staphylococcus* OTU1348) was positively correlated with the progression of IPF (Han et al., 2014). In another study, reduced diversity of the lung microbiome has been found to be associated with low forced vital capacity and early mortality in patients with IPF, and a mouse model demonstrated that bleomycin-induced lung fibrosis led to a decrease in the diversity and modification of microbiota (Takahashi et al., 2018). In addition, compared with the control group, bacterial load in BAL of IPF patients has been shown to be greater. The rate of decline in lung function and the mortality risk can be partly predicted by the baseline bacterial load (Molyneaux et al., 2014). *Haemophilus*, *Streptococcus*, *Neisseria*, and *Veillonella* have found to be more abundant in cases in comparison with controls. However, animal modeling implicated that infection of *Pseudomonas aeruginosa* did not aggravate bleomycin-induced fibrosis (Ashley et al., 2014), suggesting that there might be some microbial specificity in the progression of lung fibrosis or the bleomycin-induced mouse model cannot accurately reflect the alterations of the IPF disease course induced by bacterial infection in humans. To summarize, both viral and bacterial infections may play a crucial role in the progression of IPF and may be potential predictors of disease prognosis. Additional work will be warranted to investigate the biological mechanism of infections in the progression of IPF and further explore the potential benefit of antiviral and antimicrobial therapy.

There are several limitations in our study. First, the number of genes that can be accurately imputed in the TWAS analysis is limited by the training cohort sample size, the majority of subjects were of European ancestry, and the results cannot be directly generalized to other ethnic population. Second, there may be tissue bias using lung, peripheral blood and whole blood expression reference panels. Cell-type heterogeneity within or between tissues and cross-tissue pleiotropy may introduce tissue bias. Further investigation can be implemented to address this issue if reference panels for individual cell types or states are available (Wainberg et al., 2019). Third, TWAS significant genes cannot guarantee causality, since co-regulation may lead non-causal hits. Some fine-mapping methods such as FOCUS (fine-mapping of causal gene sets) may partly address this issue, due to its ability to directly model predicted expression correlations and use them to assign genes posterior probabilities of causality (Mancuso et al., 2019). FOCUS, as a post-TWAS analysis method, can be applied on top of the genes identified by TWAS to further reduce false discoveries.

CONCLUSION

In conclusion, we conducted a large-scale integrative analysis of TWAS and mRNA expression profiles for IPF. Our results provide novel insights into a better understanding of the genetic mechanism of IPF. Further functional biology studies are warranted to validate our findings and clarify the potential roles of identified genes and pathways in the development of IPF.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Full summary statistics for the genome-wide meta-analysis of IPF can be accessed from https://github.com/genomicsITER/PFgenetics. We would like to thank the Collaborative Group of genetic studies of IPF for providing us with the IPF GWAS summary data. The gene expression profile dataset of IPF analyzed for this study can be found in the GEO with accession number GSE110147 (https://www.ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

ZY conceived and designed the study. WG and PG performed the statistical analysis. PG wrote the manuscript. LL, QG, and ZY provided feasible advice on data analysis and drafting manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been supported by grants from National Natural Science Foundation of China (81673272 and 81872712), the Natural Science Foundation of Shandong Province (ZR2019ZD02), the Key Technology Research and Development Program of Shandong Province (2017CXGC1214), and the Young Scholars Program of Shandong University (2016WLJH23).

ACKNOWLEDGMENTS

We would like to thank Collaborative Group of genetic studies of IPF for providing us with the IPF GWAS summary data and Gene Expression Omnibus database for providing the IPF gene expression profile data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.604324/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

Allen, R. J., Guillen-Guio, B., Oldham, J. M., Ma, S. F., Dressen, A., Paynton, M. L., et al. (2020). Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 201, 564–574. doi: 10.1164/rccm.201905-1017OC

Allen, R. J., Porte, J., Braybrooke, R., Flores, C., Fingerlin, T. E., Oldham, J. M., et al. (2017). Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir. Med.* 5, 869–880. doi: 10.1016/S2213-2600(17)30387-9

Aquino-Galvez, A., Perez-Rodriguez, M., Camarena, A., Falfan-Valencia, R., Ruiz, V., Montano, M., et al. (2009). MICA polymorphisms and decreased expression of the MICA receptor NKG2D contribute to idiopathic pulmonary fibrosis susceptibility. *Hum. Genet.* 125, 639–648. doi: 10.1007/s00439-009-0666-1

Armanios, M. Y., Chen, J. J., Cogan, J. D., Alder, J. K., Ingersoll, R. G., Markin, C., et al. (2007). Telomerase mutations in families with idiopathic pulmonary fibrosis. *N. Engl. J. Med/* 356, 1317–1326. doi: 10.1056/NEJMoa066157

Ashley, S. L., Jegal, Y., Moore, T. A., van Dyk, L. F., Laouar, Y., and Moore, B. B. (2014). gamma-Herpes virus-68, but not *Pseudomonas aeruginosa* or influenza A (H1N1), exacerbates established murine lung fibrosis. *Am. J. Physiol. Lung Cell. Mol. Physio.* 307, L219–L230. doi: 10.1152/ajplung.00300.2013

Baumgartner, K. B., Samet, J. M., Coultas, D. B., Stidley, C. A., Hunt, W. C., Colby, T. V., et al. (2000). Occupational and environmental risk factors for idiopathic pulmonary fibrosis: a multicenter case-control study. Collaborating Centers. *Am. J. Epidemiol.* 152, 307–315. doi: 10.1093/aje/152.4.307

Bedard Methot, D., Leblanc, E., and Lacasse, Y. (2019). Meta-analysis of gastroesophageal reflux disease and idiopathic pulmonary fibrosis. *Chest* 155, 33–43. doi: 10.1016/j.chest.2018.07.038

Bodis, G., Toth, V., and Schwarting, A. (2018). Role of human leukocyte antigens (HLA) in autoimmune diseases. *Rheumatol. Ther.* 5, 5–20. doi: 10.1007/s40744-018-0100-z

Boomsma, D. I., de Geus, E. J., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. J., et al. (2006). Netherlands twin register: from twins to twin families. *Twin. Res. Hum. Genet.* 9, 849–857. doi: 10.1375/183242706779462426

Cecchini, M. J., Hosein, K., Howlett, C. J., Joseph, M., and Mura, M. (2018). Comprehensive gene expression profiling identifies distinct and overlapping transcriptional profiles in non-specific interstitial pneumonia and idiopathic pulmonary fibrosis. *Respir Res.* 19:153. doi: 10.1186/s12931-018-0857-1

Cogan, J. D., Kropski, J. A., Zhao, M., Mitchell, D. B., Rives, L., Markin, C., et al. (2015). Rare variants in RTEL1 are associated with familial interstitial pneumonia. *Am. J. Respir. Crit. Care Med.* 191, 646–655. doi: 10.1164/rccm.201408-1510OC

Dunsch, A. K., Hammond, D., Lloyd, J., Schermelleh, L., Gruneberg, U., and Barr, F. A. (2012). Dynein light chain 1 and a spindle-associated adaptor promote dynein asymmetry and spindle orientation. *J. Cell Biol.* 198, 1039–1054. doi: 10.1083/jcb.201202112

Enomoto, T., Usuki, J., Azuma, A., Nakagawa, T., and Kudoh, S. (2003). Diabetes mellitus may increase risk for idiopathic pulmonary fibrosis. *Chest* 123, 2007–2011. doi: 10.1378/chest.123.6.2007

Evans, C. M., Fingerlin, T. E., Schwarz, M. I., Lynch, D., Kurche, J., Warg, L., et al. (2016). Idiopathic pulmonary fibrosis: a genetic disease that involves mucociliary dysfunction of the peripheral airways. *Physiol. Rev.* 96, 1567–1591. doi: 10.1152/physrev.00004.2016

Falfan-Valencia, R., Camarena, A., Juarez, A., Becerril, C., Montano, M., Cisneros, J., et al. (2005). Major histocompatibility complex and alveolar epithelial apoptosis in idiopathic pulmonary fibrosis. *Hum. Genet.* 118, 235–244. doi: 10.1007/s00439-005-0035-7

Fingerlin, T. E., Murphy, E., Zhang, W., Peljto, A. L., Brown, K. K., Steele, M. P., et al. (2013). Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.* 45, 613–620. doi: 10.1038/ng.2609

Fingerlin, T. E., Zhang, W., Yang, I. V., Ainsworth, H. C., Russell, P. H., Blumhagen, R. Z., et al. (2016). Genome-wide imputation study identifies novel HLA locus for pulmonary fibrosis and potential role for auto-immunity in fibrotic idiopathic interstitial pneumonia. *BMC Genet.* 17:74. doi: 10.1186/s12863-016-0377-2

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, 1091–1098. doi: 10.1038/ng.3367

Gribbin, J., Hubbard, R., and Smith, C. (2009). Role of diabetes mellitus and gastro-oesophageal reflux in the aetiology of idiopathic pulmonary fibrosis. *Respir. Med.* 103, 927–931. doi: 10.1016/j.rmed.2008.11.001

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (Ldacc)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups;, Nih Common Fund, NIH/NCI; NIH/NHGRI et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

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, 245–252. doi: 10.1038/ng.3506

Han, M. K., Zhou, Y., Murray, S., Tayob, N., Noth, I., Lama, V. N., et al. (2014). Lung microbiome and disease progression in idiopathic pulmonary fibrosis: an analysis of the COMET study. *Lancet Respir. Med.* 2, 548–556. doi: 10.1016/S2213-2600(14)70069-4

Hutchinson, J., Fogarty, A., Hubbard, R., and McKeever, T. (2015). Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review. *Eur. Respir. J.* 46, 795–806. doi: 10.1183/09031936.00185114

Kim, J., Chun, J., Lee, C., Han, K., Choi, S., Lee, J., et al. (2020). Increased risk of idiopathic pulmonary fibrosis in inflammatory bowel disease: a nationwide study. *J. Gastroenterol. Hepatol.* 35, 249–255. doi: 10.1111/jgh.14838

Kim, J. S., Podolanczuk, A. J., Borker, P., Kawut, S. M., Raghu, G., Kaufman, J. D., et al. (2017). Obstructive sleep apnea and subclinical interstitial lung disease in the multi-ethnic study of atherosclerosis (MESA). *Ann. Am. Thorac. Soc.* 14, 1786–1795. doi: 10.1513/AnnalsATS.201701-091OC

Lederer, D. J., and Martinez, F. J. (2018). Idiopathic pulmonary fibrosis. *N. Engl. J. Med.* 378, 1811–1823. doi: 10.1056/NEJMra1705751

Ley, B., Collard, H. R., and King, T. E. Jr. (2011). Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 183, 431–440. doi: 10.1164/rccm.201006-0894CI

Liu, Y., Devescovi, V., Chen, S., and Nardini, C. (2013). Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Syst. Biol.* 7:14. doi: 10.1186/1752-05 09-7-14

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., et al. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* 51, 675–682. doi: 10.1038/s41588-019-0367-1

Mathai, S. K., Pedersen, B. S., Smith, K., Russell, P., Schwarz, M. I., Brown, K. K., et al. (2016). Desmoplakin variants are associated with idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 193, 1151–1160. doi: 10.1164/rccm.201509-1863OC

Molyneaux, P. L., Cox, M. J., Willis-Owen, S. A., Mallia, P., Russell, K. E., Russell, A. M., et al. (2014). The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 190, 906–913. doi: 10.1164/rccm.201403-0541OC

Moore, B. B., and Moore, T. A. (2015). Viruses in idiopathic pulmonary fibrosis. Etiology and Exacerbation. *Ann. Am. Thorac. Soc.* 12(Suppl. 2), S186–S192. doi: 10.1513/AnnalsATS.201502-088AW

Mora, A. L., Torres-Gonzalez, E., Rojas, M., Corredor, C., Ritzenthaler, J., Xu, J. G., et al. (2006). Activation of alveolar macrophages via the alternative pathway in herpesvirus-induced lung fibrosis. *Am. J. Respir. Cell Mol. Biol.* 35, 466–473. doi: 10.1165/rcmb.2006-0121OC

Mora, A. L., Torres-Gonzalez, E., Rojas, M., Xu, J., Ritzenthaler, J., Speck, S. H., et al. (2007). Control of virus reactivation arrests pulmonary herpesvirus-induced fibrosis in IFN-gamma receptor-deficient mice. *Am. J. Respir. Crit. Care Med.* 175, 1139–1150. doi: 10.1164/rccm.200610-1426OC

Mushiroda, T., Wattanapokayakit, S., Takahashi, A., Nukiwa, T., Kudoh, S., Ogura, T., et al. (2008). A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. *J. Med. Genet.* 45, 654–656. doi: 10.1136/jmg.2008.057356

Nejentsev, S., Howson, J. M., Walker, N. M., Szeszko, J., Field, S. F., Stevens, H. E., et al. (2007). Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* 450, 887–892. doi: 10.1038/nature 06406

Nogee, L. M., Dunbar, A. E. III, Wert, S. E., Askin, F., Hamvas, A., and Whitsett, J. A. (2001). A mutation in the surfactant protein C gene associated with familial interstitial lung disease. *N. Engl. J, Med.* 344, 573–579. doi: 10.1056/NEJM200102223440805

Noth, I., Zhang, Y., Ma, S. F., Flores, C., Barber, M., Huang, Y., et al. (2013). Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respir. Med.* 1, 309–317. doi: 10.1016/S2213-2600(13)70045-6

Oldham, J. M., Kumar, D., Lee, C., Patel, S. B., Takahashi-Manns, S., Demchuk, C., et al. (2015). Thyroid disease is prevalent and predicts survival in patients with idiopathic pulmonary fibrosis. *Chest* 148, 692–700. doi: 10.1378/chest.14-2714

Peljto, A. L., Zhang, Y. Z., Fingerlin, T. E., Ma, S. F., Garcia, J. G. N., Richards, T. J., et al. (2013). Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *Jama J. Am. Med. Assoc.* 309, 2232–2239. doi: 10.1001/jama.2013.5827

Qiao, J., Zhang, M., Bi, J., Wang, X., Deng, G., He, G., et al. (2009). Pulmonary fibrosis induced by H5N1 viral infection in mice. *Respir. Res.* 10:107. doi: 10.1186/1465-9921-10-107

Raghu, G., Collard, H. R., Egan, J. J., Martinez, F. J., Behr, J., Brown, K. K., et al. (2011). An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am. J. Respir. Crit. Care Med.* 183, 788–824. doi: 10.1164/rccm.2009-040GL

Raitakari, O. T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., et al. (2008). Cohort profile: the cardiovascular risk in young finns study. *Int. J. Epidemiol.* 37, 1220–1226. doi: 10.1093/ije/dym225

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler. GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* 14, 482–517. doi: 10.1038/s41596-018-0103-9

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Re.* 43:e47. doi: 10.1093/nar/gkv007

Roy, M. G., Livraghi-Butrico, A., Fletcher, A. A., McElwee, M. M., Evans, S. E., Boerner, R. M., et al. (2014). Muc5b is required for airway defence. *Nature* 505, 412–416. doi: 10.1038/nature12807

Sack, C., Vedal, S., Sheppard, L., Raghu, G., Barr, R. G., Podolanczuk, A., et al. (2017). Air pollution and subclinical interstitial lung disease: the multi-ethnic study of atherosclerosis (MESA) air-lung study. *Eur. Respir. J.* 50:1700559. doi: 10.1183/13993003.00559-2017

Seibold, M. A., Wise, A. L., Speer, M. C., Steele, M. P., Brown, K. K., Loyd, J. E., et al. (2011). A common MUC5B promoter polymorphism and pulmonary fibrosis. *N. Engl. J. Med.* 364, 1503–1512.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sheng, G., Chen, P., Wei, Y., Yue, H., Chu, J., Zhao, J., et al. (2020). viral infection increases the risk of idiopathic pulmonary fibrosis: a meta-analysis. *Chest* 157, 1175–1187. doi: 10.1016/j.chest.2019.10.032

Stuart, B. D., Choi, J., Zaidi, S., Xing, C., Holohan, B., Chen, R., et al. (2015). Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat. Genet.* 47, 512–517. doi: 10.1038/ng.3278

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Takahashi, Y., Saito, A., Chiba, H., Kuronuma, K., Ikeda, K., Kobayashi, T., et al. (2018). Impaired diversity of the lung microbiome predicts progression of idiopathic pulmonary fibrosis. *Respi. Res.* 19:34. doi: 10.1186/s12931-018-0736-9

Tang, Y. W., Johnson, J. E., Browning, P. J., Cruz-Gervis, R. A., Davis, A., Graham, B. S., et al. (2003). Herpesvirus DNA is consistently detected in lungs of patients with idiopathic pulmonary fibrosis. *J. Clin. Microbiol.* 41, 2633–2640. doi: 10.1128/Jcm.41.6.2633-2640.2003

Tilley, A. E., Walters, M. S., Shaykhiev, R., and Crystal, R. G. (2015). Cilia dysfunction in lung disease. *Annu. Rev. Physiol.* 77, 379–406. doi: 10.1146/annurev-physiol-021014-071931

Tobin, R. W., Pope, C. E. II, Pellegrini, C. A., Emond, M. J., Sillery, J., and Raghu, G. (1998). Increased prevalence of gastroesophageal reflux in patients with idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 158, 1804–1808. doi: 10.1164/ajrccm.158.6.9804105

Todd, N. W., Atamas, S. P., Luzina, I. G., and Galvin, J. R. (2015). Permanent alveolar collapse is the predominant mechanism in idiopathic pulmonary fibrosis. *Expert Rev. Respir. Med.* 9, 411–418. doi: 10.1586/17476348.2015.1067609

Vasioukhin, V., Bowers, E., Bauer, C., Degenstein, L., and Fuchs, E. (2001). Desmoplakin is essential in epidermal sheet formation. *Nat. Cell Biol.* 3, 1076–1085. doi: 10.1038/ncb1201-1076

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* 51, 592–599. doi: 10.1038/s41588-019-0385-z

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., et al. (2014). Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* 46, 430–437. doi: 10.1038/ng.2951

Xue, J., Gochuico, B. R., Feghali-Bostwick, C. A., Noth, I., Nathan, S. D., Rosen, G., et al. (2011). The HLA Class II Allele DRB1*1501 is over-represented in patients with idiopathic pulmonary fibrosis. *PLoS One* 6:e14715. doi: 10.1371/journal.pone.0014715

Yu, X.-T., and Zeng, T. (2018). "Integrative analysis of omics big data," in *Computational Systems Biology: Methods and Protocols*, ed. T. Huang (New York, NY: Springer New York), 109–135. doi: 10.1007/978-1-4939-7717-8_7

Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., et al. (2020). Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat. Commun.* 11:3861. doi: 10.1038/s41467-020-17668-6

Zhang, H. P., Zou, J., Xie, P., Gao, F., and Mu, H. J. (2015). Association of HLA and cytokine gene polymorphisms with idiopathic pulmonary fibrosis. *Kaohsiung J. Med. Sci.* 31, 613–620. doi: 10.1016/j.kjms.2015.10.007

Zhang, J., Xu, D. J., Xu, K. F., Wu, B., Zheng, M. F., Chen, J. Y., et al. (2012). HLA-A and HLA-B gene polymorphism and idiopathic pulmonary fibrosis in a Han Chinese population. *Respir. Med.* 106, 1456–1462. doi: 10.1016/j.rmed.2012.06.015

Zhang, Q. H., Wang, Y., Qu, D. H., Yu, J. Y., and Yang, J. L. (2019). The possible pathogenesis of idiopathic pulmonary fibrosis considering MUC5B. *Biomed. Res. Int.* 2019:9712464. doi: 10.1155/2019/9712464

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6

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: 05 February 2021
doi: 10.3389/fgene.2021.616686

Identifying Differentially Expressed Genes of Zero Inflated Single Cell RNA Sequencing Data Using Mixed Model Score Tests

Zhiqiang He[1], *Yueyun Pan*[2], *Fang Shao*[1]* *and Hui Wang*[3]*

[1] *Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, China,* [2] *First Clinical Medical College, Nanjing Medical University, Nanjing, China,* [3] *Department of Maternal and Child Health, School of Public Health, Peking University Health Science Center, Beijing, China*

Single cell RNA sequencing (scRNA-seq) allows quantitative measurement and comparison of gene expression at the resolution of single cells. Ignoring the batch effects and zero inflation of scRNA-seq data, many proposed differentially expressed (DE) methods might generate bias. We propose a method, single cell mixed model score tests (scMMSTs), to efficiently identify DE genes of scRNA-seq data with batch effects using the generalized linear mixed model (GLMM). scMMSTs treat the batch effect as a random effect. For zero inflation, scMMSTs use a weighting strategy to calculate observational weights for counts independently under zero-inflated and zero-truncated distributions. Counts data with calculated weights were subsequently analyzed using weighted GLMMs. The theoretical null distributions of the score statistics were constructed by mixed Chi-square distributions. Intensive simulations and two real datasets were used to compare edgeR-zinbwave, DESeq2-zinbwave, and scMMSTs. Our study demonstrates that scMMSTs, as supplement to standard methods, are advantageous to define DE genes of zero-inflated scRNA-seq data with batch effects.

Keywords: score test, generalized linear mixed model, zero inflation, observational weights, differential expression analyses, single cell RNA sequencing

Edited by:
Alfredo Pulvirenti,
University of Catania, Italy

Reviewed by:
Tiejun Tong,
Hong Kong Baptist University,
Hong Kong
Shiquan Sun,
Xi'an Jiaotong University, China

Correspondence:
Fang Shao
shaofang@njmu.edu.cn
Hui Wang
huiwang@bjmu.edu.cn

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *13 October 2020*
Accepted: *14 January 2021*
Published: *05 February 2021*

Citation:
He Z, Pan Y, Shao F and Wang H
(2021) Identifying Differentially
Expressed Genes of Zero Inflated
Single Cell RNA Sequencing Data
Using Mixed Model Score Tests.
Front. Genet. 12:616686.
doi: 10.3389/fgene.2021.616686

INTRODUCTION

In modern biology, transcriptomics has been widely used to elucidate the molecular basis of biological processes and diseases (Van den Berge et al., 2018). Previous transcriptome sequencing techniques (bulk RNA-seq) (Wang et al., 2009) might obscure the cell type heterogeneity in different samples. Because of the resolution, bulk RNA-seq hardly defines the rare cells, such as stem cells and tumor cells. Single cell RNA sequencing (scRNA-seq) enables researchers to study characteristics of gene expression in the resolution of individual cells (Kolodziejczyk et al., 2015). scRNA-seq has been treated as an effective method to study cellular heterogeneity in complex biological systems, and is being applied by more researchers in various biological processes, such as stem cell development and differentiation, embryonic organ development, tumors, immunology, and neurology (Tang et al., 2009; McEvoy et al., 2011; Zeisel et al., 2015; Chu et al., 2016; Papalexi and Satija, 2018; Sun et al., 2019). Identifying differentially expressed (DE) genes is one of the most common analysis of both bulk RNA-seq and of scRNA-seq analysis (Robinson et al., 2010; Van den Berge et al., 2017, 2018; Sun et al., 2018).

For bulk RNA-seq and scRNA-seq data, batch effects conventionally were treated as the non-biological differences that occurs when samples or cells are measured in distinct batches. The measure of transcriptome can be influenced by different environments for cells (Luecken and Theis, 2019). Various methods to correct batch effects and preserve biological variability have been presented. Some methods directly remove or correct batch effects using linear models (Johnson et al., 2007; Tung et al., 2017; Somekh et al., 2019). ComBat (Johnson et al., 2007) is an empirical Bayes method which takes batch effects into a linear regression model of gene expression. ComBat was recommended for batch correction when groups or cell types and state compositions between batches are consistent (Luecken and Theis, 2019). Mutual nearest neighbors (MNNs) (Haghverdi et al., 2018) and canonical correlation analysis (CCA) (Butler et al., 2018) remove batch effects using nonlinear models. A method comparison study showed ComBat was the best one for both bulk RNA-seq and scRNA-seq data (Büttner et al., 2019). For DE analysis, it was recommended that DE testing should be conducted on measure data with covariates including the batch information in the model design, not on batch corrected data (Luecken and Theis, 2019).

Some studies directly used traditional bulk RNA-seq DE methods (Krieg et al., 2018; Roerink et al., 2018; Li et al., 2019; Mehtonen et al., 2020). Limma-voom (Ritchie et al., 2015) applies weighted linear regression models for log-transformed count data. edgeR (Robinson et al., 2010; McCarthy et al., 2012) and DESeq2 (Love et al., 2014) model the gene expression count data based on generalized linear models (GLMs) under negative binomial (NB) distributions. It was demonstrated that NB models overestimated the dispersion parameter with excess zero counts, which influenced the power to DE analysis (Van den Berge et al., 2018). Different to bulk RNA-seq data, dropout events cause excess zeros for scRNA-seq read count data (Finak et al., 2015; Hashimshony et al., 2016). Therefore, zero inflation or an excess of zeros is a particular feature of scRNA-seq data, and it is not considered for these methods. SCDE (Kharchenko and Fan, 2019) and MAST (Finak et al., 2015; McDavid et al., 2019) model the redundant zeros of scRNA-seq data by zero inflation and hurdle models, respectively. Both zinbwave (Risso et al., 2018; Van den Berge et al., 2018) and zingeR (Van den Berge et al., 2017) estimates observational weights based on a zero-inflated negative binomial (ZiNB) model and downweight excess zeros followed by classical bulk RNA-seq DE tools (e.g., edgeR and DESeq2). The performance of two combinations, edgeR-zinbwave and DESeq2-zinbwave, outperform other DE methods (Van den Berge et al., 2018).

Here, based on isoVCT (Yang et al., 2017) and SMMATs (Chen et al., 2019), we implement a series of efficient methods, the single cell mixed model score tests (scMMSTs), to identify DE genes for defined cell types in scRNA-seq data considering batch effects and zero inflation. isoVCT, a DE method for bulk RNA-seq, uses a random effect to consider the heterogeneous isoform effects. In large-scale whole-genome sequencing (WGS) studies, SMMATs are powerful and computationally efficient variant set tests for continuous and binary traits, which integrates the burden test and SKAT (Wu et al., 2011) under the framework of generalized linear mixed models (GLMMs).

METHODS

Generalized Linear Mixed Models

For a single gene, we consider the following:

$$g\left(\mu_i\right) = \alpha + g_i \mathbf{B}_i \boldsymbol{\beta} + \mathbf{B}_i \mathbf{b},$$

where $g(\cdot)$ is a monotonic differentiable link function for GLMs, $\mu_i = E\left(y_i | g_i, \mathbf{B}_i, \mathbf{b}\right)$ denotes the mean of phenotype or count y_i for subject or cell i for a given gene with sample size n to the intercept α, g_i is the group, cluster or cell type covariate dummy variable binary value for subject i, $\mathbf{B}_i$ is the row vector of dummy variables values of the batch or individual covariate for subject i, $\boldsymbol{\beta}$ is the group effects associated with bathes and $\mathbf{b}$ is the batch effects. In the above equation, the group effects $\boldsymbol{\beta}$ are assumed to follow the normal distribution $N(\beta_0 \mathbf{1}_p, \sigma_\beta^2 \mathbf{I}_p)$, where $\mathbf{1}_p$ is the $p \times 1$ dimensional vector whose elements are all 1, $\mathbf{I}_p$ is the $p \times p$ dimensional identity matrix, β_0 and σ_β^2 are mean and variance of the normal distribution and p is the number of batches. If $\sigma_\beta^2 > 0$, group effects are associated with the batches. We assume the batch random effects $\mathbf{b} \sim N(\mathbf{0_p}, \sigma_b^2 \mathbf{I}_p)$, where $\mathbf{0}_p$ is the $p \times 1$ dimensional vector whose elements are all 0 and σ_b^2 is the variance. We consider the binomial, quasi-binomial, Poisson, quasi-Poisson, and NB distributions to model y_i. Binary phenotypes are commonly modeled by binomial and quasi-binomial distributions and counts are commonly modeled by Poisson, quasi-Poisson, and NB distributions.

For single cell RNA-seq data of a given gene, y_i is the count for cell i. We identify DE genes for each defined cell type in the form of one-against-others, so g_i, the cell type covariate for cell i, is binary. GLMMs under Poisson, quasi-Poisson and NB distributions are appropriate in this scenario.

Single Cell Mixed Model Score Tests

Testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is equivalent to testing $H_0 : \beta_0 = 0$ and $\sigma_\beta^2 = 0$. Under the null hypothesis, the reduced GLMM is as follows.

$$g\left(\mu_{0i}\right) = \alpha + \mathbf{B}_i \mathbf{b},$$

where $\mu_{0i} = E\left(y_i | \mu_0, b_i\right)$.

We construct a variance component score test statistic T derived by testing $H_0' : \sigma_\beta^2 = 0$ under the assumption $\beta_0 = 0$. SMMAT-O was also derived in the same manner. Under H_0' with the assumption $\beta_0 = 0$, we have the same reduced null model as that under $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Therefore, our derived test statistic T is applicable for testing H_0. The test statistic T is shown as follows.

$$T = \frac{\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)^T \widehat{\boldsymbol{\Phi}} \mathbf{G_B} \mathbf{G_B^T} \widehat{\boldsymbol{\Phi}} \left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)}{\widehat{\tau}},$$

where $\mathbf{y} = \left(y_1\, y_2 \cdots y_n\right)^T$ is an $n \times 1$ vector of counts or phenotypes, $\widehat{\boldsymbol{\mu}}_0 = g^{-1}\left(\widehat{\alpha} + \mathbf{B}_i \widehat{\mathbf{b}}\right)$ is the estimated mean vector of the reduced null model under H_0, $\widehat{\alpha}$ and $\widehat{\mathbf{b}}$ are estimates of the α and $\mathbf{b}$, $\boldsymbol{\Phi} = diag\left\{1/\left(1 + \left(\widehat{\mu}_{0i}/\widehat{\theta}\right)\right)\right\}$ for the NB distribution with the estimated dispersion parameter $\widehat{\theta}$ and $\widehat{\boldsymbol{\Phi}} = \mathbf{I}_n$ for other distributions mentioned, $\mathbf{B} = \left(\mathbf{B}_1^T \mathbf{B}_2^T \ \cdots \ \mathbf{B}_n^T\right)^T$ is an $n \times p$ design matrix of group covariate dummy variables values, $\mathbf{G_B} = \left(g_1\mathbf{B}_1^T g_2\mathbf{B}_2^T \ \cdots \ g_n\mathbf{B}_n^T\right)^T$ is an $n \times p$ design matrix of interactions of group and batch covariates with the multiplication of corresponding dummy variables values and $\widehat{\tau}$ is the estimate of dispersion parameter τ for quasi distributions, which is 1 for the binomial, Poisson and NB distributions and is estimated by the residual deviance divided by the degree of freedom of the reduced null model for quasi-binomial and quasi-Poisson distributions.

The asymptotic distribution of the statistic T under H_0 is derived as follows. Following the theoretical results of mixed models (Harville, 1977; Breslow and Clayton, 1993; Santos Nobre and da Motta Singer, 2007; Chen et al., 2016), we have $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0)/\sqrt{\widehat{\tau}}$ asymptotically following a n-dimensional multivariate normal distribution $\mathbf{M}VN_{\mathbf{n}}(\mathbf{0}, \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{V}}\widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{P}}\widehat{\mathbf{V}}\widehat{\mathbf{D}}^{-1})$ under H_0, where $\widehat{\mathbf{D}} = diag\left\{g'(\widehat{\mu}_{0i})\right\}$, whose diagonal elements are the first order derivative of the link function $g(\cdot)$ evaluated at $\widehat{\mu}_{0i}$, $\widehat{\mathbf{P}}$ is the $n \times n$ projection matrix of the reduced null model $\widehat{\mathbf{P}} = \widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}} - \widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\mathbf{1}_n\left(\mathbf{1}_n^{\mathbf{T}}\widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\mathbf{1}_n\right)^{-\mathbf{1}}\mathbf{1}_n^{\mathbf{T}}\widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}$ with $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{V}} + \widehat{\sigma}_b^2\mathbf{B}\mathbf{B}^{\mathbf{T}}$, $\widehat{\mathbf{V}} = diag\left\{\left(g'(\widehat{\mu}_{0i})\right)^2\widehat{Var}(y_i)\right\}$, the first order derivative function of the link function $g'(\cdot)$ and the estimated variance of y_i, $\widehat{Var}(y_i)$. For binomial and quasi-binomial distributions, $\left(g'\left(\widehat{\mu}_{0i}\right)\right)^2\widehat{Var}(y_i) = 1/[\widehat{\mu}_{0i}\left(1 - \widehat{\mu}_{0i}\right)]$. For Poisson and quasi-Poisson distributions, $\left(g'\left(\widehat{\mu}_{0i}\right)\right)^2\widehat{Var}(y_i) = 1/\widehat{\mu}_{0i}$. For NB distributions, $\left(g'\left(\widehat{\mu}_{0i}\right)\right)^2\widehat{Var}\left(y_i\right) = (1/\widehat{\mu}_{0i}) + \left(1/\widehat{\theta}\right)$. Since $\widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{P}} = \widehat{\mathbf{P}}$ and $\widehat{\boldsymbol{\Phi}} = \widehat{\mathbf{V}}^{-1}\widehat{\mathbf{D}}$, the asymptotic distribution can be simplified as $\mathbf{M}VN_{\mathbf{n}}\ (\mathbf{0}, \widehat{\boldsymbol{\Phi}}^{-\mathbf{1}}\widehat{\mathbf{P}}\widehat{\boldsymbol{\Phi}}^{-\mathbf{1}})$. Therefore, under H_0, T, a quadratic form of $\widehat{\mathbf{e}}$, asymptotically follows a mixture Chi-square distribution $\sum_{i=1}^{p}\xi_i\chi_{1,i}^2$, where $\chi_{1,i}^2$ are independent Chi-square distributions with 1 degree of freedom, and ξ_i are the eigenvalues of $\mathrm{E} = \mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\widehat{\mathbf{P}}\mathbf{G_B}$. Notably, $\widehat{\boldsymbol{\Sigma}}$ in $\widehat{\mathbf{P}}$ has a simple structure which makes $\widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}$ to be solved explicitly and E to be calculated efficiently. The *p*-value of the test can be calculated soon after the estimation of the reduced null model. More details of the computational efficiency of scMMSTs are discussed in section "Performance Evaluation". The estimation procedure of $\widehat{\mu}_{0i}$ is the same for binomial and quasi-binomial distribution pair and the Poisson and quasi-Poisson distribution pair. Thus, we implement quasi distributions to allow flexibility. In the followings, unless specified otherwise, "binomial" stands for both binomial and quasi-binomial and "Poisson" stands for both Poisson and quasi-Poisson.

There is zero inflation in scRNA-seq count data. Therefore, following the idea of ZINB-WaVE, a weighting strategy is implemented. Firstly, observational weights are calculated for all counts independently with details shown in sections "Zero-Inflated and Zero-Truncated Distributions for Counts" and "Calculations of Observational Weights for scMMSTs." Afterward, counts data with calculated weights are analyzed under the weighted GLMMs. Accordingly, a weighted version test statistic T_w for scMMSTs is proposed as follows with above notations.

$$T_w = \frac{\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)^T\widehat{\boldsymbol{\Phi}}\mathbf{W}\mathbf{G_B}\mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\mathbf{W}\widehat{\boldsymbol{\Phi}}\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)}{\widehat{\tau}},$$

where $\mathbf{W} = diag\left\{w_i\right\}$ and w_i is the given weights for count y_i. The estimation is based on the weighted GLLMs for the reduced null model. We denote $\mathbf{1}_{\mathbf{w},n} = \mathbf{W}^{\frac{1}{2}}\mathbf{1}_n$, $\mathbf{B_w} = \mathbf{W}^{\frac{1}{2}}\mathbf{B}$, $\widehat{\mathbf{V}}_{\mathbf{w}} = \mathbf{W}^{-\frac{1}{2}}\widehat{\mathbf{V}}\mathbf{W}^{-\frac{1}{2}}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}} = \widehat{\mathbf{V}} + \widehat{\sigma}_b^2\mathbf{B_w}\mathbf{B}_{\mathbf{w}}^{\mathbf{T}}$, $\widetilde{\widetilde{\widehat{\boldsymbol{\Sigma}}}}_{\mathbf{w}} = \mathbf{W}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}\mathbf{W}^{-\frac{1}{2}}$ and $\widehat{\mathbf{P}}_{\mathbf{w}} = \mathbf{W}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}^{-\mathbf{1}}\mathbf{W}^{\frac{1}{2}} - \mathbf{W}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}^{-\mathbf{1}}\mathbf{1}_{\mathbf{w},n}\left(\mathbf{1}_{\mathbf{w},n}^{\mathbf{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}^{-\mathbf{1}}\mathbf{1}_{\mathbf{w},n}\right)^{-\mathbf{1}}\mathbf{1}_{\mathbf{w},n}^{\mathbf{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}^{-\mathbf{1}}\mathbf{W}^{\frac{1}{2}} = \widetilde{\widetilde{\widehat{\boldsymbol{\Sigma}}}}_{\mathbf{w}}^{-1} - \widetilde{\widetilde{\widehat{\boldsymbol{\Sigma}}}}_{\mathbf{w}}^{-1}\mathbf{1}_n\left(\mathbf{1}_n^{\mathbf{T}}\widetilde{\widetilde{\widehat{\boldsymbol{\Sigma}}}}_{\mathbf{w}}^{-1}\mathbf{1}_n\right)^{-\mathbf{1}}\mathbf{1}_n^{\mathbf{T}}\widetilde{\widetilde{\widehat{\boldsymbol{\Sigma}}}}_{\mathbf{w}}^{-1}$. Based on the theoretical results of weighted GLMMs (Harville, 1977; Breslow and Clayton, 1993; Santos Nobre and da Motta Singer, 2007; Chen et al., 2016), if H_0 and $\mathbf{W}$ are true, we have $\widehat{\mathbf{e}}$ asymptotically normally distributed as $\mathbf{M}VN_{\mathbf{n}}(\mathbf{0}, \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{V}}_{\mathbf{w}}\widehat{\mathbf{P}}_{\mathbf{w}}\widetilde{\widetilde{\widehat{\boldsymbol{\Sigma}}}}_{\mathbf{w}}\widehat{\mathbf{P}}_{\mathbf{w}}\widehat{\mathbf{V}}_{\mathbf{w}}\widehat{\mathbf{D}}^{-1})$. Since $\widehat{\mathbf{P}}_{\mathbf{w}}\widetilde{\widetilde{\widehat{\boldsymbol{\Sigma}}}}_{\mathbf{w}}\widehat{\mathbf{P}}_{\mathbf{w}} = \widehat{\mathbf{P}}_{\mathbf{w}}$, $\widehat{\boldsymbol{\Phi}} = \widehat{\mathbf{V}}^{-1}\widehat{\mathbf{D}}$ and $\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{V}}_{\mathbf{w}} = \widehat{\mathbf{D}}^{-1}\mathbf{W}^{-\frac{1}{2}}\widehat{\mathbf{V}}\mathbf{W}^{-\frac{1}{2}} = \widehat{\boldsymbol{\Phi}}^{-\mathbf{1}}\mathbf{W}^{-1}$, where $\mathbf{W}^{-\frac{1}{2}}$, $\widehat{\mathbf{D}}^{-1}$, $\widehat{\mathbf{V}}$ are diagonal matrices, the asymptotic distribution can be simplified as $\mathbf{M}VN_{\mathbf{n}}(\mathbf{0}, \widehat{\boldsymbol{\Phi}}^{-\mathbf{1}}\mathbf{W}^{-1}\ \widehat{\mathbf{P}}_{\mathbf{w}}\mathbf{W}^{-1}\widehat{\boldsymbol{\Phi}}^{-\mathbf{1}})$. If H_0 and $\mathbf{W}$ are true, T_w, a quadratic form of $\widehat{\mathbf{e}}$, asymptotically follows a mixture Chi-square distribution $\sum_{i=1}^{p}\xi_i\chi_{1,i}^2$, where $\chi_{1,i}^2$ are independent Chi-square distributions with 1 degree of freedom, and ξ_i are the eigenvalues of $\mathrm{E}_{\mathbf{w}} = \mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\widehat{\mathbf{P}}_{\mathbf{w}}\mathbf{G_B}$. Note that $\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}$ in $\widehat{\mathbf{P}}_{\mathbf{w}}$ does not have the simple structure of $\widehat{\boldsymbol{\Sigma}}$, which makes it hard to analytically and explicitly solve $\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}^{-1}$. Therefore, we propose $\mathrm{E}'_{\mathbf{w}} = \mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\mathbf{W}\widehat{\mathbf{P}}\mathbf{W}\mathbf{G_B}$ to approximate $\mathrm{E}_{\mathbf{w}}$ for simplicity and efficiency, where we treat $\widehat{\mathbf{e}}$ as it is estimated by GLMMs without weights. Calculated weights are 1 for nonzero counts and between 0 and 1 for zero counts. Thus, this approximation performs worse when there are more redundant zeros, which might influence the performance of scMMSTs.

Zero-Inflated and Zero-Truncated Distributions for Counts

Zero-Inflated Distributions for Counts

A zero-inflated distribution for counts is a mixture distribution with two components, which are a point mass at zero and a conventional random variable distribution for counts, e.g., Poisson and NB distributions. The probability mass function (pmf) of a zero-inflated distribution for counts is as follows.

$$f_{ZI}\left(y; \boldsymbol{\theta}, \pi\right) = \pi\delta_0\left(y\right) + (1 - \pi)f\left(y; \boldsymbol{\theta}\right),\ \forall y \in \mathbb{N},$$

where $\pi \in [0, 1]$ indicates the probability of zero inflation, $\delta_0(\cdot)$ the Dirac function, $f(\cdot; \theta)$ the pmf of a conventional distribution with parameter vector $\boldsymbol{\theta}$. The observational weights of the counts can be calculated under a zero-inflated distribution model as the conditional probability that a given count y belongs to the conventional distribution with parameter estimates $\widehat{\boldsymbol{\theta}}$, $\widehat{\pi}$:

$$w = \frac{(1 - \widehat{\pi})f\left(y; \widehat{\boldsymbol{\theta}}\right)}{f_{ZI}\left(y; \widehat{\boldsymbol{\theta}}, \widehat{\pi}\right)}.$$

Note that w is 1 for nonzero counts and $\in (0, 1)$ for zeros counts. All the weights for counts under the conventional distribution are 1. Under a zero-inflated distribution, we take the weights of nonzero counts remain 1 and downweight zero counts from 1 to the conditional probability that a given count y belongs to the conventional distribution. Counts with observational weights are subsequently analyzed under the weighted version of models for the conventional distribution. In ZINB-WaVE, this weighting strategy is applied and the above formula is applied to calculate observational weights under the ZiNB distribution (Van den Berge et al., 2018).

Zero-Truncated Distributions for Counts

A zero-truncated distribution for counts is a distribution for counts with random variable values truncated at zero, i.e., only counts larger than zero can be observed. In the followings, we refer to zero-truncated distributions as truncated distributions for short. The pmf of a truncated distribution for counts is as follows.

$$f_{Tr}\left(y;\boldsymbol{\theta}\right)=\frac{f\left(y;\boldsymbol{\theta}\right)}{P_f(t>0;\boldsymbol{\theta})}=\frac{f\left(y;\boldsymbol{\theta}\right)}{\sum_{t=1}^{+\infty}f\left(t;\boldsymbol{\theta}\right)},\ \forall y\in\mathbb{N}_+,$$

where $f\left(\cdot;\boldsymbol{\theta}\right)$ denotes the pmf of a conventional distribution for counts with parameter vector$\boldsymbol{\theta}$. The observational weights of nonzero counts are 1 and weights of zero counts can be calculated under a truncated distribution model as following:

$$w=\frac{n_1 f\left(y=0;\widehat{\boldsymbol{\theta}}\right)}{n_0\sum_{t=1}^{+\infty}f\left(t;\widehat{\boldsymbol{\theta}}\right)},$$

where n_1 is the number of nonzero counts, n_0 is the number of the zero counts in the whole sample and $\widehat{\boldsymbol{\theta}}$ is the parameter vector estimate.

The derivation of the above formula is as follows. Nonzero counts follow the truncated distribution with parameter $\boldsymbol{\theta}$ which is the also the parameter for the corresponding conventional distribution. Therefore, the probability of zero counts is estimated as $f\left(y=0;\widehat{\boldsymbol{\theta}}\right)$. All the weights for counts under the conventional distribution are 1. However, since excess zeros are presented, the observational weights of nonzero counts remain 1 and zero counts are reweighted from 1 to w, so that $\frac{w\cdot n_0}{w\cdot n_0+1\cdot n_1}=f\left(y=0;\widehat{\boldsymbol{\theta}}\right)$. The resulting formula for observational weights w is derived by solving the equation. Counts are then analyzed with observational weights calculated under the weighted version of models for the conventional distribution.

Calculations of Observational Weights for scMMSTs

In ZINB-WaVE, the weighting strategy shown in the previous section is applied and observational weights are estimated by the ZiNB regression (Van den Berge et al., 2018). For our methods, the truncated Poisson (TrPois), zero-inflated Poisson (ZiPois), truncated negative binomial (TrNB), and ZiNB distributions are considered. Following the weighting strategy mentioned and $H_0:\boldsymbol{\beta}=\mathbf{0}$, we estimate parameters for counts in each batch and calculate the weights accordingly using the formulas in section "Zero-Inflated and Zero-Truncated Distributions for Counts" for simplicity with the assumption of no group effects.

For zero-inflated distributions, weights are the conditional probabilities that a count y belongs to the corresponding conventional distribution. We directly use ZINB-WaVE for the ZiNB distribution, and implement the algorithm in Appendix A of the paper (Böhning et al., 1999) for the ZiPois distribution. In ZINB-WaVE, no mixed models are involved. Thus, we treat batch effects as fixed effects in the ZiNB regression without group effects to calculate weights using all counts data, when using ZINB-WaVE. For TrPois distribution, since the pmf $f_{TrPois}\left(y\right)=\frac{f_{Pois}\left(y\right)}{1-e^{-\lambda}}=\frac{\lambda^y e^{-\lambda}}{y!1-e^{-\lambda}}$, we can derive the method of moment estimate and maximum likelihood estimate $\widehat{\lambda}$ and they are identical by numerically solve the equation $\frac{\widehat{\lambda}}{1-e^{-\widehat{\lambda}}}=\overline{y}$, where $\overline{y}$ is the sample mean for the truncated sample. For each batch, the weights are $w_i=\frac{n_1 e^{-\widehat{\lambda}}}{n_0\left(1-e^{-\widehat{\lambda}}\right)}$ for a zero count and $w_i=1$ for nonzero y_i, where n_1 is truncated sample size for the batch and n_0 is the number of the zero counts in the batch. TrPois and ZiPois perform very close to each other. For TrNB distribution, we implement the formulas in section "Results" of the paper (Rider, 1955) to estimate the mean parameter μ and the dispersion parameter θ for each batch. The common dispersion parameter θ is estimated by the harmonic mean of the estimated $\widehat{\theta}$ for each batch. However, this algorithm is not robust for small θ ($\theta<2$, based on simulations). The weights are $w_i=\frac{n_1\left(\widehat{\theta}/\left(\widehat{\theta}+\widehat{\mu}\right)\right)^{\widehat{\theta}}}{n_0\left(1-\left(\widehat{\theta}/\left(\widehat{\theta}+\widehat{\mu}\right)\right)^{\widehat{\theta}}\right)}$ for zero counts in each batch, where $\widehat{\theta}$ and $\widehat{\mu}$ are respectively the estimated dispersion and mean parameters for the NB distribution using counts in the batch, and $w_i=1$ for nonzero y_i for each corresponding batch.

After weights are calculated, counts data with weights are analyzed under weighted GLMMs shown in section "Single Cell Mixed Model Score Tests." Note that weights are calculated independently of GLMMs. Theoretically, the weights are 1 under conventional distributions. The calculated observational weights for nonzero counts remain 1. If there are calculated weights of zero counts far from 1 and closer to 0, it indicates that there are excess zeros. If calculated weights of zero counts are close to 1, the results for conventional distributions are similar to those considering zero inflation. In ZiNB-Wave, weights are calculated through the ZiNB regressions on all counts. However, the weights for TrPois, ZiPois, and TrNB are calculated using counts for each batch with smaller sample sizes. Therefore, although the calculation of weights for TrPois, ZiPois and TrNB is easier to implement and time saving, it is less accurate and less reliable than that for ZiNB-Wave and the performances of scMMSTs are affected.

Performance Evaluation

Performances of DE methods considered are assessed in terms of the per-comparison error rate (PCER), which refers to type I error rate (i.e., the proportion of false positives), line plots of the true positive rate (TPR) vs. the false discovery proportion (FDP) and the areas under the receiver operating characteristic (ROC) curves [i.e., the TPR vs. the false positive rate (FPR) curves] (AUCs) with definitions as follows.

$$TPR = \frac{TP}{P},\ FPR = \frac{FP}{N}, FDP = \frac{FP}{\max(1, FP + TP)'}$$

where we use the following abbreviations for empirical quantities: FP (the number of false positives), TP (the number of true positives), N (the number of negative samples), P (the number of positive samples). FDP-TPR curves for adjusted *p*-values are plotted by ***iCOBRA*** Bioconductor R package (version 1.12.1) (Soneson and Robinson, 2016) and AUCs for adjusted *p*-values are calculated by ***pROC*** R package (version 1.16.2) (Robin et al., 2011). Unless otherwise stated, the adjusted *p*-values for all DE methods considered are calculated by the Benjamini and Hochberg method (Benjamini and Hochberg, 1995) for FDR control.

Comparison Methods

The 12 methods considered for comparisons are Poisson, TrPois, ZiPois, NB, TrNB, NB-zinb, DESeq2, DESeq2-zinb, edgeR, edgeR-zinb, limma-voom, and MAST. The first six methods are our implemented methods of scMMSTs s under GLMMs assumptions and the last six methods are the state-of-the-art DE methods, where Tr, Zi, Pois, NB, and zinb are abbreviations of truncated, zero-inflated, Poisson, ZINB-WaVE, respectively. We follow the implementations of the last six DE methods above in the *zinbwave* paper (Van den Berge et al., 2018) and the R packages used are *edgeR* (version 3.28.1), *DESeq2* (version 1.26.0), *limma* (version 3.42.2), *MAST* (version 1.12.0), and *zinbwave* (version 1.8.0), which was developed to deal with zero inflation for scRNA-seq data by a weighting strategy and was used in edgeR-zinb, DESeq2-zinb, and NB-zinb. The binomial distribution scMMST is implemented, however, not covered in the simulations and real data analysis since only methods for count data are considered in thisarticle.

The implementations of scMMSTs are available in **Supplementary Data S1**. Codes for simulations and real data analysis are partially based on the GitHub repositories[12] of papers (Yang et al., 2017; Van den Berge et al., 2018) and the *GMMAT* R package (version 1.3.0) (Chen et al., 2016, 2019). R packages *doParallel* (version 1.0.15) (Corporation and Weston, 2019) and *BiocParallel* (version 1.20.1) (Morgan et al., 2019) are used for parallel computation. The reduced null model is estimated by *lme4* R package (version 1.1.23) and *p*-values are calculated by *CompQuadForm* R package (version 1.4.3). Simulated single cell datasets are generated by *splatter* R package (version 1.10.1) (Zappia et al., 2017). Additionally, the code to reproduce all analyses, figures and tables reported in this manuscript is attached in **Supplementary Data S1**.

Simulations

We perform simulations to evaluate performances of scMMSTs, which are our methods of association tests under the proposed GLMMs, comparing with state-of-art DE methods under a range of scenarios. We simulate the scRNA-seq data based on GLMMs directly and by the R package *splatter*. *Splatter* can directly estimate model parameters for real scRNA-seq data and generate quality controlled simulated mock datasets with DE genes easily and can add batch effects, which are not associated with group effects, to the simulated data. The simulated number of genes for one dataset by *splatter* and GLMMs is 10,000 and the number of cells is 250 with balanced two groups and five batches. In the DE genes simulations, the proportion of the DE genes is set to be 0.1.

Additional parameters of *splatter* simulations, batch.facLoc–batch factor location, batch.facScale–batch factor scale, and out.prob–the expression outlier probability, are set to be 0.5. For DE gene simulations, de.facLoc, DE factor location, is set to 2 and de.facScale, DE factor scale, is set to be 0.5.

The procedure to simulate datasets based on the proposed GLMMs is as follows. We assume that the scRNA-seq count data follow Poisson and NB distributions and generate y_i based on the GLMM shown with the parameters setting and generate a Bernoulli random variable z_i with parameter $\pi_i = logit^{-1}\,(\mu_\pi + \mathbf{B}_i\mathbf{b})$. Larger values of parameter μ_π causes smaller baseline proportions of zeros. If $z_i = 0$, then $y_i = 0$, and y_i remains the same otherwise. The parameter settings for simulations are based on the real data analysis and references (Yang et al., 2017). Seven parameters are considered: the variance of the batch or individual effects $\mathbf{b}(\sigma_b^2)$, the variance of the group or cell type effects $\boldsymbol{\beta}(\sigma_\beta^2)$, the baseline group effect (β_0), the number of batches (p), the dispersion parameter ($\theta = 1/\phi$) for NB distributions and the intercepts (μ_0) and (μ_π) for the GLMM and logstic regression for excess zeros, respectively. σ_b^2 shows the heterogeneity of batch effects in different batches. σ_β^2 shows the heterogeneity of group effects in different batches. β_0 shows the baseline group effect. The larger the $|\beta_0|$, the larger the baseline group effect is. Other parameters describe the features of the gene expression and zero inflation. σ_b^2 is set to be 0.25 and σ_β^2 varies in 0, 0.01, 0.25, and 1. β_0 varies in 0, 0.01, 0.1, 0.3, and 0.5. θ varies in 0.5, 1, and 2. μ_π varies in −1, 0, and 2. $p = 5$ and $\mu_0 = 5$.

Real Data Sets

Usoskin Dataset

This scRNA-seq dataset contains mouse neuronal cells in the dorsal root ganglion (Usoskin et al., 2015). The processed expression values were downloaded from the Github respiratory[3] of the *zinbwave* paper. Following the process procedures given in the *zinbwave* paper, the authors considered 622 cells with a classification of 11 neuronal cell-types, which were denoted as NF1 to NF5, NP1 to NP3, PEP1, PEP2 and TH. Genes with less than 20 counts were removed and a total of 12,132 genes are considered for the following analyses with 68% zero counts. The authors showed the existence of a batch effect related to the picking session for the cells. Thus, the picking session covariate (with values Cold, RT-1, and RT-2) in this dataset was considered as a batch covariate for real data analysis. The batch effect was associated with expression measures and the relationship between zero inflation and sequencing depth, which was shown

[1] https://github.com/biostat0903/RNAseq-Data-Analysis

[2] https://github.com/statOmics/zinbwaveZinger

[3] https://github.com/statOmics/zinbwaveZinger/tree/master/datasets

in **Figure 5** of the *zinbwave* paper (Hicks et al., 2015; Van den Berge et al., 2018). We repeated the results of Figures 5A,B of the *zinbwave* paper in **Supplementary Figures S1A,B**. There is a large variation in the depth of sequencing among batches, which weaken the overall association with zero inflation when pooling cells across batches (**Supplementary Figure S1A**). Zero inflation was also identified for the Usoskin dataset. Histograms of observational weights for nonzero counts, which were calculated by the ZINB-WaVE model including the cell type as a covariate with and without the batch effect as fixed effects, are shown in **Supplementary Figure S1B**. Calculated weights of nonzero counts with and without the batch effect both have high modes near zero. This suggests zero inflation in the Usoskin dataset. The real data analysis of the processed Usoskin dataset was done to identify DE genes for defined 11 cell types vs. the rest. Simulated datasets based on this dataset were generated by *spaltter* with estimated corresponding parameters. For a null dataset without DE genes, we created 10,000 genes, 250 cells, five balanced batches and two balanced groups for cells. Twelve methods were implemented to identify DE genes between the two groups for each of the 30 simulated null data sets. A gene was declared to be DE if its unadjusted *p*-value was less than or equal to 0.05. Declared DE genes were false positives for these simulated null datasets. The empirical PCER of each method was calculated as the proportion of declared DE genes and was compared to the 0.05 nominal PCER.

Tung Dataset

This scRNA-seq dataset is for induced pluripotent stem cells from three individuals from HapMap (Tung et al., 2017). Following the *splatter* paper (Zappia et al., 2017), the matrix of molecules (UMIs) was treated as counts and was used directly. This dataset is available from GEO (accession GSE77288)[4] and the Github respiratory[5] of the *splatter* paper. No batch information is available for this dataset. Genes with less than 20 counts were removed and a total of 14,893 genes with 864 cells containing 44% zero counts were considered. Zero inflation was identified for the Tung dataset. Histograms of observational weights of nonzero counts of two filtered datasets (18,726 genes with more than 0 count and 14,893 genes with more than 19 counts, respectively), which were calculated by the ZINB-WaVE model, are shown in **Supplementary Figures S1C,D**. There are moderate proportion s of calculated weights of nonzero counts close to zero. This suggests zero inflation in the Tung dataset. Comparing to the Usoskin dataset, the Tung dataset is less zero inflated. We generated 30 simulated null datasets and identified DE genes using the same procedures for the Usoskin dataset with *spaltter*.

RESULTS

Method Overview

Single cell mixed model score tests are computationally efficient DE analysis tools for scRNA-seq data considering batch effects and zero inflation. Bath effects are estimated as random effects under the reduced null models of GLMMs. A weighting strategy is implemented to characterize excess zeros. The score statistics are derived on theoretical asymptotic distributions. First, we estimated normalization factors of count matrix by the function *calcNormFactors* in *edgeR* after counts per million (CPM) normalization. Second, the estimation of the observational weights is efficient. We use *zinbwave* to fit NB-zinb which might be the most time-consumed assumption. Third, we use *lme4* for the estimation, the most efficient method to fit GLMM, to estimation the parameters in the null hypothesis (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2017). Considering the real data, the estimation procedure of mixed model is not related to the number of groups or cell types. Compared to the traditional estimation procedure, scMMSTs use three strategies to decrease memory usage and computation time. First, scMMSTs do not need to store $n \times n$ matrices $\widehat{\mathbf{P}}$ and $\widehat{\boldsymbol{\Sigma}}$ explicitly. The *p*-value is efficiently calculated by *CompQuadForm* with eigenvalues of E or $\mathrm{E}'_{\mathbf{w}}$, which is only a $p \times p$ matrix. Second, scMMSTs use an analytical form to calculate the inverse of $\widehat{\boldsymbol{\Sigma}}$ which might be the most time consumption procedure in the estimation of T or T_w. Third, scMMSTs is implemented for parallel computing. Therefore, although more complicated models GLMMs are considered, scMMSTs are computationally affordable compared to other DE methods.

Simulations by Real Datasets and *Splatter*

Simulated datasets generated by the *splatter* used parameters estimated from two publicly available real scRNA-seq datasets, the Usoskin (Usoskin et al., 2015) and Tung (Tung et al., 2017) datasets.

The FPR control was assessed by the PCER. Results are shown in **Figure 1**. For the Usoskin dataset, the estimated common dispersion parameter value of biological coefficient of variation (BCV) was $\hat{\phi} = 1/\widehat{\theta}$ =1.89. TrNB and Poisson failed to control the FPR. The PCERs of NB-zinb, DESeq2, edgeR-zinb, and edgeR were a little inflated. DESeq2-zinb and MAST controlled the FPRs with large variability, especially for MAST. Other methods were a little conservative with PCERs smaller than the nominal level 0.05. For the Tung dataset, the estimated common dispersion parameter value of BCV was $\hat{\phi} = 1/\widehat{\theta} = 0.11$. Poisson failed to control the FPR. The PCERs of TrNB and edgeR were a little inflated. Other methods conservatively controlled FPRs, especially for NB, DESeq2-zinb, and NB-zinb. We treated "NA" *p*-values of DE methods as 1, thus, there are peak bars at 1 for some methods in the unadjusted *p*-value histograms shown in **Figures 1B,D**. In summary, standard DE methods can control the FPRs and scMMSTs except Poisson and TrNB can conservatively control the FPRs. FPRs of scMMSTs increase as the dispersion parameter θ decreases.

False discovery proportion-true positive rate curves for adjusted *p*-values are shown in **Figure 2**. For the Usoskin dataset, bulk RNA-seq DE methods are shown to

[4] https://github.com/jdblischak/singleCellSeq

[5] https://github.com/Oshlack/splatter-paper

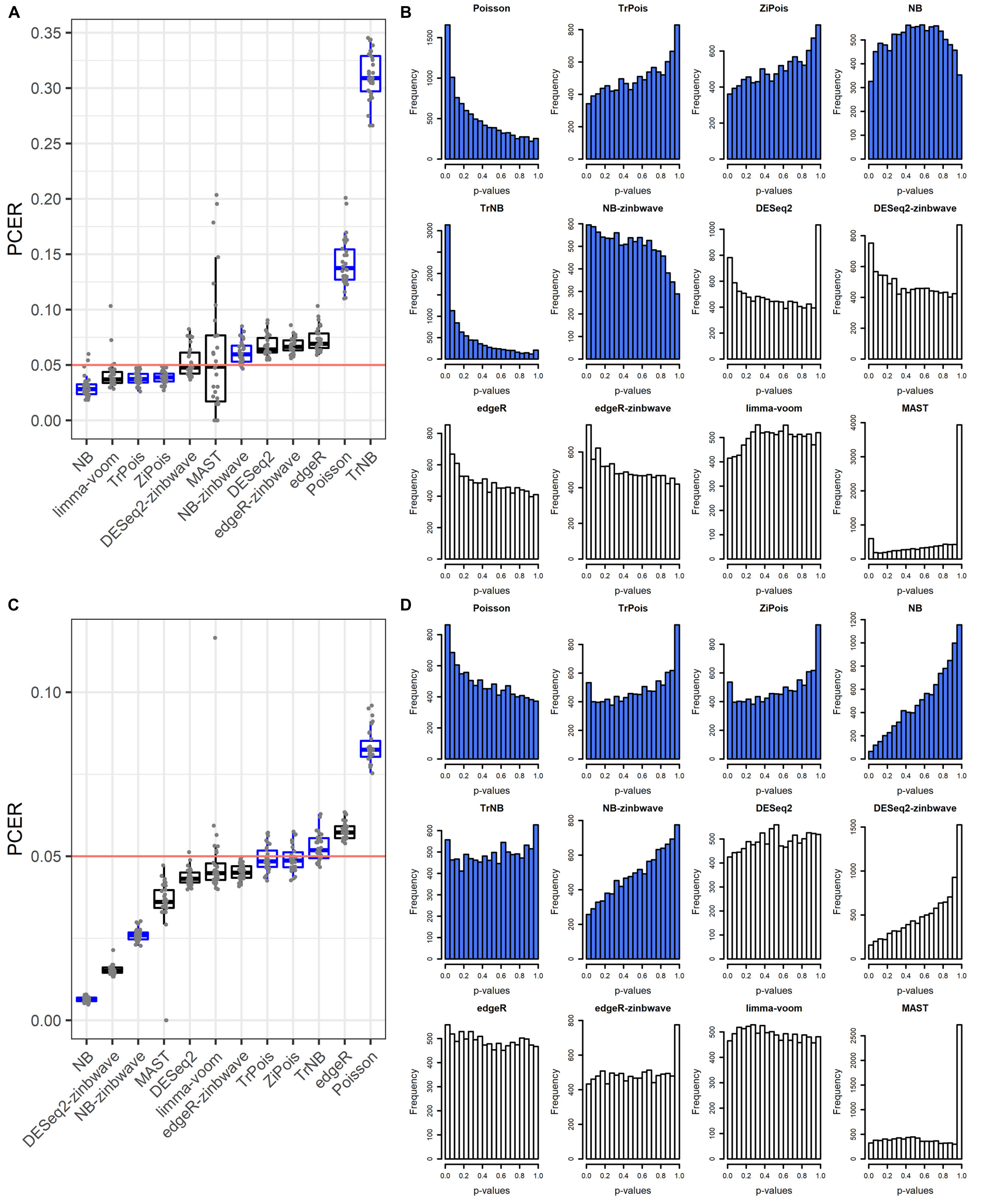

FIGURE 1 | False positive rate control on simulated null Usoskin datasets and Tung datasets. **(A)** Boxplot of PCER for 30 simulated null Usoskin datasets generated by *splatter* for each of 12 DE methods. scMMSTs are marked in blue. **(B)** Histogram of uncorrected *p*-values for one dataset in panel **A**. **(C)** Boxplot of PCER for 30 simulated null Tung datasets generated by *splatter* for each of 12 DE methods. scMMSTs are marked in blue. **(D)** Histogram of uncorrected *p*-values for one dataset in panel **C**. PCER, per-comparison error rate; DE, differential expression; scMMST, single cell mixed model score test.

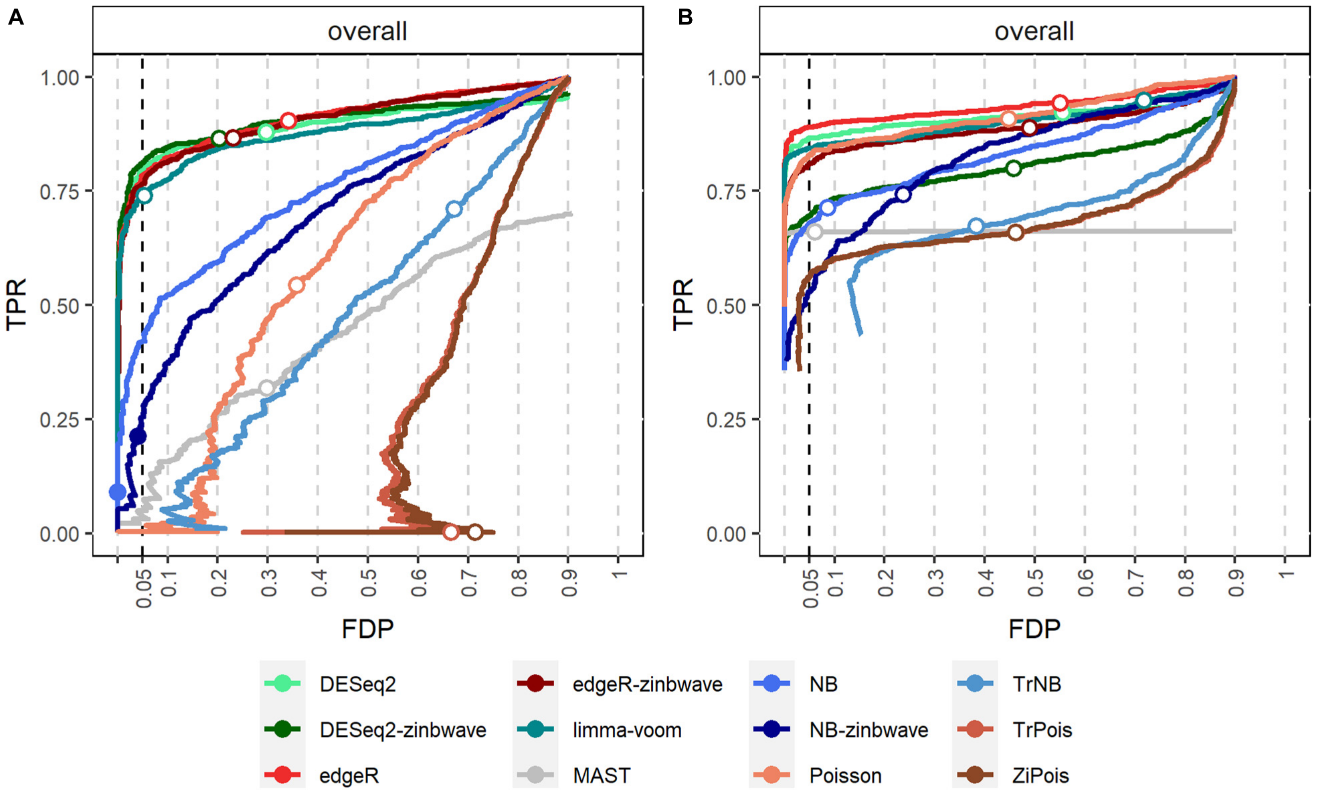

FIGURE 2 | FDP-TPR curves of DE methods on simulated Usoskin datasets and Tung datasets. **(A)** Line plot of the FDP-TPR curves for simulated Usoskin datasets generated by *splatter* for each of 12 DE methods. **(B)** Line plot of the FDP-TPR curves for simulated Tang datasets generated by *splatter* for each of 12 DE methods. Circles represent values at a 0.05 nominal FDR threshold and are filled in if the FDP (i.e., empirical FDR) is less than 0.05. DE, differential expression; TPR, true positive rate; FDP, false discovery proportion; FDR, false discovery rate.

perform well, possibly due to the high proportion of zeros and low counts (Van den Berge et al., 2018). In general, standard DE methods except MAST perform better than scMMSTs when the batch effects is not associated with group effects.

Simulations by GLMMs

Results of PCERs are shown in **Supplementary Figures S2, S3** and **Supplementary Table S1**. Methods performances of the FPR control were similar to those in simulations by *splatter*. Based on FDP-TPR curves for adjusted *p*-values shown in **Figure 3**, scMMSTs performed better than standard DE methods when batch effects were associated with weak group effects. NB-zinb was the best among all methods considered for comparisons. EdgeR-zinb and DESeq2-zinb were the best two methods among the six standard DE methods considered. TrPois and ZiPois perform very close to each other. **Figure 4** demonstrates bar plots of AUCs for adjusted *p*-values. $|\beta_0|$, σ_β^2 , θ and μ_π exhibited positive correlations with AUCs. Our scMMSTs performed better when the group effect size and its heterogeneity are larger and the counts dispersion BCV and proportion of zeros are smaller. Similar results are obtained to those of FDP-TPR curves. Therefore, our results demonstrate that scMMSTs performs better than standard DE methods when the group effect size is small with large group effect heterogeneity.

Real Data Analysis

Table 1 and **Supplementary Figure S4** show the numbers of DE genes detected by the 12 methods considered in simulations for 11 cell types in the Usoskin dataset. This dataset was also analyzed in the *zinbwave* paper. MAST failed for some cell-types, so no DE gene was detected. NB-zinb defined smallest number of DE genes in general. The results of Venn diagrams and Upset plots by R packages *VennDiagram* (version 1.6.20) (Chen, 2018) and *upsetR* (version 1.4.0) (Gehlenborg, 2019) are shown in **Supplementary Figures S5–S15**. Since NB-zinb is conservative for FDR, the DE genes only detected by NB-zinb highly likely have weak group effects with their heterogeneity across batches. In general, scMMSTs, as supplement to standard methods, are superior at selecting DE genes with weak group effects and their heterogeneity in different batches for scRNA-seq data.

Computational Time

To demonstrate the computation time scale of DE methods considered, we benchmarked two different simulated null datasets by *splatter* with parameters estimated by the Usoskin

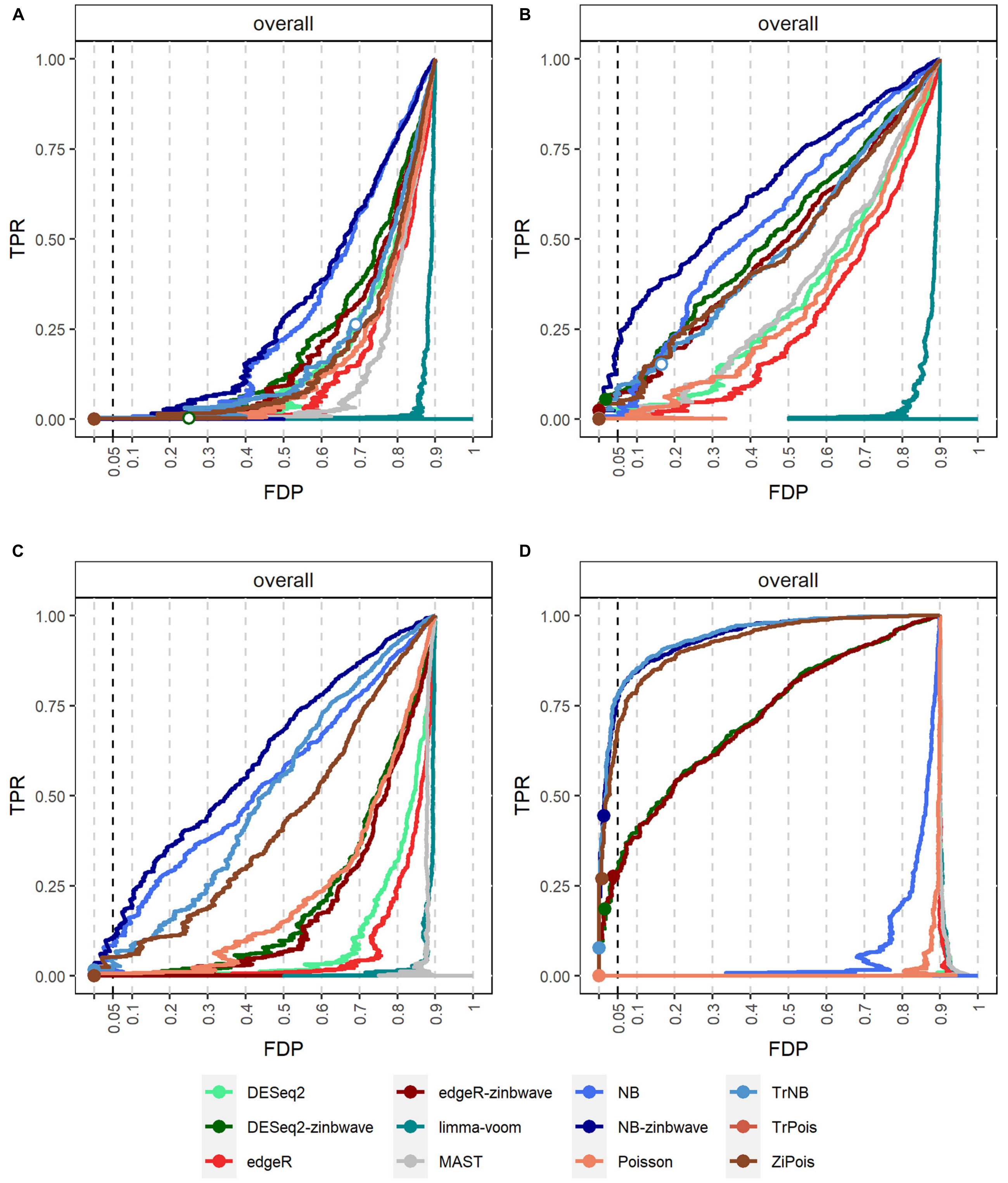

FIGURE 3 | FDP-TPR curves of DE methods on simulated datasets generated by GLMMs with $\mu_\pi =$ 0. **(A)** Line plot of the FDP-TPR curves for simulated datasets based on NB GLMMs for each of 12 DE methods with the dispersion parameter $\theta = 0.5$. **(B)** Line plot of the FDP-TPR curves for simulated datasets based on negative binomial (NB) GLMMs for each of 12 DE methods with $\theta = 1$. **(C)** Line plot of the FDP-TPR curves for simulated datasets based on NB GLMMs for each of 12 DE methods with $\theta = 2$. **(D)** Line plot of the FDP-TPR curves for simulated datasets based on Poisson GLMMs for each of 12 DE methods with $\beta_0 = \sigma_\beta^2 =$ 0.01. Circles represent values at a 0.05 nominal FDR threshold and are filled in if the FDP (i.e., empirical FDR) is less than 0.05. DE, differential expression; GLMM, generalized linear mixed model; NB, negative binomial; TPR, true positive rate; FDP, false discovery proportion; FDR, false discovery rate.

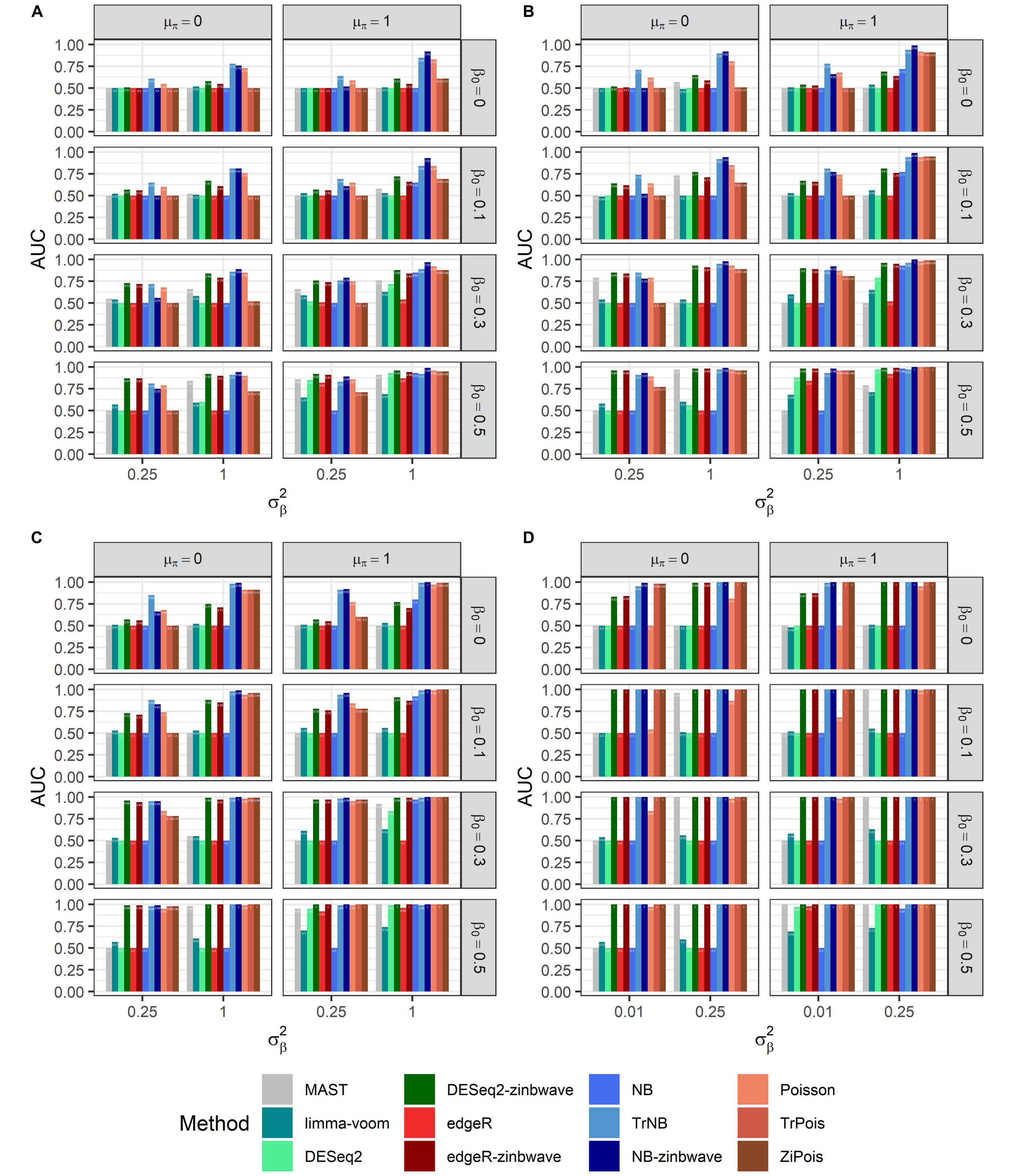

FIGURE 4 | AUCs of DE methods for simulated datasets generated by GLMMs with $\mu_\pi = 0$. Adjusted *p*-values are used as predictors. **(A)** Bar plot of AUCs for simulated datasets generated by NB GLMMs for each of 12 DE methods with the dispersion parameter $\theta = 0.5$. **(B)** Bar plot of AUCs for simulated datasets generated by NB GLMMs for each of 12 DE methods with $\theta = 1$. **(C)** Bar plot of AUCs for simulated datasets generated by NB GLMMs for each of 12 DE methods with $\theta = 2$. **(D)** Bar plot of AUCs for simulated datasets generated by Poisson GLMMs for each of 12 DE methods. AUC, area under curve; DE, differential expression; GLMM, generalized linear mixed model; NB, negative binomial.

TABLE 1 | Numbers of declared differentially expressed genes by 12 methods for 11 defined cell types vs. the rest in the Usoskin dataset (n = 622 cells).

Methods	NF1	NF2	NF3	NF4	NF5	NP1	NP2	NP3	PEP1	PEP2	TH
edgeR	826	1206	348	646	1070	1877	880	362	1833	328	2424
DESeq2	906	963	218	402	782	1988	748	407	2649	102	2387
limma-voom	5427	3762	3777	721	2572	2505	4857	203	7892	173	4800
MAST	0	0	0	2	0	85	5	2	10	0	112
edgeR-zinb	509	778	244	550	985	1871	987	486	2475	185	3225
DESeq2-zinb	555	1003	319	453	1235	1985	786	392	2249	153	3166
NB	295	517	186	365	555	462	329	218	592	145	533
TrNB	910	703	596	1763	885	1127	2139	2254	3752	537	1986
NB-zinb	192	308	77	295	364	976	467	270	2004	100	878
Pois	745	1214	410	881	1195	1401	745	583	2104	339	1942
TrPois	242	298	82	345	321	602	756	444	3353	54	708
ZiPois	337	311	81	487	376	607	1019	446	3350	137	704

and Tung datasets. Other settings remained the same as those in the simulations for PCERs. Results are shown in **Figure 5**. For both datasets, the fastest method was limma-voom. DESeq2 was slower than edgR, thus, DESeq2-zinb was also slower than edgeR-zinb. Our scMMSTs performed in the same scale of DESeq2-zinb and DESeq2-zinb. The computation times of simulated null Tung datasets were shorter than those of simulated null Usoskin datasets with the same number of cores. More cores used in the parallel computation made our scMMSTs faster. With eight cores, the computation times of Poisson related methods were close to MAST, edgeR, and DESeq2. In summary, our scMMSTs are computationally affordable compared to other DE methods especially when parallel computing is allowed. All computations were done on a cluster with 24 Intel Xeon Processor (Skylake, IBRS) at 2.60 GHz (2593 MHz) and 128 GB RAM.

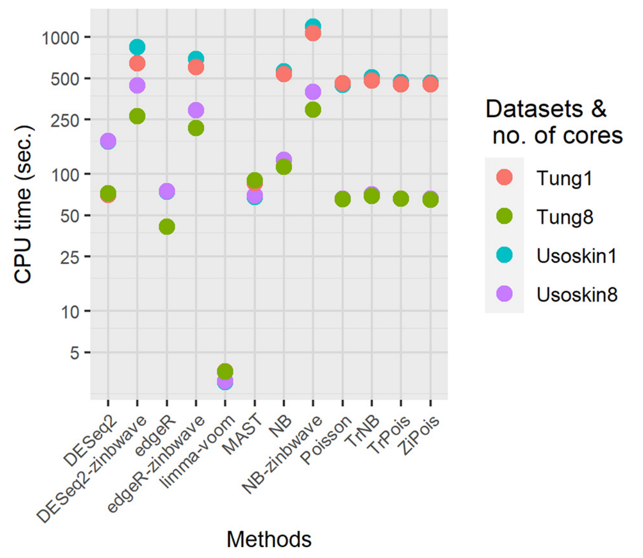

FIGURE 5 | Computational times for differential expression methods on the simulated null Usoskin and Tung datasets, which were generated by *splatter*. The number of cores were set to be 1 and 8 on a cluster with 24 Intel Xeon Processor (Skylake, IBRS) at 2.60 GHz (2593 MHz) and 128 GB RAM.

DISCUSSION

We proposed scMMSTs to identify DE genes, considering batch effect and zero inflation of scRNA-seq data. Both simulations and real data indicated that these methods have advantages in selecting DE genes with weak group effects and their heterogeneity in different batches. In simulations, scMMSTs conservatively controlled FPRs or type I error rates in each setting under assumptions of NB and Poisson distributions, except TrNB and Poisson assumption. However, TrNB controlled FPRs when θ is large. Second, following the model assumption, scMMST was the best one when $|\beta_0|$ was small and σ_β^2 was large, especially when θ was large. In real data analysis, the Venn diagrams and Upset plots of DE genes (**Supplementary Figures S5–S15**) directly indicated the relationships among the DE methods. scMMATs defined smaller numbers of DE genes and NB-zinb defined the smallest. Since scMMATs are conservative, the DE genes only defined by NB-zinb are likely to have the small group effect size with its heterogeneity across batches.

Furthermore, scMMSTs exhibited three innovations. First, scMMSTs derived the association test score statistics and their theoretical null distributions in the framework of GLMMs under the binomial, Poisson and NB assumptions. Second, the group effect $\boldsymbol{\beta}$ was modeled as random effects associated with batches in the framework of GLMMs. Third, scMMSTs verified their effectiveness to detect DE genes with the weak group effect and its heterogeneity in different batches. However, scMMSTs have some limitations. scMMSTs performed worse than other standard DE methods to detect DE genes without group effect heterogeneity across batches. scMMSTs performed worse when the dispersion parameter θ was small, especially for the TrNB method, this may due to the non-robust estimation of θ. scMMSTs, in fact, are derived to test H_0' under the assumption $\beta_0 = 0$, not to jointly test $\beta_0 = 0$ and $\sigma_\beta^2 = 0$. This decreases the power of testing H_0 for scMMSTs. For association tests, the Mixed effects Score Test (MiST), which jointly tests H_0, is more powerful. Therefore, scMMSTs may be extended using the framework of GLMM-MiST (Sun et al., 2013) in future work to overcome these drawbacks. E_w' is used to approximate E_w for the statistic $\mathbf{T_w}$ of scMMSTs. This approximation performs worse when there are more excess zeros. Better approximations of $E_{\mathbf{w}}$ or methods to efficiently calculate $E_{\mathbf{w}}$ may improve the performance of scMMSTs. The weighting strategy implemented may be explained in a Bayesian framework and scMMSTs may be extended accordingly. In addition, following the idea of PEA (Shao et al., 2019), scMMSTs may be extended to efficiently identify gene-pathway interactions without permutations of test statistics. In conclusion, scMMSTs, supplements to standard single cell DE methods, are advantageous at selecting genes with the weak group effect and its heterogeneity across batches for scRNA-seq data analysis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: the dataset (Usoskin) analyzed for this study can be found in the [Github respiratory of the *zinbwave* paper (Van den Berge et al., 2018)] (https://github.com/statOmics/zinbwaveZinger/blob/master/datasets/esetUsoskin.RData); the dataset (Tung) can be found in the [Github respiratory of the splatter paper (Zappia et al., 2017)](https://github.com/Oshlack/splatter-paper/blob/master/data.tar.gz).

AUTHOR CONTRIBUTIONS

FS and HW conceived and supervised the study. ZH and FS implemented the software, conducted the simulations, analyzed the data, and wrote the manuscript. ZH and YP prepared figures and tables. ZH, YP, HW, and FS modified and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 81703321 and 81502888), the Jiangsu Shuangchuang Plan and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank reviewers for their valuable feedback and comments which significantly improved the article's quality. The authors are grateful to YiDuCloud Tech. Ltd., for the support on high performance computation resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.616686/full#supplementary-material

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L., and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. R. Stat. Soc. Ser. A* 162, 195–209. doi: 10.1111/1467-985X.00130

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88, 9–25. doi: 10.2307/2290687

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

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* 16, 43–49. doi: 10.1038/s41592-018-0254-1

Chen, H. (2018). *VennDiagram: Generate High-Resolution Venn and Euler Plots*. Available online at: https://CRAN.R-project.org/package=VennDiagram (accessed June 8, 2020).

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., et al. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am. J. Hum. Genet.* 104, 260–274. doi: 10.1016/j.ajhg.2018.12.012

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* 98, 653–666.

Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17:173. doi: 10.1186/s13059-016-1033-x

Corporation, M., and Weston, S. (2019). *doParallel: Foreach Parallel Adaptor for the "Parallel" Package*. Available online at: https://CRAN.R-project.org/package=doParallel (accessed June 8, 2020).

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. New York, NY: Springer, doi: 10.1007/978-1-4614-6868-4

Eddelbuettel, D., and Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ. Prepr.* 5:e3188v1. doi: 10.7287/peerj.preprints.3188v1

Eddelbuettel, D., and François, R. (2011). Rcpp: Seamless R and C++ Integration. *J. Stat. Softw.* 40, 1–18. doi: 10.18637/jss.v040.i08

Finak, G., Mcdavid, A., Yajima, M., Deng, J., Gersuk, V. H., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278–278. doi: 10.1186/s13059-015-0844-5

Gehlenborg, N. (2019). *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. Available online at: https://CRAN.R-project.org/package=UpSetR (accessed June 8, 2020).

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi: 10.1038/nbt.4091

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.10480998

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., De Leeuw, Y., Anavy, L., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17, 77–77. doi: 10.1186/s13059-016-0938-8

Hicks, S. C., Teng, M., and Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *BioRxiv*[Preprint] 025528. doi: 10.1101/025528

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127. doi: 10.1093/biostatistics/kxj037

Kharchenko, P., and Fan, J. (2019). *scde: Single Cell Differential Expression*. Available online at: http://pklab.med.harvard.edu/scde (accessed June 8, 2020).

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005

Krieg, C., Nowicka, M., Guglietta, S., Schindler, S., Hartmann, F. J., Weber, L. M., et al. (2018). High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat. Med.* 24:144. doi: 10.1038/nm.4466

Li, Q., Cheng, Z., Zhou, L., Darmanis, S., Neff, N. F., Okamoto, J., et al. (2019). Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell RNA sequencing. *Neuron* 101, 207–223.e10. doi: 10.1016/j.neuron.2018.12.006

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

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15:e8746. doi: 10.15252/msb.20188746

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

McDavid, A., Finak, G., and Yajima, M. (2019). *MAST: Model-based Analysis of Single Cell Transcriptomics*. Available online at: https://github.com/RGLab/MAST/ (accessed June 8, 2020).

McEvoy, J., Flores-Otero, J., Zhang, J., Nemeth, K., Brennan, R., Bradley, C., et al. (2011). Coexpression of normally incompatible developmental pathways in retinoblastoma genesis. *Cancer Cell* 20, 260–275. doi: 10.1016/j.ccr.2011.07.005

Mehtonen, J., Teppo, S., Lahnalampi, M., Kokko, A., Kaukonen, R., Oksa, L., et al. (2020). Single cell characterization of B-lymphoid differentiation and leukemic cell states during chemotherapy in ETV6-RUNX1 positive pediatric leukemia identifies drug-targetable transcription factor activities. *bioRxiv*[Preprint] doi: 10.1186/s13073-020-00799-2

Morgan, M., Obenchain, V., Lang, M., Thompson, R., and Turaga, N. (2019). *BiocParallel: Bioconductor Facilities for Parallel Evaluation*. Available online at: https://github.com/Bioconductor/BiocParallel (accessed June 8, 2020).

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18, 35–45. doi: 10.1038/nri.2017.76

Rider, P. R. (1955). Truncated binomial and negative binomial distributions. *J. Am. Stat. Assoc.* 50, 877–883. doi: 10.1080/01621459.1955.10501973

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9:284. doi: 10.1038/s41467-017-02554-5

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

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

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

Roerink, S. F., Sasaki, N., Lee-Six, H., Young, M. D., Alexandrov, L. B., Behjati, S., et al. (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* 556, 457–462. doi: 10.1038/s41586-018-0024-3

Santos Nobre, J., and da Motta Singer, J. (2007). Residual analysis for linear mixed models. *Biom. J. J. Math. Methods Biosci.* 49, 863–875. doi: 10.1002/bimj.200610341

Shao, F., Wang, Y., Zhao, Y., and Yang, S. (2019). Identifying and exploiting gene-pathway interactions from RNA-seq data for binary phenotype. *BMC Genet.* 20:36. doi: 10.1186/s12863-019-0739-7

Somekh, J., Shenorr, S. S., and Kohane, I. S. (2019). Batch correction evaluation framework using a-priori gene-gene associations: applied to the GTEx dataset. *BMC Bioinformatics* 20:268. doi: 10.1186/s12859-019-2855-9

Soneson, C., and Robinson, M. D. (2016). iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat. Methods* 13:283. doi: 10.1038/nmeth.3805

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol.* 37, 334–344. doi: 10.1002/gepi.21717

Sun, S., Zhu, J., Mozaffari, S., Ober, C., Chen, M., and Zhou, X. (2018). Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. *Bioinformatics* 35, 487–496. doi: 10.1093/bioinformatics/bty644

Sun, X., Sun, S., and Yang, S. (2019). An efficient and flexible method for deconvoluting bulk RNA-Seq data with single-cell RNA-seq data. *Cells* 8:1161. doi: 10.3390/cells8101161

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C. C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Tung, P., Blischak, J. D., Hsiao, C. J., Knowles, D., Burnett, J. E., Pritchard, J. K., et al. (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* 7, 39921–39921. doi: 10.1038/srep39921

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lonnerberg, P., Lou, D., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18, 145–153. doi: 10.1038/nn.3881

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., et al. (2018). Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19:24. doi: 10.1186/s13059-018-1406-4

Van den Berge, K., Soneson, C., Love, M. I., Robinson, M. D., and Clement, L. (2017). zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications. *bioRxiv*[Preprint] doi: 10.1101/157982

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

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

Yang, S., Shao, F., Duan, W., Zhao, Y., and Chen, F. (2017). Variance component testing for identifying differentially expressed genes in RNA-seq data. *PeerJ* 5:e3797. doi: 10.7717/peerj.3797

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18:174. doi: 10.1186/s13059-017-1305-0

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi: 10.1126/science.aaa1934

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: 08 March 2021
doi: 10.3389/fgene.2021.637322

Leveraging Single-Cell RNA-seq Data to Uncover the Association Between Cell Type and Chronic Liver Diseases

Xiangyu Ye[1†], *Julong Wei*[2†], *Ming Yue*[3], *Yan Wang*[1], *Hongbo Chen*[4], *Yongfeng Zhang*[4], *Yifan Wang*[4], *Meiling Zhang*[4], *Peng Huang*[1*] *and Rongbin Yu*[1*]

[1] *Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China,* [2] *Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, United States,* [3] *Department of Infectious Diseases, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China,* [4] *Department of Infectious Disease, Jurong Hospital Affiliated to Jiangsu University, Jurong, China*

Background: Components of liver microenvironment is complex, which makes it difficult to clarify pathogenesis of chronic liver diseases (CLD). Genome-wide association studies (GWASs) have greatly revealed the role of host genetic background in CLD pathogenesis and prognosis, while single-cell RNA sequencing (scRNA-seq) enables interrogation of the cellular diversity and function of liver tissue at unprecedented resolution. Here, we made integrative analysis on the GWAS and scRNA-seq data of CLD to uncover CLD-related cell types and provide clues for understanding on the pathogenesis.

Methods: We downloaded three GWAS summary data and three scRNA-seq data on CLD. After defining the cell types for each scRNA-seq data, we used *RolyPoly* and *LDSC-cts* to integrate the GWAS and scRNA-seq. In addition, we analyzed one scRNA-seq data without association to CLD to validate the specificity of our findings.

Results: After processing the scRNA-seq data, we obtain about 19,002–32,200 cells and identified 10–17 cell types. For the HCC analysis, we identified the association between B cell and HCC in two datasets. *RolyPoly* also identified the association, when we integrated the two scRNA-seq datasets. In addition, we also identified natural killer (NK) cell as HCC-associated cell type in one dataset. In specificity analysis, we identified no significant cell type associated with HCC. As for the cirrhosis analysis, we obtained no significant related cell type.

Conclusion: In this integrative analysis, we identified B cell and NK cell as HCC-related cell type. More attention and verification should be paid to them in future research.

Keywords: chronic liver diseases, GWAS, scRNA-seq, integrated analysis, cell type

OPEN ACCESS

Edited by:
Shiquan Sun,
Xi'an Jiaotong University, China

Reviewed by:
Xingjie Hao,
Huazhong University of Science and Technology, China
Zhongshang Yuan,
Shandong University, China

****Correspondence:***
Peng Huang
huangpeng@njmu.edu.cn
Rongbin Yu
rongbinyu@njmu.edu.cn

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

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *03 December 2020*
Accepted: *25 January 2021*
Published: *08 March 2021*

Citation:
Ye X, Wei J, Yue M, Wang Y, Chen H, Zhang Y, Wang Y, Zhang M, Huang P and Yu R (2021) Leveraging Single-Cell RNA-seq Data to Uncover the Association Between Cell Type and Chronic Liver Diseases. Front. Genet. 12:637322. doi: 10.3389/fgene.2021.637322

INTRODUCTION

Chronic liver disease (CLD) is a public health topic of global concern. As estimated, about 844 million people worldwide are suffering from CLD and 2 million deaths each year (Asrani et al., 2019). Starting with diverse etiology-related chronic hepatitis, CLD might develop into cirrhosis and hepatocellular carcinoma after repetitive liver damage (Gadd et al., 2020). Environment risk factors associated with CLD are virus, diet, drug, and autoimmune (Marcellin and Kutala, 2018). With the development of molecular biology, the role of host genetic background in CLD has also gained wide attention (Anstee et al., 2020). Genome-wide association studies (GWASs) have contributed greatly to our understanding of the genetic roles in CLD pathogenesis and prognosis (Matsuura et al., 2017). A number of associated polymorphisms, including variants on *CDK14*, *SH2B3*, *CARD10*, *TLL1*, *PNPLA3*, and *HLA*, have been reported (De Boer et al., 2014; Sudlow et al., 2015; Matsuura et al., 2017; Nicoletti et al., 2017; Li et al., 2018; Ishigaki et al., 2020; Schwantes-An et al., 2020). Nevertheless, the current understanding of CLD is far from enough, and it is still of great significance to further clarify the pathological process of CLD and explore new treatment strategy for CLD patients (Marcellin and Kutala, 2018).

As the largest internal organ of the body, the liver consists of many cell types, including not only epithelial cells and some non-parenchymal cells (e.g., endothelial and mesenchymal cells) but also a variety of immune cells (MacParland et al., 2018; Aizarani et al., 2019; Ramachandran et al., 2019; Sharma et al., 2020). Different cell types vary greatly in abundance and function, leading to their completely distinct roles in the physiological and pathophysiological processes of liver diseases (Ramachandran et al., 2020). Single-cell genomics technologies are transforming our understanding on diseases like CLD, enabling interrogation of cellular diversity and function at unprecedented resolution, and adding a new dimension to traditional bulk transcriptomic techniques (Giladi and Amit, 2018). Single-cell RNA sequencing (scRNA-seq) has been used to feature the fundamental liver biology and the cellular mechanisms underpinning liver regeneration (Aizarani et al., 2019). It also has been used to uncover the pathophysiological changes of hepatic fibrosis and hepatocellular carcinoma, where the heterogeneity and changes of T cells (Zheng C. et al., 2017), macrophages (Ramachandran et al., 2019), and endothelial cells (Sharma et al., 2020) residing within the liver tissue may be critical in driving disease states.

Both GWAS and scRNA-seq have thrown light on the way to indepthly understand the pathogenesis of CLD and further laid a foundation for the development of precision treatment strategy (Saviano et al., 2020). Integrating GWAS summary data and scRNA-seq data to identify the cell types associated to CLD might provide new clues for understanding the pathogenesis of CLD (Calderon et al., 2017; Finucane et al., 2018; Hao et al., 2020). Here, we used *RolyPoly* and *LDSC-cts* to ensure the robustness and confidence of the result. Especially, we first processed the scRNA-seq data to derive averaged expression vector and differential expression gene (DEG) list of each cell type for *RolyPoly* and *LDSC-cts*, respectively. Then, we used the Ensembl database to obtain the position relationship between SNPs and gene (Yates et al., 2019). Finally, with GWAS data, scRNA-seq data and block annotation in place, as well as accounting for linkage disequilibrium (LD) of related population, we applied *RolyPoly* and *LDSC-cts* to identify and prioritize CLD-relevant cell types.

MATERIALS AND METHODS

Genome-Wide Association Studies Data

The first category of summary statistics is Asian ancestry GWAS. The datasets are from the Biobank of Japan (BBJ)[1] (Ishigaki et al., 2020). We focus on the CLD-related phenotype that contain allele information and variant ID and that contain effect size and its standard error. With the two criteria, we obtained two GWAS summary statistics: cirrhosis (n = 212,453, prevalence = 1.03%) and HCC (n = 197,611, prevalence = 0.94%). Here, cirrhosis and HCC in BBJ were adjusted for age, sex, and top five genotype PCs (Ishigaki et al., 2020). The details of the two GWAS data are provided in **Supplementary Table 1**. Based on Asian ancestry from the 1000 Genome Project (1000 GP), we filtered out variants with minor allele frequency (MAF) < 0.01 and Hardy–Weinberg equilibrium (HWE) $< 10^{-6}$ (Auton et al., 2015). After these quality control (QC) steps, we finally obtained 7,246,475 and 7,246,543 SNPs from the two datasets.

The second category of GWAS summary statistics is from European ancestry. The dataset is from GeneATLAS website[2] (Canela-Xandri et al., 2018). We focus on the CLD-related phenotype that contain allele information and variant ID and that contain effect size and standard error. With the two criteria, we obtain one GWAS summary statistics: cirrhosis (n = 452,264, prevalence = 1.99%). This cirrhosis GWAS data was adjusted for sex, array batch, UK Biobank Assessment Center, age, age2 (Sudlow et al., 2015), and the top 20 genotype PCs as computed by UK Biobank. The details of these data are also provided in **Supplementary Table 1**. Based on European ancestry from the 1000 Genome Project, we filtered out variants with MAF < 0.01 and HWE $< 10^{-6}$ (Auton et al., 2015). After these QC steps, we finally obtained 7,636,847 SNPs from this dataset.

We treated the phase 3 of the 1000 Genome Project as the reference panel (Auton et al., 2015). Here, we collected 503 European individuals and 504 East Asian individuals with 81,271,745 SNPs. We used *PLINK* to calculate Pearson's r^2 values of pairwise SNPs for *RolyPoly* with the default 1 MB window size (Chang et al., 2015). In *LDSC-cts*, we set the window size to 1 centiMorgan to estimate LD scores (Finucane et al., 2018).

Four Single-Cell Data

Considering the cirrhosis and HCC data acquired from GWAS, we searched the GEO database for related scRNA-seq data and obtained one data for liver cirrhosis and two for HCC, whose raw counts data are available (Barrett et al., 2012;

[1] http://jenger.riken.jp/en/

[2] http://geneatlas.roslin.ed.ac.uk/

Ramachandran et al., 2019; Losic et al., 2020). In addition, to verify the specificity of the outcomes, we also downloaded an idiopathic Parkinson's disease (IPD) data. The details are provided in **Supplementary Table 2**. Following the original study, we performed QC and clustering for each scRNA-seq data. Note that scRNA-seq data usually have the potential to have its clusters continuously subdivided, but we just controlled the cell type number of each data within 10–20 depending on the features and quality of each data. The specific processing details of each data are as follows: After demultiplexing, aligning, and estimating cell-containing partitions and associated UMIs, a cirrhosis dataset (GSE136103) consisting of CD45 + and CD45-, blood and liver, healthy and cirrhosis, and human and mice samples were downloaded (Ramachandran et al., 2019). Here, we only chose nine human cirrhotic samples, including five CD45 + and four CD45- samples, for downstream analysis.

For scRNA-seq data analysis, we first removed potential doublets, and then excluded the cells that expressed fewer than 300 genes or mitochondrial gene content >30% of the total UMI count (Ramachandran et al., 2019). We also excluded genes expressed in fewer than three cells. We followed the analysis flow in *Seurat* (Stuart et al., 2019): (1) used *SCTransform*, a new strategy to remove the influence of technical characteristics while preserving biological heterogeneity via regularized negative binomial regression, to normalize and scale scRNA-seq data (Hafemeister and Satija, 2019); (2) used default setting of *IntegrateData* to remove the batch effect (Butler et al., 2018); (3) performed unsupervised clustering and differential gene expression analyses on the integrated data; (4) used principal component analysis (PCA) for linear dimension reduction, and then used shared nearest neighbor (SNN) graph-based clustering, in which the graph was constructed using the top 30 principal components; and (5) used UMAP to visualize by the same number of principal components (PCs) as the associated clustering, with perplexity ranging from 30 to 300 according to the number of cells in the dataset or lineage. The details of data processing are shown in **Figure 1**.

In cell type definition, we referred to marker genes that are widely recognized and those from the original research. We used *BuildClusterTree* to assess cluster similarity by constructing the phylogenetic tree (Stuart et al., 2019). Totally, we identified 20 clusters on 23,184 cells (**Supplementary Table 2** and **Figure 2**). Marker genes used for cell type definition are shown in **Supplementary Table 3**.

The first HCC dataset (GSE149614) contains 21 primary tumor, portal vein tumor thrombus (PVTT), metastatic lymph node, and non-tumor liver samples from 10 HCC patients. We downloaded the raw count data, which have been

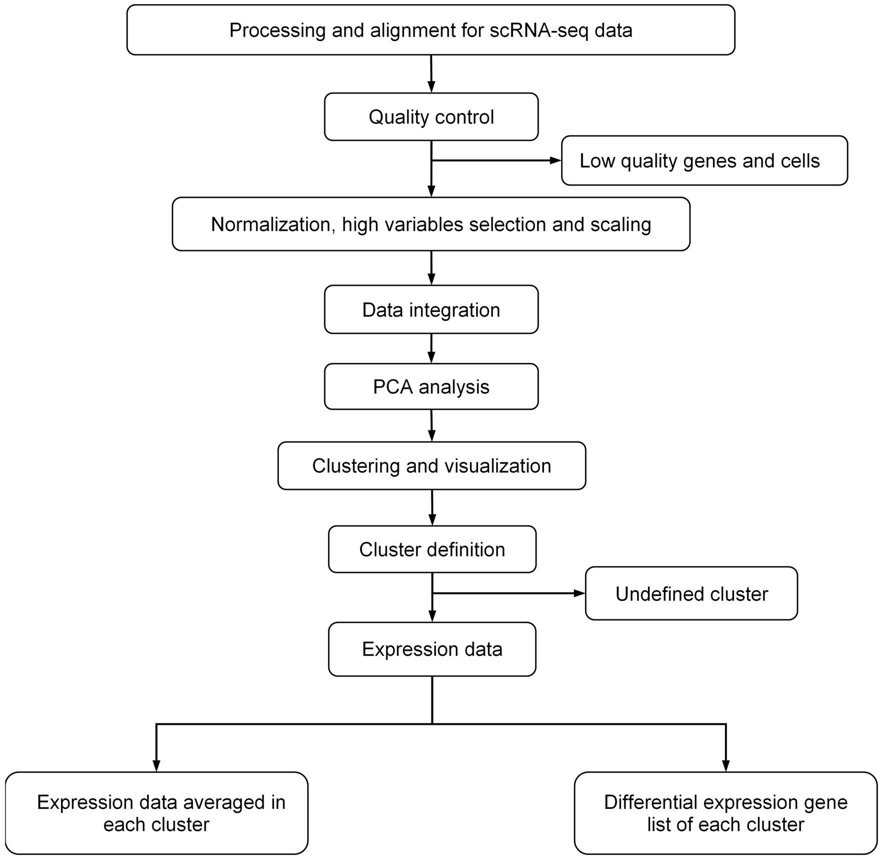
Processing and alignment for scRNA-seq data
Quality control
Low quality genes and cells
Normalization, high variables selection and scaling
Data integration
PCA analysis
Clustering and visualization
Cluster definition
Undefined cluster
Expression data
Expression data averaged in each cluster
Differential expression gene list of each cluster

FIGURE 1 | General procedure for scRNA-data processing.

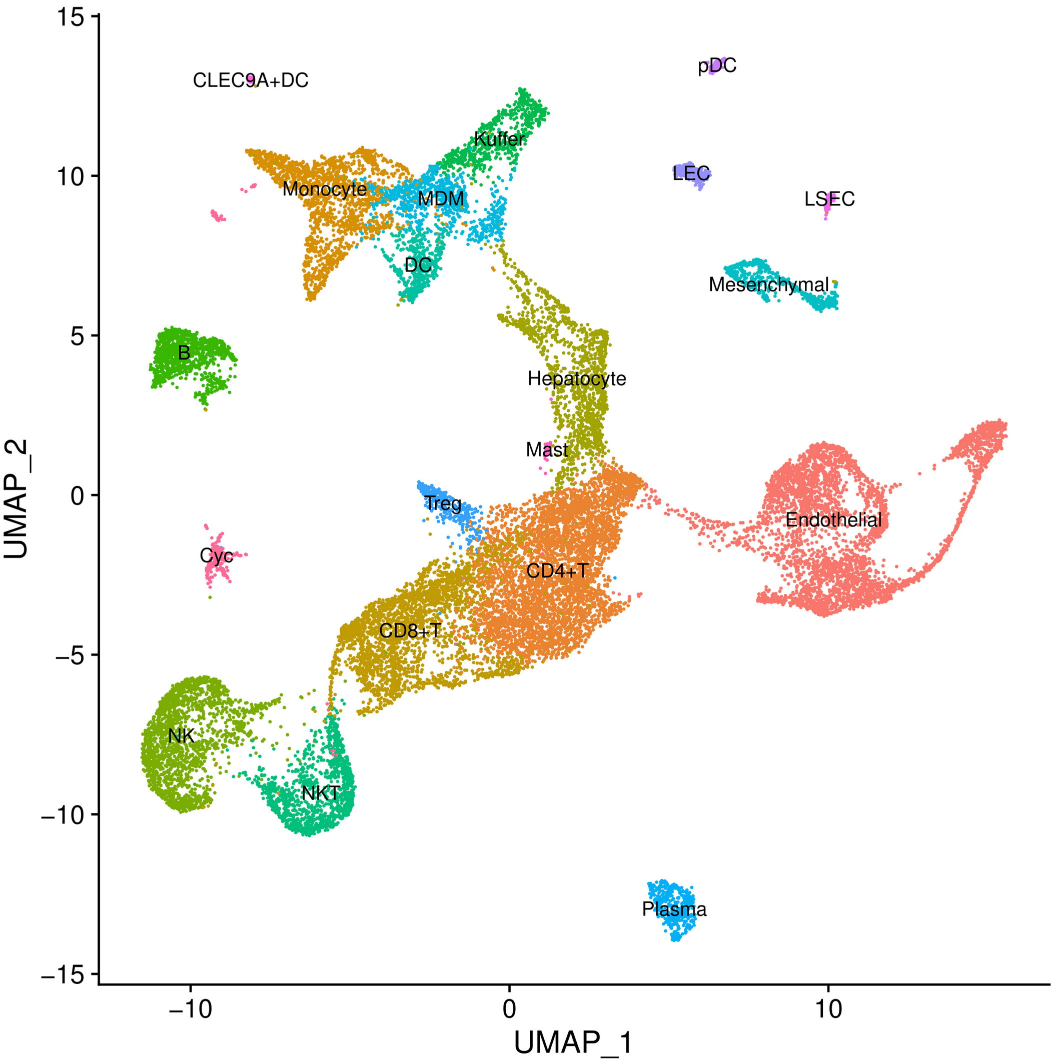

FIGURE 2 | Cell types inferred from expression of marker gene signatures in GSE136103. NKT, natural killer T cells; Pdc, plasmacytoid dendritic cell; Treg, regulatory T cell; LEC, lymphatic endothelial cell; MDM, monocyte-derived macrophage; NK, natural killer cell; LSEC, liver sinusoids endothelial cell; DC, dendritic cell.

processed and aligned by Cell Ranger, and chose only 10 primary tumor samples for downstream analysis (Zheng G.X.Y. et al., 2017). After processing and clustering, we totally identified 14 cell types on 30,983 cells in this dataset (**Supplementary Table 2**).

Another HCC dataset (GSE112271) contains three and four tumor samples coming from different regions of two different individuals, and we included all seven samples for downstream analysis. After data processing, we totally identified 13 clusters on 32,200 cells in this dataset (Losic et al., 2020; **Supplementary Table 2**).

We downloaded the processed and aligned IPD dataset (GSE157783), which contains samples from six control and five idiopathic Parkinson's disease cases. We chose only five disease samples for downstream analysis and totally identified 12 clusters on 19,002 cells following our procedure (**Supplementary Table 2**).

Defining the Specific Cell Types Associated With Cirrhosis and HCC

We used *RolyPoly* and *LDSC-cts* to define the specific cell types associated with cirrhosis and HCC (Calderon et al., 2017; Finucane et al., 2018). Based on polygenic model, *RolyPoly* treats the variance of each gene as the linear combination of each cell type and estimates the coefficients by method-of-moment. Then, *RolyPoly* uses block bootstrap to estimate the variance for the cell type effects, then construct t-statistics to test them (Efron and Tibshirani, 1986). By utilizing GWAS summary statistics for all SNPs near protein-coding genes, the model performed joint analysis with gene expression of a variety of cell types simultaneously, to define prioritized trait-relevant cell types (Calderon et al., 2017). We extracted the log-normalized matrix from each processed data and averaged the expression across each identified cell-type classes. We also scaled the expression data, and then took the absolute expression values, so as to form the input of *RolyPoly* (Calderon et al., 2017). We referred to the Ensembl database (GRCh37) and defined a 10-kb window center around the transcription start site (TSS) of a gene as its transcribed region, to construct a block annotation as recommended that could link the location of GWAS variants with related genes. Of note, we only retained genes on autosomes (Calderon et al., 2017). We used the default parameters and set 1,000 times bootstrap to obtain robust standard errors.

Based on partition heritability, *LDSC-cts* needs the top upregulated genes list of each cell type rather than the expression data (Finucane et al., 2018). Here, we used Wilcoxon rank sum test embedded in *Seurat* to find the DEGs for each cell type with all remaining clusters as control. Following Finucane et al. (2018), we extracted the top 10% upregulated genes ranked by *P* value from each cell type. DEGs were identified as genes expressed in at least 0.1% total cells and with log-transformed fold change above 0 in the target cluster under comparison, so as to ensure a sufficient number of genes could be obtained from each cluster. DEGs lists of each scRNA-seq data used for *LDSC-cts* analysis are summarized in **Supplementary Tables 4,8**. We referred to the Ensembl database (GRCh37) and defined the region from the TSS to the transcription end sites (TES) of a gene as its transcribed region (Yates et al., 2019). We also added 100-kb windows on either side of the transcribed region of each gene. Finally, we applied *LDSC-cts* by jointly modeling the annotation that corresponded to each cell type, a common annotation that included all of the genes, and the 52 annotations in the default "baseline model," to identify CLD-specific cell types (Finucane et al., 2018).

We also made a sensitivity analysis. Specifically, we changed the resolution used in clustering to obtain a coarser cell type list for analysis. In particular, since *LDSC-cts* is sensitive to the gene list used for analysis, we simultaneously changed the number of genes included in *LDSC-cts* to the top 5% upregulated ones.

Bonferroni correction was used for multiple tests ($P < 0.1/n$, where $n = 4$ or three is the number of cell type groups, including epithelial cell, non-parenchymal cell, lymphatic immune cell, myeloid immune cell for liver tissue, or gliocyte, neuron, and vascular cell for the brain tissue, **Supplementary Table 9**) (Hao et al., 2020).

Statistical Software

We used *scDblFinder* package (version 1.4.0), *Seurat* package (version 1.4.0), *biomaRt* package (version 2.45.6), and *RolyPoly* package (version 0.1.0) in R software (version 3.6.3) (R Core Team, 2020). We used *PLINK* (version 2.0) (Chang et al., 2015) to analyze GWAS data. We also used *LDSC-cts* (version 1.0.1) in python software (version 2.7.18) (Van Rossum and De Boer, 1991).

RESULTS

HCC Datasets Analysis

For the HCC GWAS data from BBJ, we totally retained 7,246,543 variants with HWE $< 10^{-6}$ and MAF > 0.01, as well as their annotation. For the scRNA-seq data (GSE149614), we identified 14 cell types on 30,983 cells (**Supplementary Table 2** and **Supplementary Figures 1,2**). We further excluded cluster with less than 100 cells (63 mast cells) to avoid the interference of their unstable signal on the results. We also excluded the circulating cluster (2,510 cells), since it usually contains various immune cells from the circulation and may represent a mixed signal. Finally, we retained a total of 28,410 cells from 12 cell types. After integrative analysis, we identified B cell ($\beta = 2.956 \times 10^{-4}$, se $= 1.442 \times 10^{-4}$, $P = 0.0228$) as cell type associated with HCC in *RolyPoly* (**Figure 3**), whereas natural killer cell (NK), monocyte, CD4 + T cell, plasma, macrophage, hepatocyte, regulatory T cell (Treg), endotheliocyte, mesenchymal cell, CD8 + T cell, and dendritic cell (DC) showed no significance ($P > 0.05$). In *LDSC-cts* analysis, we also obtained B cell ($\beta = 2.475 \times 10^{-9}$, se $= 1.116 \times 10^{-9}$, $P = 0.0133$) as the significant cell type.

We used another HCC scRNA-seq data from GEO for verification. Totally, we recognized 12 cell types on 30,931 cells from the GSE112271 data with one circulating (1,192 cells) and one small cluster (77 liver sinusoids endothelial cells) excluded (**Supplementary Table 2** and **Supplementary Figures 3,4**). We identified monocyte-derived macrophage (MDM, $\beta = 1.665 \times 10^{-4}$, se $= 6.098 \times 10^{-5}$, $P = 0.0031$), T cell ($\beta = 1.732 \times 10^{-4}$, se $= 7.170 \times 10^{-5}$, $P = 0.0076$), and natural killer cell (NK, $\beta = 1.458 \times 10^{-4}$, se $= 6.976 \times 10^{-5}$, $P = 0.0191$) as cell types significantly associated with HCC in *RolyPoly* (**Figure 3**), whereas the obtained NK ($\beta = 2.331 \times 10^{-9}$, se $= 1.118 \times 10^{-9}$, $P = 0.0186$) and B cell ($\beta = 2.255 \times 10^{-9}$, se $= 1.134 \times 10^{-9}$, $P = 0.0234$) as the significant cell types in *LDSC-cts* analysis.

We also integrated the two HCC scRNA-seq data and obtained a combined data consisting of 60,120 cells and 13 cell types for further analysis (**Supplementary Figures 5,6**). The *RolyPoly* analysis showed that B cell ($\beta = 2.451 \times 10^{-4}$, se $= 9.240 \times 10^{-5}$, $P = 0.0040$) was significantly associated with HCC (**Figure 3**), whereas the *LDSC-cts* identified no significant cell type.

HCC Dataset Specificity and Sensitivity Analysis

We used scRNA-seq data from other disease to verify the specificity of our findings. To be specific, we downloaded one IPD (GSE157783) scRNA-seq data, and identified 12 cell types on 19,002 cells (**Supplementary Table 2** and **Supplementary Figures 7,8**). After excluding clusters with too few cells (47 fibroblasts and 26 T cells), we identified no cell type significantly associated with HCC in either *RolyPoly* or *LDSC-cts* analysis (**Figure 4**).

We also made a sensitivity analysis by changing the resolution used in clustering and got nine, eight, and nine cell types for GSE149614, GSE112271, and their integrated data, respectively. Sensitivity analysis showed that B cell was still significantly associated with HCC in *RolyPoly* analysis on GSE149614 and the integrated data, as well as in *LDSC-cts* analysis on the integrated data. It also showed nominal significance ($P < 0.1$) in *LDSC-cts* analysis on GSE112271, and was the top cell type ($P = 0.119$) in the analysis on GSE149614 (**Supplementary Figure 9**).

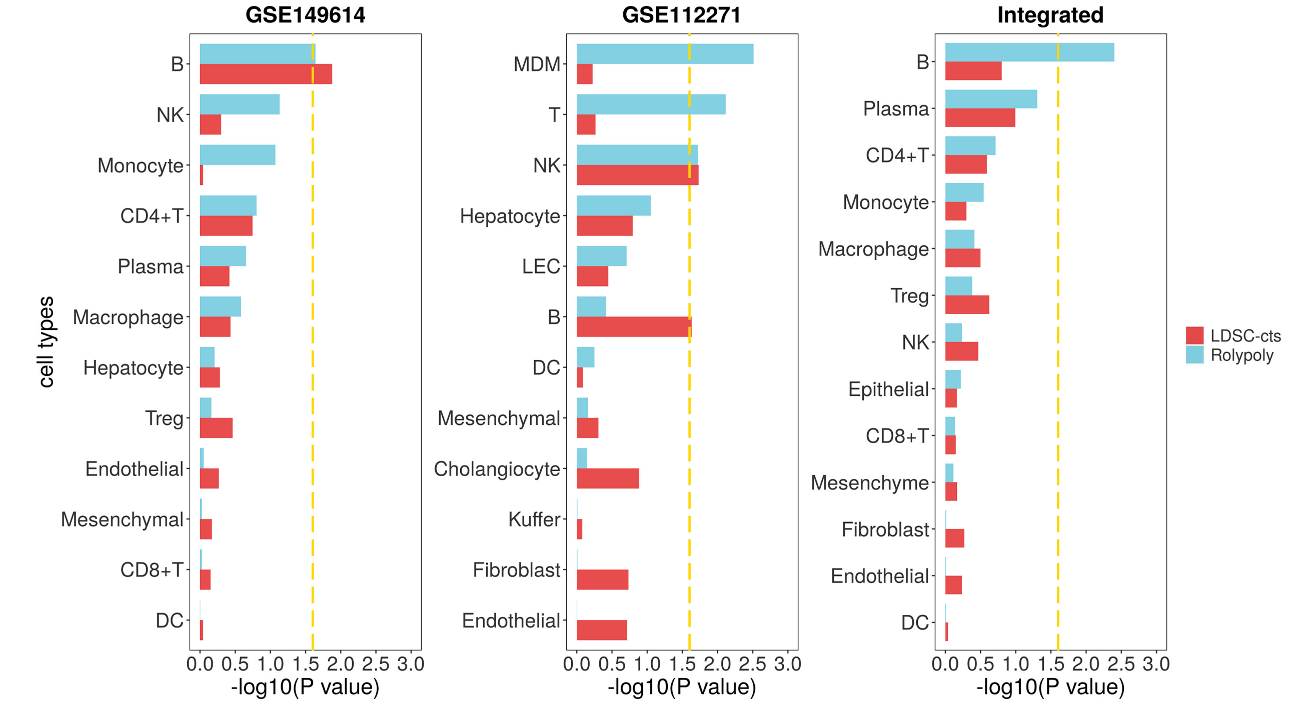

FIGURE 3 | Association between HCC scRNA-seq data and HCC GWAS data from *RolyPoly* and *LDSC-cts*. Using *RolyPoly* and *LDSC-cts* to detect the association of cell types in GSE149614, GSE112271, and their combined HCC scRNA-seq data with EAS HCC GWAS data. Dashed lines in each panel represent a threshold of *P* = 0.1/4. NK, natural killer cell; Treg, regulatory T cell; DC, dendritic cell; MDM, monocyte-derived macrophage; LEC, lymphatic endothelial cell.

Cirrhosis Data Analysis

For the cirrhosis GWAS data from BBJ of East Asian population, we totally retained 7,246,475 variants with their annotation. For the scRNA-seq data (GSE136103), we identified 20 cell types on 23,184 cells (**Supplementary Table 2**; **Figure 2**; **Supplementary Figure 10**), but further excluded circulating cluster (309 cells) and clusters with less than 100 cells (56 CLEC9A + dendritic cells and 31 mast cells). Finally, we retained a gene expression data of 17 cell types. *RolyPoly* showed that CD4 + T cell ($\beta = 2.278 \times 10^{-4}$, se = 1.149×10^{-4}, $P = 0.0259$) was significantly associated with cirrhosis, whereas *LDSC-cts* identified no significant cell type (**Figure 5**).

We also used a cirrhosis GWAS summary data of European population from GeneATLAS website to verify the stability of our outcomes, in which a total of 7,636,847 variants was retained after QC. We identified natural killer T cell (NKT, $\beta = 6.535 \times 10^{-10}$, se = 2.423×10^{-10}–1.110×10^{-9}, $P = 0.0038$) and hepatocyte ($\beta = 2.891 \times 10^{-10}$, se = 1.364×10^{-10}, $P = 0.0149$) as cell types significantly associated with cirrhosis in *RolyPoly*, while we obtained no significant cell type in the *LDSC-cts* analysis (**Figure 5**).

DISCUSSION

Identifying disease-specific cell types has important implications to understand the mechanisms of disease, to guide research, and to develop more precise therapies (Calderon et al., 2017). In this study, using two separate methods and based on available data, we explored the CLD-related cell types through an integrative analysis on GWAS and scRNA-seq data.

In the analysis of HCC, both *RolyPoly* and *LDSC-cts* identified B cell as significant associated with HCC ($P = 0.0228$ and $P = 0.0133$, respectively). B cell mainly exerts its humoral immunity function through the antibody production and antigen presentation, and can also regulate T cells and innate immune responses (Tsou et al., 2016). Recently, the regulation role of resident B cell in tumor has been investigated (Garaud et al., 2018; Lechner et al., 2019; Wang et al., 2019). The balance between B cells in different states and their activities may have the potential to affect pro- or anti-tumor functions (Largeot et al., 2019; Liu et al., 2019). A similar phenomenon has also been observed in liver disease. In a Hras12V HCC mouse models, B cells were found to have a potential role in suppressing hepatic tumorigenesis (Wang et al., 2017), whereas in another mouse model with inflammation-associated HCC, infiltrating B cells was correlated with increased tumor aggressiveness and mortality (Faggioli et al., 2018). In addition, activated FcγRII$^{low/-}$ B cells from HCC tumor may also suppress host anti-tumor immune response via IL-10 signals (Ouyang et al., 2016; Jin et al., 2017). Nevertheless, the depth of research on tumor-associated B cells and their subsets is far less than that of T cells. As for the liver diseases, existing several unbiased scRNAseq research on CLD have not revealed major alterations in the composition or transcriptional profile of liver B cells in disease state (MacParland et al., 2018; Ramachandran

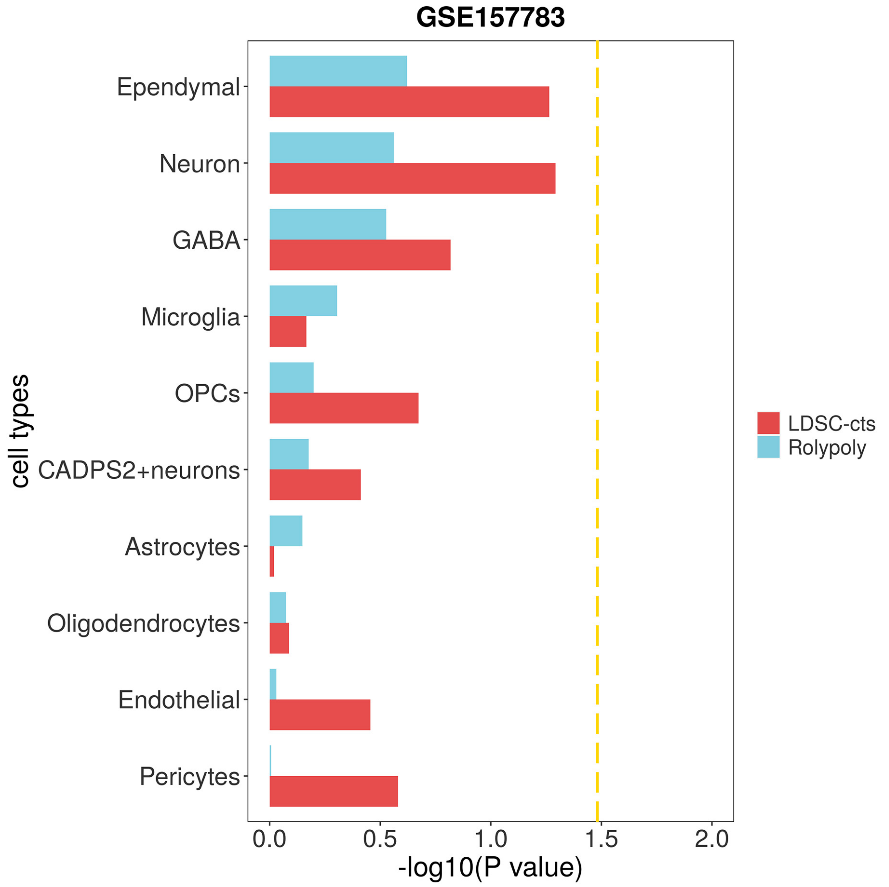

FIGURE 4 | Association of IPD scRNA-seq data with HCC GWAS data form *RolyPoly* and *LDSC-cts*. Using *RolyPoly* and *LDSC-cts* to detect the association of cell types in GSE157783 IPD scRNA-seq data with EAS HCC GWAS data. Dashed lines in each panel represent a threshold of *P* = 0.1/3. OPC, Oligodendrocyte precursor cell; GABA, GABAergic neurons.

et al., 2019; Losic et al., 2020; Sharma et al., 2020). Separate single-cell research has not been conducted specifically on the relationship between B cells and liver disease. However, with the development of single-cell technology, the combination of single-cell transcriptomics and immunomics (B cell receptor) is expected to further reveal the exact role of B cells in HCC and other CLD, and explore B cell-based immunotherapy (Setliff et al., 2019).

We also used another HCC-related scRNA-seq data to verify our findings. *RolyPoly* identified MDM, T cell, and NK cell, rather than B cell, as significant cell types, whereas B cell remained significant together with NK cell in *LDSC-cts* analysis. This might have resulted from *LDSC-cts* using DEGs, which may be conserved but more robust among different studies for a specific disease. Although we have averaged the expression for each identified cell type and taken a scale on the averaged data, differences in data structure arising from the different angles of the two original studies may also be a probable interpretation (Losic et al., 2020). Therefore, we further integrated the two data and repeated these analyses, and found that B cell regained its significance in the integrated data under *RolyPoly* method. In addition, we used the IPD scRNA-seq data (GSE157783) from brain tissue to make specificity analysis, and found that neither *RolyPoly* nor *LDSC-cts* method identified significant cell types. The above results jointly indicated that B cells may be a significant cell type for HCC, and more attention should be paid to them in future research.

Of note, outcomes from the second HCC data also suggested that NK cells might be HCC-related cells, which was significant in both *RolyPoly* and *LDSC-cts* analysis. Although this result has not been verified in our analysis, a previous study has identified the contribution of NK cell in liver injury (Luci et al., 2019), NK cell composition alteration and an interaction with other clusters was also observed in HCC (Zhang et al., 2019). Thus, it is also of meaning to further explore the relationship between NK cell and HCC.

As for the analysis on cirrhosis, we have not obtained an overlap cell type in the two methods, with CD4 + T cell significant in *RolyPoly* analysis using the GWAS data on East Asian population, while NKT and hepatocyte are significant in *RolyPoly* analysis on European population. That might be caused by the different linkage disequilibrium and minor allele frequency (MAF) for different ancestry, cross-population correlations of

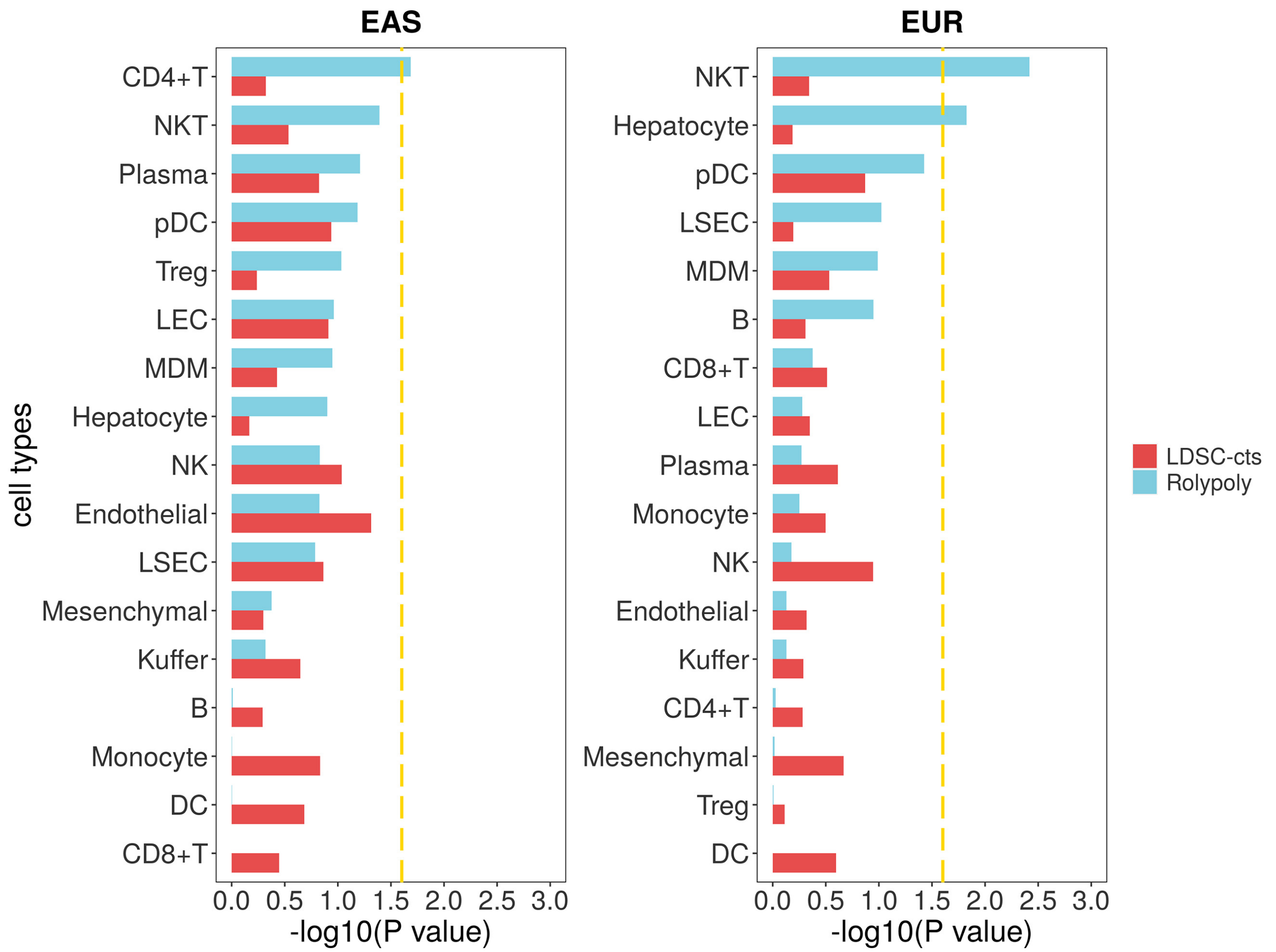

FIGURE 5 | Association between cirrhosis scRNA-seq data and cirrhosis GWAS data from *RolyPoly* and *LDSC-cts*. Using *RolyPoly* and *LDSC-cts* to detect the association of cell types in GSE136103 cirrhosis scRNA-seq data with EAS and EUR cirrhosis GWAS data. Dashed lines in each panel represent a threshold of *P* = 0.1/4. NKT, natural killer T cell; pDC, plasmacytoid dendritic cell; Treg, regulatory T cell; LEC, lymphatic endothelial cell; MDM, monocyte-derived macrophage; NK, natural killer cell; LSEC, liver sinusoids endothelial cell; DC, dendritic cell.

causal SNP effects, and heritability (Mather and Thalamuthu, 2020; Wang et al., 2020; Yang and Zhou, 2020). For example, there are 1,558 SNPs and 76 SNPs with $P < 10^{-6}$ in EAS and EUR datasets, respectively (**Supplementary Table 10**).

Certainly, several limitations remain in our study. First, all data used came from public databases, and external experiments were not conducted to verify our findings; but alternatively, we used other available GWAS and scRNA-seq data to make verification as well as specificity analysis, which would also ensure the reliability of our results to some extent. Second, *SCTransform* is a relative powerful normalization method, which may weaken the heterogeneity among samples when used for integration (Butler et al., 2018; Tran et al., 2020). Since we were aimed to apply similar cell type definition strategy in different samples and focused mainly on the similarity rather than heterogeneity, it may offer more help than interference to our analysis. In addition, since current research advances have limited ability in cell type definition and explanation, we applied a relative conservation cell subdivided strategy in the current study. With the in-depth research on various cell subtypes and the development of single-cell technology, similar research is expected be carried out in a larger sample with a higher resolution and precision, and more novel findings with biological explanation would be obtained.

In summary, we performed integrative analysis on GWAS summary data and single scRNA-seq data of CLD, and identified B cell as a potential HCC-related cell type. Since we have made verification from multiple angles, our outcomes are of relative reliability. In addition, as the single-cell atlas of different tissues and diseases has been completed, more targeted researches are expected, and our study would provide valuable clues for further research on CLD.

CODE AVAILABILITY

Code used for data processing, integrated analysis, and plotting could be found at (https://github.com/XiangyuYe/CLD-specific-celltype-identification).

DATA AVAILABILITY STATEMENT

Asian ancestry CLD GWAS summary data was downloaded from BBJ (http://jenger.riken.jp/en/). European ancestry CLD GWAS summary data was downloaded from GeneATLAS website (http://geneatlas.roslin.ed.ac.uk/). ScRNA-seq datasets used (GSE136103, GSE149614, GSE112271, and GSE157783) were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

RBY and PH designed the study. JLW, YFW, and MLZ performed the datasets quality control. XYY and YW performed the data analysis. PH, HBC, and YFZ interpreted the analysis results. XYY and JLW wrote the draft manuscript. RBY, PH, and MY revised the article. All authors accepted the final manuscript.

FUNDING

This work was supported by the Natural Science Foundation of Jiangsu Province (Nos. BK20171054 and BK20190106), National Natural Science Foundation of China (No. 81703273), Key Research and Development Program of Zhenjiang, China (Nos. SH2019035 and SH2018058), and People's Livelihood, Science and Technology Program of Jurong, China (Nos. SF2018453282, SF2019758956, and 2020SA00107).

ACKNOWLEDGMENTS

We acknowledge the participants and investigators of Biobank Japan Project and GeneATLAS for making the summary data publicly available for us.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.637322/full#supplementary-material

REFERENCES

Aizarani, N., Saviano, A., Sagar, Mailly, L., Durand, S., Herman, J. S., et al. (2019). A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* 572, 199–204. doi: 10.1038/s41586-019-1373-2

Anstee, Q. M., Darlay, R., Cockell, S., Meroni, M., Govaere, O., Tiniakos, D., et al. (2020). Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort(☆). *J. Hepatol.* 73, 505–515. doi: 10.1016/j.jhep.2020.04.003

Asrani, S. K., Devarbhavi, H., Eaton, J., and Kamath, P. S. (2019). Burden of liver diseases in the world. *J. Hepatol.* 70, 151–171. doi: 10.1016/j.jhep.2018.09.014

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

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

Calderon, D., Bhaskar, A., Knowles, D. A., Golan, D., Raj, T., Fu, A. Q., et al. (2017). Inferring relevant cell types for complex traits by using single-cell gene expression. *Am. J. Hum. Genet.* 101, 686–699. doi: 10.1016/j.ajhg.2017.09.009

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. *Nat. Genet.* 50, 1593–1599. doi: 10.1038/s41588-018-0248-z

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

De Boer, Y. S., Van Gerven, N. M. F., Zwiers, A., Verwer, B. J., Van Hoek, B., Van Erpecum, K. J., et al. (2014). Genome-Wide association study identifies variants associated with autoimmune hepatitis type 1. *Gastroenterology* 147, 443.e–452.e. doi: 10.1053/j.gastro.2014.04.022

Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1, 54–75. doi: 10.1214/ss/1177013815

Faggioli, F., Palagano, E., Di Tommaso, L., Donadon, M., Marrella, V., Recordati, C., et al. (2018). B lymphocytes limit senescence-driven fibrosis resolution and favor hepatocarcinogenesis in mouse liver injury. *Hepatology* 67, 1970–1985. doi: 10.1002/hep.29636

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* 50, 621–629. doi: 10.1038/s41588-018-0081-4

Gadd, V. L., Aleksieva, N., and Forbes, S. J. (2020). Epithelial plasticity during liver injury and regeneration. *Cell Stem Cell* 27, 557–573. doi: 10.1016/j.stem.2020.08.016

Garaud, S., Zayakin, P., Buisseret, L., Rulle, U., Silina, K., De Wind, A., et al. (2018). Antigen specificity and clinical significance of IgG and IgA autoantibodies produced in situ by tumor-infiltrating B cells in breast cancer. *Front. Immunol.* 9:2660. doi: 10.3389/fimmu.2018.02660

Giladi, A., and Amit, I. (2018). Single-Cell genomics: a stepping stone for future immunology discoveries. *Cell* 172, 14–21. doi: 10.1016/j.cell.2017.11.011

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 20:296. doi: 10.1186/s13059-019-1874-1

Hao, X., Wang, K., Dai, C., Ding, Z., Yang, W., Wang, C., et al. (2020). Integrative analysis of scRNA-seq and GWAS data pinpoints periportal hepatocytes as the relevant liver cell types for blood lipids. *Hum. Mol. Genet.* 29, 3145–3153. doi: 10.1093/hmg/ddaa188

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

Jin, Y., Lang, C., Tang, J., Geng, J., Song, H. K., Sun, Z., et al. (2017). CXCR5+CD8+ T cells could induce the death of tumor cells in HBV-related hepatocellular carcinoma. *Int. Immunopharmacol.* 53, 42–48. doi: 10.1016/j.intimp.2017.10.009

Largeot, A., Pagano, G., Gonder, S., Moussay, E., and Paggetti, J. (2019). The B-side of cancer immunity: the underrated tune. *Cells* 8:449. doi: 10.3390/cells8050449

Lechner, A., Schlößer, H. A., Thelen, M., Wennhold, K., Rothschild, S. I., Gilles, R., et al. (2019). Tumor-associated B cells and humoral immune response in head and neck squamous cell carcinoma. *Oncoimmunology* 8, 1535293–1535293. doi: 10.1080/2162402X.2018.1535293

Li, Y., Zhai, Y., Song, Q., Zhang, H., Cao, P., Ping, J., et al. (2018). Genome-Wide association study identifies a new locus at 7q21.13 associated with hepatitis B virus–related hepatocellular carcinoma. *Clin. Cancer Res.* 24, 906–915. doi: 10.1158/1078-0432.CCR-17-2537

Liu, M., Sun, Q., Wang, J., Wei, F., Yang, L., and Ren, X. (2019). A new perspective: exploring future therapeutic strategies for cancer by understanding the dual role of B lymphocytes in tumor immunity. *Int. J. Cancer* 144, 2909–2917. doi: 10.1002/ijc.31850

Losic, B., Craig, A. J., Villacorta-Martin, C., Martins-Filho, S. N., Akers, N., Chen, X., et al. (2020). Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat. Commun.* 11:291. doi: 10.1038/s41467-019-14050-z

Luci, C., Vieira, E., Perchet, T., Gual, P., and Golub, R. (2019). Natural killer cells and type 1 innate lymphoid cells are new actors in non-alcoholic fatty liver disease. *Front. Immunol.* 10:1192. doi: 10.3389/fimmu.2019.01192

MacParland, S. A., Liu, J. C., Ma, X.-Z., Innes, B. T., Bartczak, A. M., Gage, B. K., et al. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* 9:4383. doi: 10.1038/s41467-018-06318-7

Marcellin, P., and Kutala, B. K. (2018). Liver diseases: a major, neglected global public health problem requiring urgent actions and large-scale screening. *Liver Int.* 38(Suppl. 1), 2–6. doi: 10.1111/liv.13682

Mather, K. A., and Thalamuthu, A. (2020). Unraveling the genetic contributions to complex traits across different ethnic groups. *Nat. Med.* 26, 467–469. doi: 10.1038/s41591-020-0834-3

Matsuura, K., Sawai, H., Ikeo, K., Ogawa, S., Iio, E., Isogawa, M., et al. (2017). Genome-wide association study identifies TLL1 variant associated with development of hepatocellular carcinoma after eradication of hepatitis C virus infection. *Gastroenterology* 152, 1383–1394. doi: 10.1053/j.gastro.2017.01.041

Nicoletti, P., Aithal, G. P., Bjornsson, E. S., Andrade, R. J., Sawle, A., Arrese, M., et al. (2017). Association of liver injury from specific drugs, or groups of drugs, with polymorphisms in HLA and other genes in a genome-wide association study. *Gastroenterology* 152, 1078–1089. doi: 10.1053/j.gastro.2016.12.016

Ouyang, F. Z., Wu, R. Q., Wei, Y., Liu, R. X., Yang, D., Xiao, X., et al. (2016). Dendritic cell-elicited B-cell activation fosters immune privilege via IL-10 signals in hepatocellular carcinoma. *Nat. Commun.* 7:13453. doi: 10.1038/ncomms13453

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Ramachandran, P., Dobie, R., Wilson-Kamanori, J. R., Dora, E. F., Henderson, B. E. P., Luu, N. T., et al. (2019). Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* 575, 512–518. doi: 10.1038/s41586-019-1631-3

Ramachandran, P., Matchett, K. P., Dobie, R., Wilson-Kamanori, J. R., and Henderson, N. C. (2020). Single-cell technologies in hepatology: new insights into liver biology and disease pathogenesis. *Nat. Rev. Gastroenterol. Hepatol.* 17, 457–472. doi: 10.1038/s41575-020-0304-x

Saviano, A., Henderson, N. C., and Baumert, T. F. (2020). Single-cell genomics and spatial transcriptomics: discovery of novel cell states and cellular interactions in liver physiology and disease biology. *J. Hepatol.* 73, 1219–1230. doi: 10.1016/j.jhep.2020.06.004

Schwantes-An, T.-H., Darlay, R., Mathurin, P., Masson, S., Liangpunsakul, S., Mueller, S., et al. (2020). Genome-wide association study and meta-analysis on alcohol-related liver cirrhosis identifies novel genetic risk factors. *Hepatology [Online ahead of print]* doi: 10.1002/hep.31535

Setliff, I., Shiakolas, A. R., Pilewski, K. A., Murji, A. A., Mapengo, R. E., Janowska, K., et al. (2019). High-throughput mapping of b cell receptor sequences to antigen specificity. *Cell* 179, 1636.e–1646.e. doi: 10.1016/j.cell.2019.11.003

Sharma, A., Seow, J. J. W., Dutertre, C.-A., Pai, R., Blériot, C., Mishra, A., et al. (2020). Onco-fetal reprogramming of endothelial cells drives immunosuppressive macrophages in hepatocellular carcinoma. *Cell* 183, 377.e–394.e. doi: 10.1016/j.cell.2020.08.040

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

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

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., et al. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biol.* 21, 12–12. doi: 10.1186/s13059-019-1850-9

Tsou, P., Katayama, H., Ostrin, E. J., and Hanash, S. M. (2016). The emerging role of B cells in tumor immunity. *Cancer Res.* 76, 5597–5601. doi: 10.1158/0008-5472.CAN-16-0431

Van Rossum, G., and De Boer, J. (1991). Interactively testing remote servers using the Python programming language. *CWI Q.* 4, 283–304.

Wang, K., Nie, X., Rong, Z., Fan, T., Li, J., Wang, X., et al. (2017). B lymphocytes repress hepatic tumorigenesis but not development in Hras12V transgenic mice. *Int. J. Cancer* 141, 1201–1214. doi: 10.1002/ijc.30823

Wang, S.-S., Liu, W., Ly, D., Xu, H., Qu, L., and Zhang, L. (2019). Tumor-infiltrating B cells: their role and application in anti-tumor immunity in lung cancer. *Cell. Mol. Immunol.* 16, 6–18. doi: 10.1038/s41423-018-0027-x

Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* 11:3865. doi: 10.1038/s41467-020-17719-y

Yang, S., and Zhou, X. (2020). Accurate and scalable construction of polygenic scores in large biobank data sets. *Am. J. Hum. Genet.* 106, 679–693. doi: 10.1016/j.ajhg.2020.03.013

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al. (2019). Ensembl 2020. *Nucleic Acids Res.* 48, D682–D688. doi: 10.1093/nar/gkz966

Zhang, Q., He, Y., Luo, N., Patel, S. J., Han, Y., Gao, R., et al. (2019). Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* 179, 829.e–845.e. doi: 10.1016/j.cell.2019.10.003

Zheng, C., Zheng, L., Yoo, J.-K., Guo, H., Zhang, Y., Guo, X., et al. (2017). Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* 169, 1342.e–1356.e. doi: 10.1016/j.cell.2017.05.035

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communi.* 8, 14049. doi: 10.1038/ncomms14049

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: 29 March 2021
doi: 10.3389/fgene.2021.649196

A Fast Multi-Locus Ridge Regression Algorithm for High-Dimensional Genome-Wide Association Studies

*Jin Zhang[1,2]†, Min Chen[1]†, Yangjun Wen[1], Yin Zhang[1], Yunan Lu[1], Shengmeng Wang[1] and Juncong Chen[3]**

[1] College of Science, Nanjing Agricultural University, Nanjing, China, [2] Postdoctoral Research Station of Crop Science, Nanjing Agricultural University, Nanjing, China, [3] College of Finance, Nanjing Agricultural University, Nanjing, China

The mixed linear model (MLM) has been widely used in genome-wide association study (GWAS) to dissect quantitative traits in human, animal, and plant genetics. Most methodologies consider all single nucleotide polymorphism (SNP) effects as random effects under the MLM framework, which fail to detect the joint minor effect of multiple genetic markers on a trait. Therefore, polygenes with minor effects remain largely unexplored in today's big data era. In this study, we developed a new algorithm under the MLM framework, which is called the fast multi-locus ridge regression (FastRR) algorithm. The FastRR algorithm first whitens the covariance matrix of the polygenic matrix K and environmental noise, then selects potentially related SNPs among large scale markers, which have a high correlation with the target trait, and finally analyzes the subset variables using a multi-locus deshrinking ridge regression for true quantitative trait nucleotide (QTN) detection. Results from the analyses of both simulated and real data show that the FastRR algorithm is more powerful for both large and small QTN detection, more accurate in QTN effect estimation, and has more stable results under various polygenic backgrounds. Moreover, compared with existing methods, the FastRR algorithm has the advantage of high computing speed. In conclusion, the FastRR algorithm provides an alternative algorithm for multi-locus GWAS in high dimensional genomic datasets.

Keywords: genome-wide association study, mixed linear model, multi-locus algorithm, statistical power, polygenic background, minor effect

Edited by:
Sheng Yang,
Nanjing Medical University, China

Reviewed by:
HaiYan Lü,
Henan Agricultural University, China
Qidi Feng,
Broad Institute, United States
Yuan Niu,
School of Life Sciences and Food Engineering, Huaiyin Institute of Technology, China

***Correspondence:**
Juncong Chen
jcchen@njau.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

***Received:** 04 January 2021*
***Accepted:** 01 March 2021*
***Published:** 29 March 2021*

Citation:
Zhang J, Chen M, Wen Y, Zhang Y, Lu Y, Wang S and Chen J (2021) A Fast Multi-Locus Ridge Regression Algorithm for High-Dimensional Genome-Wide Association Studies. Front. Genet. 12:649196. doi: 10.3389/fgene.2021.649196

INTRODUCTION

Genome-wide association study (GWAS) has been widely used in the genetic dissection of quantitative traits in human, animal, and plant genetics. GWAS typically searches for the correlations between genetic variants and hundreds or thousands of individuals. However, a complete characterization of the biological mechanism for most quantitative traits remains elusive

Abbreviations: MLM, the mixed linear model; GWAS, genome-wide association study; FastRR, fast multi-locus ridge regression; SNP, single nucleotide polymorphism; QTN, quantitative trait nucleotide; EMMA, efficient mixed model association; DRR, deshrinking ridge regression; ORR, ordinary ridge regression; BLUP, best linear unbiased prediction; lasso, least absolute shrinkage and selection operator; SCAD, smoothly clipped absolute deviation; DEMMA, Decontaminated efficient mixed model association; MAF, minor allele frequency; LD, days to flowering under long days; SD, days to flowering under short days; SDV, days to flowering under short days with vernalization; MSE, mean squared error.

(Dahl et al., 2016) and a number of polygenes with minor effects are unexplored (Zhang and Xu, 2005; Wen et al., 2019). This may be because the GWAS approach is still quite crude, and most of the minor biological associations between sequence and phenotype remain unmeasured. Recently, advanced biotechnology has generated large-scale single nucleotide polymorphisms (SNPs) and phenotypes, which have been valuable for genetic analysis. A large number of statistical methodologies for GWAS have been proposed (Atwell et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012; Wen et al., 2018, 2020; Sun et al., 2019; Wang et al., 2020).

Since the introduction of the Q + K (Q represents the population structure and K represents the kinship matrix) mixed linear model (MLM) approach (Yu et al., 2006) to the concept of GWAS, the power of quantitative trait nucleotide (QTN) detection has been significantly increased. On this basis, the compressed MLM (Zhang et al., 2010) and enriched compressed MLM (Li et al., 2014) have been proposed to improve computational efficiency. Meanwhile, an efficient mixed model association (EMMA) (Kang et al., 2008) was regarded as the milestone improvement in the MLM approach, which treated the polygenic effect as the random effect to fit the mixed model. Currently, this concept has become more and more popular in genomic analysis. A number of methods based on this concept are continually emerging, such as EMMAX (Kang et al., 2010), FaST-LMM (Lippert et al., 2011), and GEMMA (Zhou and Stephens, 2012). Because of the dissection of genetic variants and computational speed, all these methods have been successfully applied in MLM. For all the above methods, they comprise a one-dimensional genome scan by testing one marker at a time, more importantly, the SNP effect is considered as the fixed effect, which may be disadvantageous to the detection of QTN in GWAS (Goddard et al., 2009; Zhang et al., 2017; Wen et al., 2018, 2020).

Although the current single variant methods of GWAS have succeeded in identifying QTNs associated with the interested traits, these approaches fail to consider the joint minor effect of multiple genetic markers on a trait (Tamba et al., 2017); furthermore, they do not match the internal genetic mechanism of these quantitative traits (Tamba et al., 2017; Zhang et al., 2017; Sun et al., 2019; Wen et al., 2019). To overcome this drawback, multi-locus methodologies have been developed, such as least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996; Xu, 2010; Zhang et al., 2012), Bayesian lasso (Yi and Xu, 2008), adaptive mixed lasso (Wang et al., 2011), and empirical Bayes (Xu, 2007). All SNPs can be included in the model and can be simultaneously estimated by using multi-locus methodologies. If the number of SNPs (p) is many times larger than the number of individuals (n), the approaches will fail to analyze this oversaturated model. Under this circumstance, a natural response is to consider reducing the number of SNP effects in the multi-locus genetic model. Zhou et al. (2013) and Moser et al. (2015) proposed the Bayesian model, which estimates only a few variance components instead of considering all. It is an alternative approach to solve the "big p, small n" problem. Currently, two-stage methodologies (Tamba et al., 2017; Zhang et al., 2017; Wen et al., 2018) borrowed this idea and have been proposed for multi-locus GWAS. All these methodologies provide the tools for high-dimensional genetic data analysis. It is known that the quantitative traits are controlled by a few genes with large effects and numerous polygenes with minor effects. Nevertheless, the dissection of the polygenes with minor effects needs to be improved in above mentioned multi-locus approaches.

In this study, we propose a multi-stage flexible approach for GWAS to detect the associated (large and minor effects) variables/SNPs. In our model, the fast multi-locus ridge regression algorithm (FastRR), all SNP effects are considered as random effects. The FastRR algorithm first whitens the covariance matrix of the polygenic matrix K and environmental noise. Subsequently, the FastRR algorithm reduces the number of SNPs according to correlation, the variables of which significantly correlate with the response are retained for the next stage. In the final stage, deshrinking ridge regression (DRR) is applied to implement parametric estimation and significance tests of variables. In this study, a series of simulated and real dataset analyses are used to validate this new method. For comparison, five established methods – lasso, adaptive lasso, smoothly clipped absolute deviation (SCAD), EMMA, and decontaminated efficient mixed model association (DEMMA) are used for analysis.

MATERIALS AND METHODS

Genetic Model

Let $y_i (i = 1, 2, ..., n)$ be the phenotypic value of the i-th individual in a sample of size n from a natural population, and the genetic model can be described as:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{Z}\gamma + \mathbf{u} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y} = (y_1, ..., y_n)^T$; $\boldsymbol{\alpha}$ is a $c \times 1$ vector of the fixed effects, such as the intercept, population structure effect and so on, $\mathbf{W}$ is the corresponding designed matrix for $\boldsymbol{\alpha}$; $\mathbf{Z}$ is an $n \times 1$ vector of marker genotypes, and $\gamma \sim N(0, \sigma_\gamma^2)$ is a random effect of putative QTN. σ_γ^2 is the variance of the putative QTN; $\mathbf{u} \sim MVN(\mathbf{0}, {\sigma_g}^2\mathbf{K})$ is an $n \times 1$ random vector of polygenic effects, ${\sigma_g}^2$ is the variance of polygenic background, $\mathbf{K}$ is a known $n \times n$ relatedness matrix; $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of residual errors with an assumed $MVN(\mathbf{0}, \sigma^2\mathbf{I}_n)$ distribution; σ^2 is the variance of residual error; and $\mathbf{I}_n$ is a $n \times n$ identity matrix. MVN denotes multivariate normal distribution.

As γ is treated as being a random effect, the variance of $\boldsymbol{y}$ in the model (1) is:

$$var(\mathbf{y}) = \sigma_\gamma^2\mathbf{Z}\mathbf{Z}^T + {\sigma_g}^2\mathbf{K} + \sigma^2\mathbf{I}_n = \sigma^2(\lambda_\gamma\mathbf{Z}\mathbf{Z}^T + \lambda_g\mathbf{K} + \mathbf{I}_n) \quad (2)$$

where $\lambda_\gamma = {\sigma_\gamma}^2/\sigma^2, \lambda_g = {\sigma_g}^2/\sigma^2$.

Fast Multi-Locus Ridge Regression Algorithm

The FastRR algorithm is a multi-stage flexible approach for GWAS, which simultaneously implements estimation and testing to detect associated variables/SNPs. We describe it with the following stages:

The Polygenic and Residual Noise Whitening Stage

The key point of solving the model (1) is to estimate two ratios of variance components, λ_γ and λ_g, which cause expensive computational burden. It is noted that polygenic variance is always larger than zero, while variance components for most SNPs are zero because these markers are not associated with the interested trait, which is $\lambda_\gamma = 0$ for most SNPs. Therefore, in the first step, we estimate $\hat{\lambda}_g$ by the reduced form of the model (1), which deleted $\mathbf{Z}\gamma$ with only polygenic background, and replace λ_g in (2) by the $\hat{\lambda}_g$ (Wen et al., 2018, 2020), avoiding re-estimate λ_g for each single marker scanning. Thus,

$$var(\mathbf{y}) = \sigma^2(\lambda_\gamma \mathbf{ZZ}^T + \hat{\lambda}_g\mathbf{K} + \mathbf{I}_n) = \sigma^2(\lambda_\gamma \mathbf{ZZ}^T + \mathbf{B}) \quad (3)$$

An eigen (or spectral) decomposition of the positive definite matrix $\mathbf{B} = \hat{\lambda}_g\mathbf{K} + \mathbf{I}_n$ is:

$$\mathbf{B} = \mathbf{Q}\Lambda\mathbf{Q}^\mathbf{T} = (\mathbf{Q}\Lambda^{\frac{1}{2}}\mathbf{Q}^\mathbf{T})(\mathbf{Q}\Lambda^{\frac{1}{2}}\mathbf{Q}^\mathbf{T}) \quad (4)$$

where $\mathbf{Q}$ is orthogonal and Λ is a diagonal matrix with positive eigenvalues. Let $\mathbf{C} = \mathbf{Q}\Lambda^{-\frac{1}{2}}\mathbf{Q}^\mathbf{T}$, the model (1) is changed to:

$$\mathbf{y}_c = \mathbf{W}_c\boldsymbol{\alpha} + \mathbf{Z}_c\gamma + \boldsymbol{\varepsilon}_c \quad (5)$$

where, $\mathbf{y}_c = \mathbf{C}\mathbf{y}$, $\mathbf{W}_c = \mathbf{C}\mathbf{W}$, $\mathbf{Z}_c = \mathbf{C}\mathbf{Z}$, $\boldsymbol{\varepsilon}_c = \mathbf{C}\mathbf{u} + \mathbf{C}\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_n)$ (Wen et al., 2018, 2020).

Variable Reduction Stage

A number of studies have illustrated that most quantitative traits are controlled by a small portion of genes, including a few genes with large effects and polygenes with minor effects (Zhang et al., 2017; Wen et al., 2019). It is critical to dissect all associated loci from large-scale genetic markers. Herein, we conduct a variable reduction stage, whose purpose is dimension reduction. At this stage, the FastRR algorithm detects a subset of putative variables associated with the phenotype, and thus avoids the intractable computational problems of high-dimensional datasets analysis.

We calculate the marginal correlation coefficients between $\mathbf{Z}_c$ (variables after polygenic background correction) and $\mathbf{y}_c$ (phenotype after polygenic background correction) under model (5), R function *cor.test* returns the *p*-value of the correlation test. The critical value for significance was set at p-value < 0.01 (Tamba et al., 2017). For the threshold of 0.01, even the slight correlations between predictors and the response will be captured (Tamba et al., 2017), and the unassociated loci will be removed. All the most potential QTNs are selected to construct the reduced multi-locus model for the next stage. Essentially, this marginal correlation step is similar to the single marker scanning, which combined with the polygenic background without considering variance components σ_γ^2.

Parameter Estimation Stage

In the multi-locus model,

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad (6)$$

where $\mathbf{y}$ is the phenotypic value of the quantitative trait, which is the same as that in the model (1); $\boldsymbol{\alpha}$ is a vector of fixed effects, $\boldsymbol{\gamma}$ is a $q \times 1$ random effect vector of the selected q markers from the above stage, and $\gamma_k \sim N\left(0, \phi^2\right), k = 1, \ldots, q$; $\mathbf{W}$ and $\mathbf{Z}$ are the corresponding design matrices for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Here, polygenic background correction is not considered in model (6), because the above two steps under the polygenic background model had already selected all potential associated QTNs. All the parameters in model (6) are estimated by DRR proposed by Wang et al. (2020).

Before introducing the DRR, let us briefly recall the ordinary ridge regression (ORR). According to the best linear unbiased prediction (BLUP) of the marker effects and the prediction error variances using the conditional expectation and conditional variance, the estimates of ORR are as follows,

$$\widehat{\boldsymbol{\gamma}}^{ORR} = E\left(\boldsymbol{\gamma}|\mathbf{y}\right) = \lambda\mathbf{Z}^T\mathbf{H}^{-1}(\boldsymbol{y} - \mathbf{W}\boldsymbol{\alpha}) \quad (7)$$

$$var\left(\widehat{\boldsymbol{\gamma}}^{ORR}|\mathbf{y}\right) = \left(\lambda\mathbf{I} - \lambda\mathbf{Z}^T\boldsymbol{H}^{-1}\mathbf{Z}\lambda\right) \quad (8)$$

where $\lambda = \frac{\phi^2}{\sigma^2}$, $\mathbf{H} = \left(\mathbf{ZZ}^T\right)\lambda + I_n$.

Ordinary ridge regression is inflexible and inaccurate for GWAS (Wang et al., 2020). Therefore, we apply the following DRR method, which can bring both the accurate effects and tests back. The essential difference between ORR and DRR is the well-measurement-factor (also called degree of freedom), which is

$$d_k = 1 - \frac{var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right)}{\phi^2} = \lambda\mathbf{Z}_k^T\mathbf{H}^{-1}\mathbf{Z}_k \quad (9)$$

$\widehat{\gamma}_k^{ORR}$ is the k-th element of $\widehat{\gamma}^{ORR}$, where $\phi^2$and $var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right)$ are prior and posterior variances for γ_k, respectively.

$$\widehat{\gamma}_k^{DRR} = \frac{\phi^2}{\phi^2 - var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right)}\widehat{\gamma}_k^{ORR} = d_k^{-1}\,\widehat{\gamma}_k^{ORR} \quad (10)$$

$$\begin{aligned} var\left(\widehat{\gamma}_k^{DRR}\right) &= \frac{\phi^2}{\phi^2 - var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right)} var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right) \\ &= d_k^{-1}\, var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right) \quad (11) \end{aligned}$$

$$W_k = \frac{\left(\widehat{\gamma}_k^{DRR}\right)^2}{var\left(\widehat{\gamma}_k^{DRR}\right)} = \frac{\left(\widehat{\gamma}_k^{ORR}/d_k\right)^2}{var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right)/d_k} = d_k^{-1}\frac{\left(\widehat{\gamma}_k^{ORR}\right)^2}{var\left(\widehat{\gamma}_k^{ORR}|\mathbf{y}\right)} \quad (12)$$

The test statistic of DRR, W_k, follows a Chi-square distribution with one degree of freedom under the null model, $H_0 : \gamma_k = 0$. The DRR method deshrinks both the estimated effects of markers and their estimated variances from the ORR, resulting in deshrunk Wald test statistics.

Comparison Methods

LASSO

Lasso regression (Tibshirani, 1996) is a type of linear regression that implements shrinkage by performing L_1 regularization and selects the most correlated with response variables. It is a popular method for simultaneous estimation and variable selection. The method was implemented by the R software package *lars*[1].

Adaptive Lasso

Similar to the lasso, the adaptive lasso (Zou, 2006) is a mainstream method of variable selection, in which the adaptive weights are used for penalizing different coefficients in the L_1 penalty. Adaptive lasso shows more consistence for variable selection than lasso in data analysis. The method was implemented by the R software package *glmnet*[2].

SCAD

SCAD (Fan and Li, 2001) as the variable selection has the nice oracle property. The estimator of SCAD attempts to alleviate bias from variable selection, while also retaining a continuous penalty that encourages sparsity. The method was implemented by the R software package *ncvreg*[3].

EMMA

Efficient mixed-model association (Kang et al., 2008) is an established genome-wide single-marker scan methodology under the framework of MLM, in which the polygenic background and population structure are controlled. The method was implemented by the R software package EMMA[4].

DEMMA

The polygenic effect (the sum of all marker effects) is treated as a random effect in EMMA. On the other side, EMMA already included the marker effect as the fixed effect. Thus, there are two effects for each marker, which lead to a reduced power for testing. Wang et al. (2020) proposed DEMMA to overcome the above drawback. The method was implemented by the R code[5].

Experimental Materials

The Simulation Data

Three Monte Carlo simulation experiments were conducted to evaluate the performances of the FastRR algorithm and other methods. We generated genotypes according to the minor allele frequency (MAF) in the interval (0.1, 0.5) under Hardy–Weinberg equilibrium. The simulation datasets contained n = 2000 individuals with p = 10,000 genetic variants, which were generated with MLM. The total average was set at 10.0 and residual variance was set at 10.0. We considered three scenarios for each simulation, including two times polygenic background, five times polygenic background, and ten times polygenic background.

Only one QTN with a fixed position (**Table 1**) was simulated and placed on the SNPs with 0.1 heritability for the first simulation; five QTNs with fixed positions were assigned and placed on the SNPs for the second simulation, the heritabilities of the QTNs were set as 0.02, 0.05, 0.05, 0.08, and 0.10, respectively.

TABLE 1 | Comparison of lasso, adaptive lasso, SCAD, EMMA, DEMMA, and FastRR methods in the first simulation experiment (three scenarios).

Polygenic back	True value			Lasso			Adaptive lasso			SCAD			EMMA			DEMMA			FastRR		
ground	Position	Effect	r^2	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE
2K	98	0.7398	5%	100.0	0.476 (0.092)	7.768	83.0	0.374 (0.155)	13.079	100.0	0.474 (0.156)	9.446	100.0	0.736 (0.091)	0.818	100.0	0.736 (0.091)	0.818	100.0	0.734 (0.091)	0.817
5K	98	0.7398	5%	100.0	0.404 (0.111)	12.527	59.0	0.315 (0.224)	13.585	100.0	0.390 (0.164)	14.915	98.0	0.735 (0.103)	1.040	99.0	0.733 (0.105)	1.089	100.0	0.729 (0.109)	1.188
10K	98	0.7398	5%	91.0	0.337 (0.134)	16.386	32.0	0.380 (0.247)	6.048	87.0	0.324 (0.168)	17.446	70.0	0.795 (0.094)	0.829	84.0	0.765 (0.110)	1.052	99.0	0.729 (0.131)	1.693
False positive rate of 2K (‰)					0.453			0.004			0.288			0.030			0.014			0.450	
False positive rate of 5K (‰)					0.555			0.001			0.460			0.090			0.018			0.498	
False positive rate of 10K (‰)					0.636			0.019			0.550			0.050			0.026			0.436	

Three scenarios, including two times polygenic background, five times polygenic background, and ten times polygenic background. MSE, mean squared error. The numbers in parentheses represent the standard deviation.

[1] https://cran.r-project.org/web/packages/lars/index.html

[2] https://glmnet.stanford.edu/

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

[4] http://mouse.cs.ucla.edu/emma/

[5] https://doi.org/10.1093/bioinformatics/btaa345/

TABLE 2A | Comparison of lasso, adaptive lasso, SCAD, EMMA, DEMMA, and FastRR methods in the second simulation experiment (scenarios 1: two times polygenic background).

QTN	True value			Lasso			Adaptive lasso			SCAD			EMMA			DEMMA			FastRR		
	Position	Effect	r^2	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE
Polygenic background (2K)																					
1	98	0.5451	2%	99.0	0.298 (0.091)	6.833	96.0	0.416 (0.149)	3.703	99.0	0.269 (0.122)	9.011	91.0	0.600 (0.087)	0.956	94.0	0.596 (0.089)	0.978	99.0	0.587 (0.094)	1.035
2	301	0.8622	5%	100.0	0578 (0.100)	9.080	100.0	0.782 (0.114)	1.924	100.0	0.683 (0.174)	6.221	100.0	0.822 (0.095)	1.044	100.0	0.822 (0.095)	1.044	100.0	0.820 (0.094)	1.054
3	540	0.8598	5%	100.0	0.605 (0.093)	7.350	100.0	0.811 (0.101)	1.240	100.0	0.730 (0.150)	3.906	100.0	0.852 (0.089)	0.788	100.0	0.852 (0.089)	0.788	100.0	0.850 (0.089)	0.788
4	801	1.0789	8%	100.0	0.807 (0.099)	8.34	100.0	1.030 (0.105)	1.333	100.0	1.025 (0.139)	2.211	100.0	1.061 (0.094)	0.914	100.0	1.061 (0.094)	0.914	100.0	1.059 (0.094)	0.911
5	1000	1.2093	10%	100.0	0.957 (0.095)	7.276	100.0	1.118 (0.098)	1.023	100.0	1.207 (0.251)	10.129	100.0	1.223 (0.094)	0.886	100.0	1.223 (0.094)	0.886	100.0	1.220 (0.094)	0.878
False positive rate (‰)					0.461			0.024			0.355			0.000			0.007			0.422	

Three scenarios, including two times polygenic background, five times polygenic background, and ten times polygenic background.
MSE, mean squared error.
The numbers in parentheses represent the standard deviation.

TABLE 2B | Comparison of lasso, adaptive lasso, SCAD, EMMA, DEMMA, and FastRR methods in the second simulation experiment (scenarios 2: five times polygenic background)

QTN	True value			Lasso			Adaptive lasso			SCAD			EMMA			DEMMA			FastRR		
	Position	Effect	r^2	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE
Polygenic background (5K)																					
1	98	0.5451	2%	89.0	0.239 (0.091)	9.048	71.0	0.375 (0.179)	4.297	88.0	0.216 (0.098)	10.367	52.0	0.656 (0.072)	0.943	73.0	0.622 (0.082)	0.910	96.0	0.587 (0.095)	1.029
2	301	0.8622	5%	100.0	0.527 (0.119)	12.673	100.0	0.764 (0.166)	3.703	100.0	0.606 (0.200)	10.515	99.0	0.841 (0.106)	1.140	99.0	0.841 (0.106)	1.140	100.0	0.820 (0.126)	1.283
3	540	0.8598	5%	100.0	0.518 (0.117)	13.063	100.0	0.754 (0.153)	3.439	100.0	0.591 (0.191)	10.812	99.0	0.831 (0.107)	1.195	100.0	0.828 (0.110)	1.297	100.0	0.826 (0.109)	1.299
4	801	1.0789	8%	100.0	0.755 (0.116)	11.824	100.0	1.029 (0.126)	1.811	100.0	0.957 (0.186)	4.911	100.0	1.077 (0.117)	1.336	100.0	1.077 (0.116)	1.336	100.0	1.075 (0.116)	1.334
5	1000	1.2093	10%	100.0	0.897 (0.109)	10.937	100.0	1.176 (0.117)	1.480	100.0	1.165 (0.150)	2.428	100.0	1.234 (0.101)	1.063	100.0	1.234 (0.101)	1.063	100.0	1.232 (0.100)	1.049
False positive rate (‰)					0.510			0.102			0.473			0.040			0.014			0.431	

Three scenarios, including two times polygenic background, five times polygenic background, and ten times polygenic background.
MSE, mean squared error.
The numbers in parentheses represent the standard deviation.

TABLE 2C | Comparison of lasso, adaptive lasso, SCAD, EMMA, DEMMA, and FastRR methods in the second simulation experiment (scenarios 3: ten times polygenic background).

QTN	True value			Lasso			Adaptive lasso			SCAD			EMMA			DEMMA			FastRR		
	Position	Effect	r^2	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE	Power (%)	Effect (SD)	MSE
Polygenic background (10K)																					
1	98	0.5451	2%	56.0	0.223 (0.092)	6.283	46.0	0.393 (0.188)	4.297	51.0	0.240 (0.092)	5.165	20.0	0.757 (0.047)	0.943	36.0	0.706 (0.069)	1.102	76.0	0.644 (0.095)	1.160
2	301	0.8622	5%	97.0	0.437 (0.126)	19.080	93.0	0.718 (0.212)	6.046	98.0	0.488 (0.195)	17.444	89.0	0.860 (0.102)	0.923	93.0	0.851 (0.108)	1.088	100.0	0.830 (0.126)	1.668
3	540	0.8598	5%	97.0	0.459 (0.141)	17.520	97.0	0.726 (0.235)	1.240	98.0	0.516 (0.210)	15.874	88.0	0.873 (0.119)	1.242	94.0	0.858 (0.128)	1.529	99.0	0.842 (0.140)	1.960
4	801	1.0789	8%	100.0	0.682 (0.147)	17.912	99.0	1.020 (0.173)	3.287	100.0	0.855 (0.251)	11.254	100.0	1.085 (0.141)	1.962	100.0	1.085 (0.141)	1.962	100.0	1.083 (0.141)	1.958
5	1000	1.2093	10%	100.0	0.783 (0.159)	20.627	99.0	1.129 (0.174)	3.592	100.0	1.012 (0.251)	10.129	100.0	1.206 (0.152)	2.297	100	1.206 (0.153)	2.297	100.0	1.204 (0.152)	2.290
False positive rate (‰)					0.673			0.209			0.788			0.050			0.026			0.490	

Three scenarios, including two times, five times and ten times the polygenic background.
MSE, mean squared error.
The numbers in parentheses represent the standard deviation.

Their positions and effects are listed in **Tables 2A–C**. For the third simulation experiment, we randomly selected 100 QTNs, and the sum contribution of QTNs to the total phenotypic variance was 0.5. Each simulation experiment was repeated 100 times. The power for each QTN was defined as the proportion of samples over the threshold to the total number of replicates (100), the criterion for lasso, adaptive lasso, and SCAD was set as LOD $\geq$ 3.0, the criterion for ORR, EMMA, DEMMA, and the FastRR algorithm was set as $0.05/p$, where p was the number of markers in the genetic model. The false positive rate was calculated as the ratio of the number of false positive effects to the total number of zero effects.

The Rice Data

To validate the FastRR algorithm, the rice data that was used in this study for GWAS demonstration consists of 524 inbred varieties, which were collected from China and southeast Asia (Chen et al., 2014; Wei et al., 2018). A total of 6.5 million high-quality SNPs covering 90% of total SNPs were analyzed by Chen et al. (2014). A total of 314,393 SNPs and grain width traits (Wang et al., 2020) were analyzed in this study. These data were downloaded from the link.[6]

The *Arabidopsis* Data

To further evaluate the performance of FastRR, we reanalyzed the genetic data sets of *Arabidopsis* published by Atwell et al. (2010). Both phenotypes and genotypes were obtained from the link[7]. A total of 199 *Arabidopsis* lines and 216,130 SNPs were used for analysis. Among all traits, we analyzed three traits related to flowering time: (1) LD: days to flowering under long days; (2) SD: days to flowering under short days; and (3) SDV: days to flowering under short days with vernalization.

RESULTS

Simulation Studies

Statistical Power for QTN Detection

In the first simulation experiment, only one QTN with a fixed position is simulated, and the power in the detection of the QTN is higher for the FastRR algorithm than for the others (**Figure 1** and **Table 1**). The FastRR algorithm has a dramatically higher statistical power for 10 times polygenic background especially. When five QTNs with the fixed position are simulated in the second experiment, a similar trend is observed (**Figure 2** and **Tables 2A–C**). Three minor effect QTNs (QTL 1 and QTL 2 for three scenarios; QTL 3 for the third scenario) are illustrated in **Figure 2**, the power of each QTN is less than 100%. Notably, the FastRR algorithm has the highest power for the 98th marker (minor effect locus, r^2 = 2%) under different polygenic backgrounds. One hundred random QTNs are simulated in the third experiment and the total heritabilities are 50%. As the genetic background increases, the power of the FastRR algorithm is getting increasingly high (**Figure 3**).

[6] https://doi.org/10.1093/bioinformatics/btaa345/
[7] http://www.arabidopsis.usc.edu/

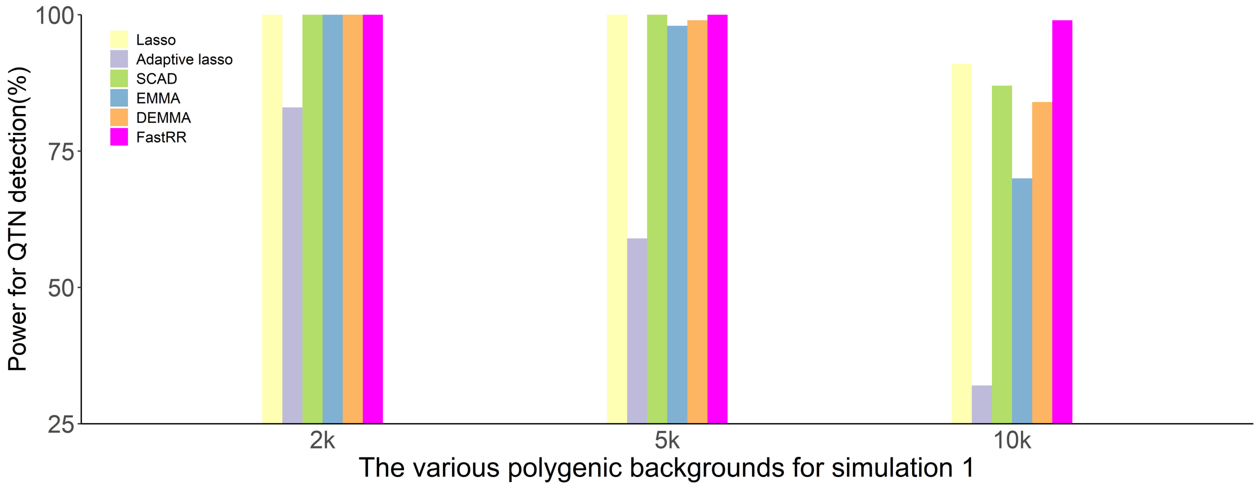

FIGURE 1 | The statistical powers for the fixed position QTN in the first simulation experiment using six methods (lasso, adaptive lasso, SCAD, EMMA, DEMMA, and the FastRR algorithm).

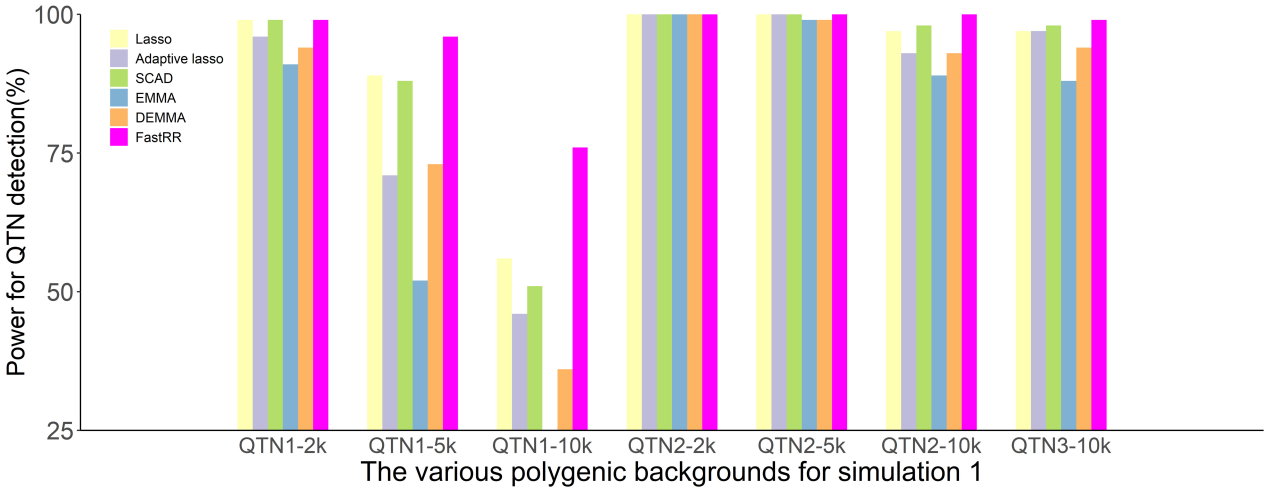

FIGURE 2 | The statistical powers for the minor effect QTNs in the second simulation experiment using six methods (lasso, adaptive lasso, SCAD, EMMA, DEMMA, and the FastRR algorithm).

The results illustrate that the trends are similar to the above experiments (**Figure 3**). In summary, the FastRR algorithm retains an obviously advantageous performance for the random loci experiment. These results demonstrate the highest power of the FastRR algorithm across all the approaches under various genetic backgrounds.

Accuracy for the Estimated QTN Effects

The average effect and mean squared error (MSE) are used to measure the accuracy of an estimated QTN effect. We evaluated the accuracies for the (fixed positions, including simulation experiment 1 and 2) estimates using all six methods (**Tables 1**, **2A–C**). As a result, the estimates for each QTN effect for EMMA, DEMMA, and FastRR are much closer to the true value, and EMMA and DEMMA are slightly better than the FastRR algorithm, nevertheless, EMMA and DEMMA methods have relatively lower power than FastRR. The performance of SCAD, adaptive lasso, and lasso are unsatisfactory. The MSE shows a similar trend to the average effect. On these occasions, the FastRR algorithm, EMMA, and DEMMA methods are recommended for the estimation of QTN effects.

The false positive rate is a crucial index in GWAS. All the false positive rate results of simulation experiment 1 and 2 are listed in **Tables 1**, **2A–C**. Obviously, the false positive rate becomes increasingly high along with the stronger polygenic background. EMMA, DEMMA, and adaptive lasso have a relatively lower false positive rate followed by FastRR, SCAD, and lasso. The false positive rates of all six methods are under control.

Computing Time

We compare the computing time of 100 repeated simulated analyses by using six approaches. In each of the three simulation experiments, computing times are recorded and are shown in **Figure 4** and **Supplementary Figures 1**, **2** (Intel Xeon E5-2630

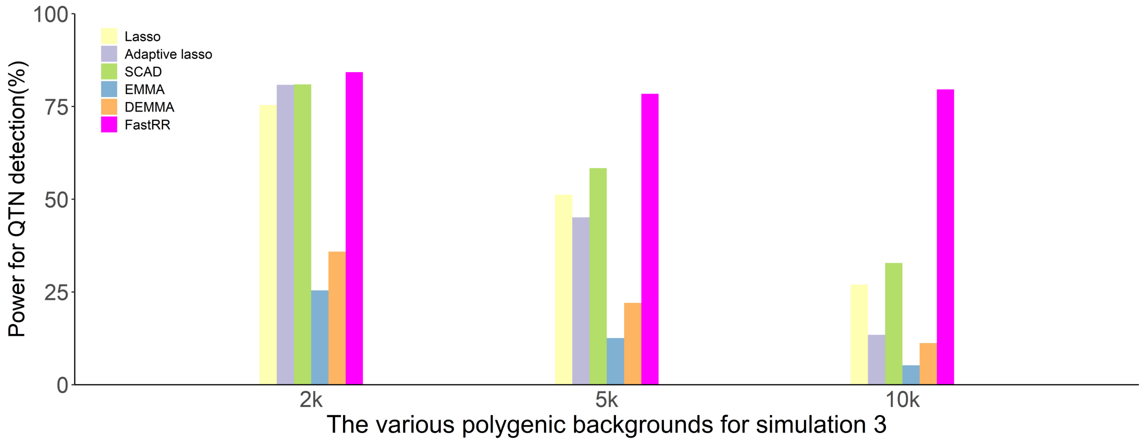

FIGURE 3 | The average statistical powers for all QTNs in the third simulation experiment using six methods (lasso, adaptive lasso, SCAD, EMMA, DEMMA, and the FastRR algorithm).

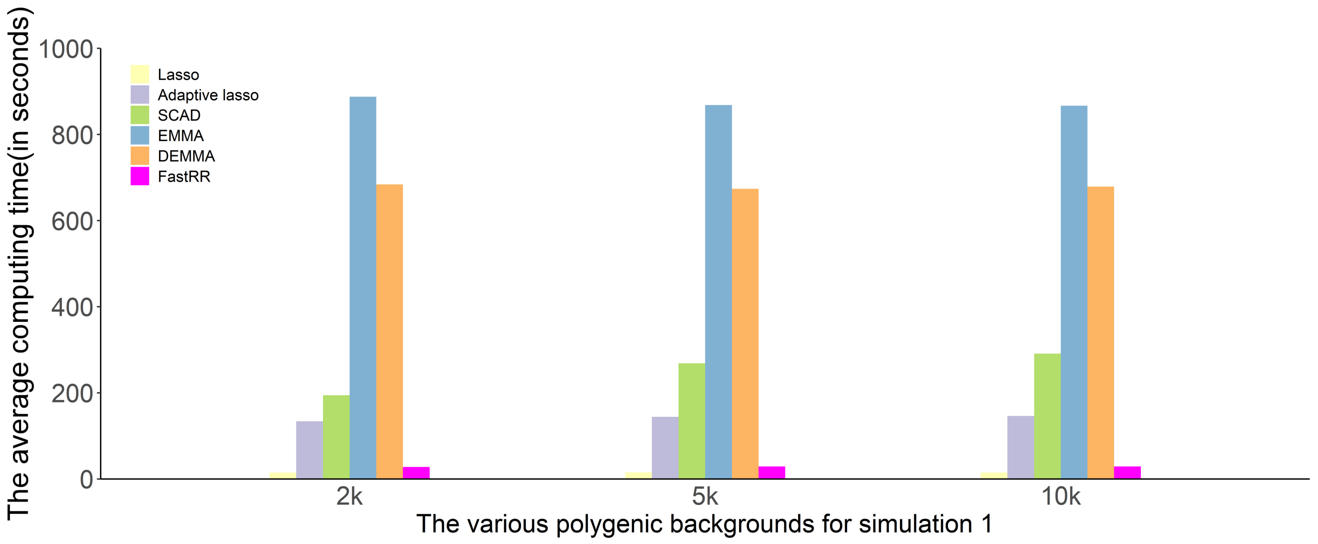

FIGURE 4 | Comparison of computing times to analyze simulation experiment 1 using all six methods (lasso, adaptive lasso, SCAD, EMMA, DEMMA, and the FastRR algorithm).

v4, CPU 2.20 GHz, Memory 64G). The computing time of the LASSO and FastRR algorithm have a faster computing speed than the other methods, which are on the same order of magnitude. They are followed by the adaptive lasso and SCAD. DEMMA and EMMA methods take the most expensive computing time at about 600 min, which is nearly seven times more than the FastRR algorithm.

Analysis of the Rice Data Set

To validate the FastRR algorithm, the grain width trait of rice data is analyzed by using six methods: lasso, adaptive lasso, SCAD, EMMA, DEMMA, and the FastRR algorithm. The rice dataset contains 310,000 SNPs genotyped for 524 inbred varieties. **Supplementary Figure 3** shows the LOD plot for three variable selection methods and Manhattan plots for the other three methods. Obviously, DEMMA method and the FastRR algorithm have the identical detected regions, two significant peaks on chromosome 5 and 9. Both DEMMA and FastRR detect the cloned gene *GW5* (Weng et al., 2008) that controls grain width trait. The test statistics of SNP135176 (the most significant SNP) for the DEMMA method and FastRR algorithm are 2.31×10^{-26} and 1.92×10^{-20}, respectively; the *p*-value for the DEMMA method is lower than for the FastRR algorithm. However, the test statistics for the EMMA method do not reach the Bonferroni correction threshold. In addition, three variable selection methods, lasso, adaptive lasso, and SCAD, show unsatisfactory performance according to the LOD scores.

The average computing times are listed in **Table 3**. The relatively fast methods, lasso, SCAD, and FastRR, are 235.33, 455.31, and 561.31 s, respectively. Lasso is the fastest method

TABLE 3 | The computation times (seconds) for analyzing *Arabidopsis* flowering time traits and rice grain width by using lasso, adaptive lasso, SCAD, EMMA, DEMMA, and FastRR methods.

Traits	Lasso	Adaptive lasso	SCAD	EMMA	DEMMA	FastRR
Rice						
Grain width	235.33	1067.22	455.31	60813.82	26417.71	561.31
Arabidopsis						
LD	36.11	189.36	128.79	1362.55	1117.49	105.17
SD	37.17	159.00	114.17	1350.19	4114.88	112.75
SDV	44.47	140.96	112.34	1665.94	4123.34	107.36

among all six methods, which is followed by SCAD and FastRR. In **Table 3**, the adaptive lasso is different from the above simulation experiments, which consumes much computing time in the cross-validation along with the increasing number of SNPs. The EMMA method takes more than ten times the computing time than the FastRR algorithm.

Analysis of the *Arabidopsis* Data Set

To further validate the FastRR algorithm, this new algorithm FastRR along with lasso, adaptive lasso, SCAD, EMMA, and DEMMA methods are used to reanalyze the *Arabidopsis* data for three traits related to flowering time (LD, SD, and SDV). The results are illustrated in **Supplementary Figures 4–6**. Each putative QTN (over the threshold) is used to mine the candidate genes by The *Arabidopsis* Information Resource[8]. The FastRR algorithm detects the confirmed genes *AGL*17 and *CDKG*1, which are detected by SCAD and DEMMA as well. From the analysis results, lasso shows several false positive loci in the detection of SD and SDV, meanwhile the adaptive lasso and SCAD methods are inflexible in dissecting the SNPs associated with the target traits. The statistical tests of EMMA are under the Bonferroni corrected threshold. The FastRR algorithm shows a similar pattern as the DEMMA method for all results of three traits, the statistics of part SNPs using the DEMMA method are slightly more significant than the FastRR algorithm, which is similar to the results of the rice datasets.

In terms of the computing speed for all three traits, lasso is computationally much faster than the other methods. The computing times of FastRR, SCAD, and adaptive lasso are on the same order of magnitude, which require less than 200 s. The DEMMA and EMMA methods have much more computational burden than the other methods, both of which require over ten times the computing time required by the FastRR algorithm. Overall, the FastRR algorithm is recommended from the perspective of detection and computing speed across all experiments.

DISCUSSION

The FastRR algorithm is a multi-stage flexible approach for QTNs dissection in GWAS, and displays high power for detecting QTN of large and minor effects, even under the ten times polygenic background. We aimed to understand the performance of regression analysis methods, thus the following three regression analysis methods, ORR, DRR, and FastRR, are used to analyze simulation experiment 1 and 2. As the results show (**Supplementary Tables 1**, **2A–C**), ORR has the worst detection ability, and even major QTN with large effects are not identified. This explains why ORR is rarely used in GWAS. DRR performs well in simulation 1 and 2, and shows slightly lower power for the major QTNs than FastRR. However, DRR loses power in detecting QTNs with minor effects, and this difference becomes more and more obvious with the increase of the polygenic background. Among three regression analysis methods, the FastRR performs well in the simulation experiment and has the highest statistical power.

Currently, the two-stage methodologies (Tamba et al., 2017; Zhang et al., 2017; Wen et al., 2018) are more popular in GWAS, which are the alternative approaches to solve the “big P, small N” problem. The FASTmrEMMA (Wen et al., 2018; Wen et al., 2020) algorithm is a fast and accurate two-stage methodology for QTNs detection. We further compare the FastRR and FASTmrEMMA algorithm in this study. The results of simulation experiment 1 and 2 are listed in **Supplementary Tables 1**, **2A–C**. Observably, the FastRR and FASTmrEMMA algorithm are powerful in QTNs detection from the perspective of statistical power. However, the estimation of FASTmrEMMA is slightly worse than FastRR, which has a relatively larger MSE. In addition, FASTmrEMMA consumes a median computing time (∼150 s for each replication) among all methods, and much more than FastRR. Therefore, the FastRR algorithm was shown to be a good alternative method for multi-locus GWAS.

Mixed linear model methodologies are mainstream in GWAS; most of them treat QTN effects as fixed effects. In this study, the QTN effects are viewed as random, and it is more consistent with genetic mechanisms (Wen et al., 2018). In order to avoid the influence of the increase of computational complexity, several acceleration techniques have been incorporated into the algorithm. Firstly, we estimate and fix the polygenic-to-residual variance ratio, and then transform the phenotypes and genotypes in the first stage. This technique was adopted in pLARmEB (Zhang et al., 2017) and FASTmrEMMA (Wen et al., 2018), avoiding re-estimating this ratio for each marker. Secondly, the marginal correlation in the second step is similar to the single marker scanning, which quickly filters the unassociated SNPs. The number of SNPs reduces from tens of thousands to hundreds of putative QTNs in the simulation and real data analysis. Thirdly, in the multi-locus model (6), we assume

[8] https://www.arabidopsis.org/

all $\sigma_\gamma{}^2 = \phi^2$, thus only two variance components (ϕ^2 and σ^2) requires DRR to estimate. The results from simulation and real data analysis indicate that the estimation under this simple assumption has achieved better performance for QTN detection and fast computational speed. Lastly, multithreaded marginal correlation is implemented in the FastRR.

Efficient mixed model association and DEMMA as popular single-locus genome scan approaches have been successfully used in GWAS to dissect quantitative traits. However, single-locus approaches ignore the potential information of neighboring markers and fail to consider the joint minor effect of multiple genetic markers on a trait. The FastRR algorithm overcomes this shortcoming. From the results of the simulation, FastRR is more powerful in the detection of QTNs (**Figures 2**, **3**). Although the three popular variable selection approaches, lasso, adaptive lasso, and SCAD, utilize the potential information of markers, the detection and estimation are not accurate (**Tables 1**, **2A–C**). This may be due to the over shrinkage of QTNs, and therefore the effect of QTN is smaller than the true effect; specifically, the minor effect of QTN is shrunk to 0. Consequently, the FastRR algorithm is shown to be more robust in data analysis.

The analysis of large-scale genetic data in GWAS is a hot topic at present. In this study, the correlation coefficients are employed to reduce the dimension of potentially related variables, which are then included in the subsequent multi-locus analysis. The threshold of the correlation coefficient test is set to 0.01 (Tamba et al., 2017), and even the slight correlations between predictors and the response are easily captured. The other thresholds are used, such as 0.001 and 0.0001, which are more rigorous and allows the filtering out of the minor effect loci that will not be included in the multi-locus model. The threshold equal to 0.05 is too loose and includes a large number of SNPs over the threshold; the putative loci are included in the subsequent multi-locus analysis, and furthermore, it is time consuming and results in intractable calculations. Thus, it is reasonable to choose 0.01 as the threshold value in the selection of variables.

DATA AVAILABILITY STATEMENT

The rice data used for the analysis described in this manuscript was obtained from https://doi.org/10.1093/bioinformatics/btaa345; The Arabidopsis data used for the analysis described in this manuscript was obtained from http://www.arabidopsis.usc.edu/.

AUTHOR CONTRIBUTIONS

JZ and JC conceived and supervised the study. JZ, MC, YW, and YL performed all experiments and analyzed the data and revised the manuscript. YZ, MC, SW, and JC made all figures and forms. JZ, YW, and JC also wrote and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by grant 2020Z330 from the Postdoctoral Science Foundation of Jiang Su, grants 31301229 and 32070688 from the National Natural Science Foundation of China, and grant KJQN201414 from the Fundamental Research Funds for the Central Universities.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.649196/full#supplementary-material

REFERENCES

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631.

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721. doi: 10.1038/ng.3007

Dahl, A., Iotchkova, V., Baud, A., Johansson, A., Gyllensten, U., Soranzo, N., et al. (2016). A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* 48, 466–472. doi: 10.1038/ng.3513

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96, 1348–1360. doi: 10.1198/016214501753382273

Goddard, M. E., Wray, N. R., Verbyla, K., and Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* 24, 517–529. doi: 10.1214/09-sts306

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y. M., Todhunter, R. J., et al. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol.* 12:73. doi: 10.1186/s12915-014-0073-5

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835. doi: 10.1038/nmeth.1681

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* 11:e1004969. doi: 10.1371/journal.pgen.1004969

Sun, J., Wu, Q., Shen, D., Wen, Y., Liu, F., Gao, Y., et al. (2019). TSLRF: two-stage algorithm based on least angle regression and random forest in genome-wide association studies. *Sci. Rep.* 9:18034.

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Wang, D., Eskridge, K. M., and Crossa, J. (2011). Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *J. Agric. Biol. Environ. Stat.* 16, 170–184. doi: 10.1007/s13253-010-0046-2

Wang, M., Li, R., and Xu, S. (2020). Deshrinking ridge regression for genome-wide association studies. *Bioinformatics* 36, 4154–4162. doi: 10.1093/bioinformatics/btaa345

Wei, J., Wang, A., Li, R., Qu, H., and Jia, Z. (2018). Metabolome-wide association studies for agronomic traits of rice. *Heredity (Edinb)* 120, 342–355. doi: 10.1038/s41437-017-0032-3

Wen, Y., Zhang, Y., Zhang, J., Feng, J., and Zhang, Y. (2020). The improved FASTmr EMMA and GCIM algorithms for genome-wide association and linkage studies in large mapping populations. *Crop J.* 8, 733–744.

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Wen, Y. J., Zhang, Y. W., Zhang, J., Feng, J. Y., Dunwell, J. M., and Zhang, Y. M. (2019). An efficient multi-locus mixed model framework for the detection of small and linked QTLs in F2. *Brief. Bioinform.* 20, 1913–1924. doi: 10.1093/bib/bby058

Weng, J., Gu, S., Wan, X., Gao, H., Guo, T., Su, N., et al. (2008). Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Res.* 18, 1199–1209. doi: 10.1038/cr.2008.307

Xu, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* 63, 513–521. doi: 10.1111/j.1541-0420.2006.00711.x

Xu, S. (2010). An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity (Edinb)* 105, 483–494. doi: 10.1038/hdy.2009.180

Yi, N., and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179, 1045–1055. doi: 10.1534/genetics.107.085589

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity (Edinb)* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, J., Yue, C., and Zhang, Y. M. (2012). Bias correction for estimated QTL effects using the penalized maximum likelihood method. *Heredity (Edinb)* 108, 396–402. doi: 10.1038/hdy.2011.86

Zhang, Y. M., and Xu, S. (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity (Edinb)* 95, 96–104. doi: 10.1038/sj.hdy.6800702

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* 9:e1003264. doi: 10.1371/journal.pgen.1003264

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101, 1418–1429. doi: 10.1198/016214506000000735

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: 26 April 2021
doi: 10.3389/fgene.2021.648831

AGNEP: An Agglomerative Nesting Clustering Algorithm for Phenotypic Dimension Reduction in Joint Analysis of Multiple Phenotypes

*Fengrong Liu[1,2], Ziyang Zhou[1], Mingzhi Cai[1], Yangjun Wen[1] and Jin Zhang[1,3]**

[1] College of Science, Nanjing Agricultural University, Nanjing, China, [2] School of Data Science, University of Science and Technology of China, Hefei, China, [3] Postdoctoral Research Station of Crop Science, Nanjing Agricultural University, Nanjing, China

Genome-wide association study (GWAS) has identified thousands of genetic variants associated with complex traits and diseases. Compared with analyzing a single phenotype at a time, the joint analysis of multiple phenotypes can improve statistical power by taking into account the information from phenotypes. However, most established joint algorithms ignore the different level of correlations between multiple phenotypes; instead of that, they simultaneously analyze all phenotypes in a genetic model. Thus, they may fail to capture the genetic structure of phenotypes and consequently reduce the statistical power. In this study, we develop a novel method agglomerative nesting clustering algorithm for phenotypic dimension reduction analysis (AGNEP) to jointly analyze multiple phenotypes for GWAS. First, AGNEP uses an agglomerative nesting clustering algorithm to group correlated phenotypes and then applies principal component analysis (PCA) to generate representative phenotypes for each group. Finally, multivariate analysis is employed to test associations between genetic variants and the representative phenotypes rather than all phenotypes. We perform three simulation experiments with various genetic structures and a real dataset analysis for 19 *Arabidopsis* phenotypes. Compared to established methods, AGNEP is more powerful in terms of statistical power, computing time, and the number of quantitative trait nucleotides (QTNs). The analysis of the *Arabidopsis* real dataset further illustrates the efficiency of AGNEP for detecting QTNs, which are confirmed by The *Arabidopsis* Information Resource gene bank.

Keywords: genome-wide association study, statistical power, clustering algorithms, principal component analysis, genetic structure

OPEN ACCESS

Edited by:
Sheng Yang,
Nanjing Medical University, China

Reviewed by:
Wenlong Ren,
Nantong University, China
Qidi Feng,
Broad Institute, United States

****Correspondence:***
Jin Zhang
zhangjin@njau.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *02 January 2021*
Accepted: *01 April 2021*
Published: *26 April 2021*

Citation:
Liu F, Zhou Z, Cai M, Wen Y and Zhang J (2021) AGNEP: An Agglomerative Nesting Clustering Algorithm for Phenotypic Dimension Reduction in Joint Analysis of Multiple Phenotypes. Front. Genet. 12:648831. doi: 10.3389/fgene.2021.648831

Abbreviations: GWAS, genome-wide association study; SNP, single nucleotide polymorphism; QTN, quantitative trait nucleotide; PCA, principal component analysis; AGNES, agglomerative nesting clustering algorithm; AGNEm, AGNES with mean representative phenotypes; AGNEmed, AGNES with median representative phenotypes; AGNEP, AGNES for phenotypic dimension reduction analysis; ANOVA, analysis of variance; MANOVA, multivariate analysis of variance; CLC, cluster linear combination; HCMM, a hierarchical clustering method with mean representative phenotypes.

INTRODUCTION

Genome-wide association study (GWAS) is a powerful tool for exploring associations between genetic variants and phenotypes. To date, GWAS has been successfully applied to human, plant and animal genetic research, to identify thousands of genetic variants related to phenotypes or diseases. Common statistical methods only test the relationships between a single phenotype and loci, that is, only one phenotype is analyzed at a time. Compared to univariate analysis, joint analysis of multiple phenotypes can improve the accuracy and efficiency of the test by using more information from multiple phenotypes (Allison et al., 1998; Zhou and Stephens, 2014), which can be very advantageous for two reasons (Allison et al., 1998; Zhou and Stephens, 2014). First, it promotes computing efficiency. Most of the multi-phenotype methods perform the test for association with all traits, instead of analyzing phenotypes one by one. Joint analysis greatly reduces calculating time and promotes analytical efficiency. Second, multivariate analysis increases statistical power by using genetic structure and potential information among different traits rather than ignoring them as in univariate analysis (Ferreira and Purcell, 2009; Huang et al., 2011). Currently, more and more multivariate analyses have been put forward to analyze the related phenotypes.

The previous studies illustrated that more than 4.6% of single nucleotide polymorphism (SNPs) and 16.9% of genes are reported to be significantly associated with more than one trait (Solovieff et al., 2013). Due to the fact that the joint analysis of multiple phenotypes is more consistent with biological theory (van der Sluis et al., 2013), many multivariate methods have been proposed (Galesloot et al., 2014). O'Brien's method (O'Brien, 1984), one of the earliest methods of jointly analyzing multiple phenotypes, can be used to integrate the results of univariate association tests. If the means of individual statistics are homogeneous, O'Brien's method is more effective among linear combination statistics. Multivariate analysis of variance (MANOVA) (Cole et al., 1994) is a classic method of analyzing multiple phenotypes that jointly tests whether the independent variables explain the variance of the dependent variables statistically significant at the same time. Subsequently, Multiphen (O'Reilly et al., 2012) and TATES (van der Sluis et al., 2013) are powerful to test associations between genetic variants and corresponding multiple traits. Under the framework of linear mixed models, multi-trait mixed model (Korte et al., 2012) and multivariate linear mixed model (Zhou and Stephens, 2014) are proposed, which take into account the variance components of multiple phenotypes and the population structure in GWAS.

However, established procedures for analyzing multiple phenotypes face several challenges from the following perspectives. First, computing is infeasible. Hundreds and thousands of phenotypes are being collected in biological experiments and surveys. However, most methods become computationally intractable or hard to implement as the number of phenotypes increases (Dahl et al., 2016). Second, estimates are inaccurate. The complexity and the number of parameters increase sharply in joint analysis of more than 10 phenotypes, and hence accuracy and statistical stability decrease (Solovieff et al., 2013). Finally, most multivariate algorithms simultaneously analyze all phenotypic data and thus might ignore different level of correlation or homogeneous genetic basis among traits, resulting in an unsatisfactory power (Liang et al., 2018).

Clustering algorithm is an alternative method of overcoming these challenges. It aims to maximize homogeneity within a cluster so that similarity is greater between elements in the same cluster than those in different clusters. As the dimension of the data is reduced by clustering, temporal and spatial complexity decreases. In addition, the intragroup phenotypic correlation is stronger than the intergroup correlation, which improves the efficiency and accuracy of the statistical test. Therefore, clustering is great importance to the study of the joint analysis of high-dimensional phenotypes. Recently, Sha et al. (2019) proposed the cluster linear combination (CLC) method, which groups phenotypes and then analyzes quadratic combination of individual data. CLC takes full advantage of similar genetic information in the same group. However, CLC does not work well with negative or mixed correlations.

In this study, we propose a new method agglomerative nesting clustering algorithm for phenotypic dimension reduction analysis (AGNEP), which uses an agglomerative nesting (AGNES) clustering algorithm to group multiple correlated phenotypes and then applies principal component analysis (PCA) to generate representative phenotypes for each group. Finally, MANOVA is employed to test associations between genetic variants and the representative phenotypes rather than all phenotypes. In three simulation experiments, we consider six scenarios under three kinds of genetic structures to compare the performance of different methods: MANOVA, analysis of variance (ANOVA), a hierarchical clustering method with mean representative phenotypes (HCMM), AGNEP, AGNES with mean representative phenotypes (AGNEm), and AGNES with median representative phenotypes (AGNEmed). All of these methods are applied to analyze 19 traits of *Arabidopsis* real dataset. AGNEP is validated by the analysis of real dataset and the series of simulation experiments.

MATERIALS AND METHODS

Genetic Model

Consider the multivariate linear model:

$$Y_{(d\times n)} = \alpha W_{(d\times n)} + B_{(d\times 1)} X_{(1\times n)} + E_{(d\times n)} \tag{1}$$

where $Y_{d\times n} = (Y_1, \ldots, Y_d)^T$ is a $d \times n$ matrix of phenotypes, n is the number of individuals and d is the number of phenotypes; $Y_i = (y_{i1}, \ldots, y_{in})^T$ is the i^{th} phenotype of n individuals. α is the intercept and $W_{d\times n}$ is a $d \times n$ matrix with elements of 1. B is a d-vector of effect sizes for the d phenotypes, which are considered as fixed effects. $X_{1\times n} = (x_1, \ldots, x_n)$ is an n-vector of genotypes for a particular marker, and x_j is denoted as the number of minor alleles that the j^{th} individual carries at the variant. $E_{(d\times n)} \sim MN_{(d\times n)}(0, V, I_n)$ is a $d \times n$ matrix of residual error. $MN_{d\times n}(0, V, I_n)$ denotes the $d \times n$ matrix normal distribution with mean 0, row covariance matrix V (a $d \times d$ symmetric matrix of environmental variance component) and column covariance matrix I_n (an $n \times n$ identity matrix).

Clustering Algorithms

Generally, hundreds or even thousands of phenotypes are cataloged from biological experiments and surveys. However, either these phenotypic data are analyzed separately by univariate analysis, or all phenotypes are analyzed without distinction. This creates some challenges for the statistical analysis, such as a reduction in statistical power, inflexibility in the computational analysis, a high computing time, and so on. From the perspective of multi-phenotype joint analysis, grouping high-dimensional phenotypic data by clustering algorithms is an alternative to overcome above challenges (Fung, 2001). Here we integrate clustering algorithms, AGNES into analysis of multiple phenotypes.

Hierarchical clustering algorithm creates a tree-like cluster structure based on the similarity between samples. In general, two partitioning strategies are possible according to the direction of hierarchical decomposition, that is, agglomerative (bottom up) and divisive (top down). The agglomerative method starts with all samples in their own clusters and then groups two clusters with the greatest similarity until only one cluster remains. The divisive method adopts an inverse procedure with agglomerative method (Liang et al., 2018).

AGNES is a typical hierarchical clustering algorithm, which implements bottom-up strategy until a preset criterion is satisfied (Deng et al., 2018). The similarity between Y_i and Y_j is evaluated by formula (2). The minimum distance is calculated by formula (3) to measure the similarity of clusters c_i and c_j (Murtagh and Legendre, 2014).

$$dist\left(Y_i, Y_j\right) = ||Y_i - Y_j||_2 = \sqrt{\sum\nolimits_{t=1}^{n} |y_{(it)} - y_{(jt)}|^2} \tag{2}$$

$$dist_{min}\left(c_i, c_j\right) = \underset{p\in c_i, q\in c_j}{min} dist(p, q) \tag{3}$$

where Y_i is the i^{th} phenotype; $c_i = (c_{i1}, ..., c_{in})^T$ is the i^{th} cluster; p is a sample belonging to cluster c_i, and q is a sample belonging to cluster c_j.

The Optimal Number of Clusters *K*

In this study, the optimal number of clusters K is calculated according to the maximum silhouette coefficient s, which is an index used to evaluate the clustering algorithm (Rousseeuw, 1987). The silhouette coefficient combines two factors, cohesion and resolution. Assuming all phenotypes are divided into K clusters by using AGNES, for each sample, we assume that Y_i belongs to the cluster c_k, we can calculate the silhouette coefficient s as formula (4):

$$s\left(i\right) = \frac{b(i) - a(i)}{max\left(b(i), a(i)\right)} \tag{4}$$

$$a(i) = \begin{cases} \frac{1}{|c_k|-1} \sum\nolimits_{p\in c_k, p\neq Y_i} dist(Y_i, p), & |c_k| > 1 \\ 0, & |c_k| = 1 \end{cases}$$

$$b(i) = \underset{c_d\neq c_k}{min}\ dist(Y_i, c_d) = \frac{1}{|c_d|} \sum\nolimits_{q\in c_d}, (Y_i, q)$$

where $s\left(i\right)$ is the silhouette coefficient of the sample Y_i, $s\left(i\right)$ ranges from −1 to 1, and $|c_k|$ is the number of phenotypes in cluster c_k.

Obviously, $s\left(i\right)$ close to 1 indicates that the distance within a cluster is small and the distance between clusters is large, that is, relatively better clustering results. The silhouette coefficient s is the average of silhouette coefficient of all samples, $s = d^{-1} \sum_{i=1}^{d} s\left(i\right)$. The optimal classification, say K clusters, is determined according to the maximum characteristics of the silhouette coefficient. In this study, the number of clusters K ranges from 2 to d−1, which means two situations are not considered, each phenotype is a cluster, and all phenotypes are clustered into one cluster.

Representative Phenotypes of Clusters

In the following multivariate analysis, representative phenotype(s) are analyzed instead of all phenotypes by three ways: (i) the mean of each group (AGNEm), (ii) the median of each group (AGNEmed), and (iii) the top principal components of each group (AGNEP).

We scale each phenotype for each cluster and define the representative phenotype for the k^{th} cluster as the average or median phenotypic value within the group using formula (5) and (6):

$$Y_{mean}^{k} = \frac{1}{|c_k|} \sum_{Y_i\in c_k} Y_i \tag{5}$$

$$Y_{median}^{k} = \underset{Y_i\in c_k}{median}\, Y_i \tag{6}$$

In addition, top m principal components $Y_{PCA}^{k} = \left(Y_{PCA}^{k1}, ..., Y_{PCA}^{km}\right)$ with a cumulative contribution rate over 85% (Xue, 2007) are regarded as the representative phenotypes for the k^{th} cluster.

Experimental Materials

Three simulation experiments are conducted to evaluate the performances of AGNEP and other methods. We generate genotypes according to the minor allele frequency in the interval [0.1, 0.5] under Hardy–Weinberg equilibrium. The simulation datasets contain n = 5000 individuals with m = 10,000 genetic variants, which are generated by using the factor model (Sha et al., 2019). We consider two scenarios for each simulation, including 10 quantitative trait nucleotides (QTNs) for scenario 1 and 50 QTNs for scenario 2.

In simulation experiment I, 20 phenotypes are divided into five independent clusters (**Table 1**). Each cluster consists of four phenotypes based on genetic correlation (**Figure 1A**). In simulation experiment II, we consider a pervasive genetic structure. The adjacent clusters have overlapping phenotypes, and the overlapped phenotypes share the same or similar genetic basis. Twenty phenotypes are divided into five correlated clusters (**Table 1**). Group 1 and group 2 share two phenotypes, group 3 is independent

TABLE 1 | Different genetic structures for three simulation experiments, including five independent clusters (simulation I), five dependent clusters (simulation II), and eight dependent clusters of high-dimensional phenotypes (simulation III).

Simulation experiments	Simulation setting								
	Clustering	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
I	No. of phenotypes	1–4	5–8	9–12	13–16	17–20			
II	No. of phenotypes	1–5	4–8	9–12	13–18	16–20			
III	No. of phenotypes	1–15	10–30	31–40	41–60	61–75	70–90	85–95	90–100

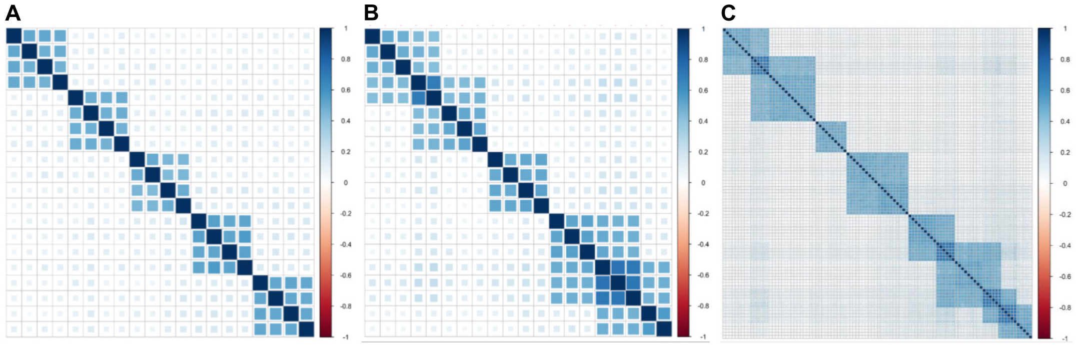

FIGURE 1 | Genetic correlations for three simulation experiments. **(A)** Five independent clusters (simulation I). **(B)** Five dependent clusters (simulation II). **(C)** Eight dependent clusters of high-dimensional phenotypes (simulation III).

with the other groups, and group 4 shares three phenotypes with group 5 (**Figure 1B**). In simulation experiment III, we focus on high-dimensional phenotypes with more complex correlations. All 100 phenotypes are divided into eight phenotypic groups. The genetic correlations are exhibited in **Figure 1C**. The high-dimensional correlations are more complicated than the correlations in the previous two simulation experiments.

Arabidopsis Real Dataset

We reanalyze the *Arabidopsis thaliana* (Atwell et al., 2010) dataset, including 199 diverse inbred lines, each of which has 216,130 SNPs and 107 phenotypes. To evaluate the performance of different methods, we focus on 19 quantitative phenotypes: days to flowering under long days (LD), days to flowering under LD with vernalization (LDV), days to flowering under short days (SD), days to flowering under SD with vernalization (SDV), days to flowering at 10, 16, and 22°C (FT10, FT16, and FT22), days to flowering with 8 weeks vernalization in greenhouse (8WGHFT), leaf number at flowering with 8 weeks vernalization in greenhouse (8WGHLN), days to flowering in field (FTF), diameter of plants at flowering in field (FTD), leaf number at 10, 16, and 22°C (LN10, LN16, and LN22), plant diameter at 10, 16, and 22°C (Width10, Width16, and Width22), and presence of leaf serration at 16 and 22°C (Leafserr16 and Leafserr22). We filter out SNPs with minor allele frequency less than 5% and each individual with missing phenotypic data. After quality control, the data consist of 206,603 SNPs and 137 individuals. The genetic structure of the phenotypic data is shown in **Figure 2**.

RESULTS

Simulation Results

To evaluate the performance of the following multivariate methods (MANOVA, HCMM, AGNEP, AGNEm, and AGNEmed) and univariate method (ANOVA), we conduct three simulations: independent phenotypic groups in simulation I (**Figure 1A**), correlated groups in simulation II (**Figure 1B**), and high-dimensional phenotypes divided into eight groups in simulation III (**Figure 1C**).

Statistical Power for Detection

In the three simulations, 10 (scenario 1) and 50 (scenario 2) QTNs are simulated in each dataset. For simulation I (independent groups), **Figures 3A,B** show the significant advantages of all multivariate analysis over the univariate analysis (ANOVA). According to the optimal silhouette coefficient of clustering algorithm (**Supplementary Figure 1**), the power under various FDR is higher for AGNEP than the other methods in simulation I. MANOVA easily captures the independent genetic structure of 10 QTNs (**Figure 3A**) and has slightly higher power than HCMM, AGNEm, and AGNEmed. In scenario 2, the multivariate analysis based on clustering algorithm obviously outperforms than MANOVA (**Figure 3B**). The clustering results for AGNEm and HCMM are completely consistent with the optimal silhouette coefficient, thus, these two methods have the same power, and their curves are overlapping in **Figures 3A,B**. From the results of simulation I, we conclude that AGNEP seems slightly more robust and multivariate algorithms easily capture genetic information for independent groups.

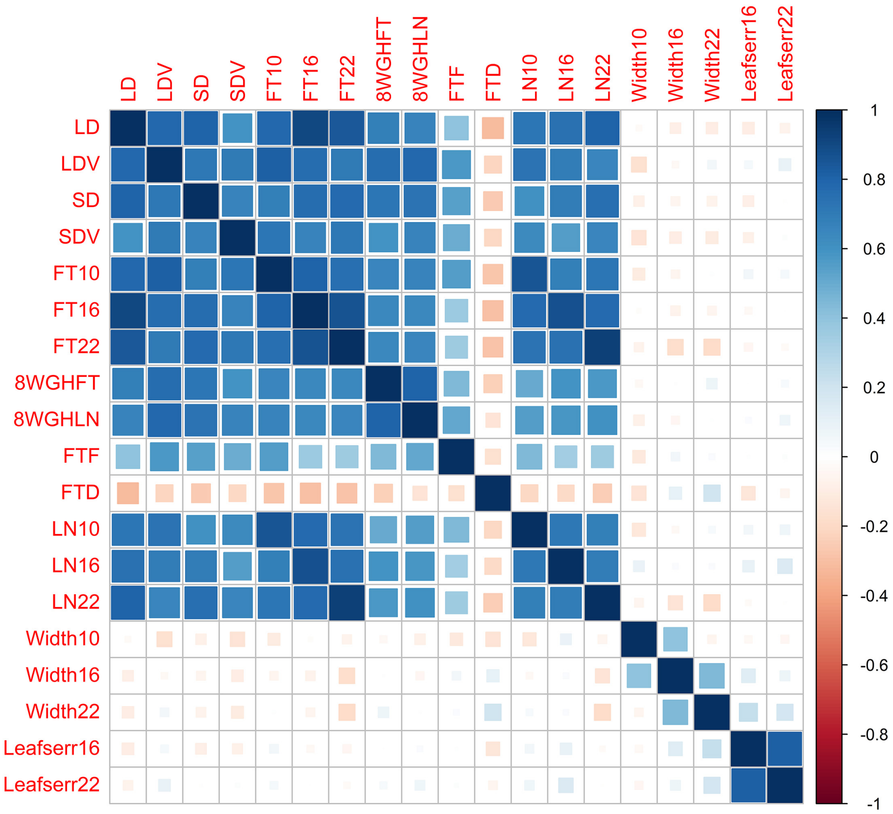

FIGURE 2 | Genetic correlations between 19 phenotypes in the *Arabidopsis* dataset.

For simulations II (related groups) and III (high-dimensional related groups), the powers of almost all multivariate algorithms are significantly higher than that of the univariate analysis (ANOVA; **Figures 3C–F**). AGNEP has higher power and more significant detection in simulations II and III, which is followed by HCMM, MANOVA, AGNEm, AGNEmed, and ANOVA. In addition, the results of simulations II and III show that the power of AGNEm and AGNEmed are even worse than MANOVA and similar to ANOVA. It is evident that different representative phenotypes achieve significantly different results under the same clustering algorithm, and PCA appears to be a powerful tool for flexibly taking full advantage of potential information. Moreover, this difference becomes more and more obvious with the increase in the number of phenotypes, the complexity of the genetic structure, and the number of QTNs. The results of the three simulations demonstrate the superior power of AGNEP over all the other methods under various genetic structures.

Computing Time

The computing times of the different methods in the three simulations are shown in **Figure 4**. For analyses of multiple phenotypes based on different clustering algorithms, the computing times are in the same magnitude, which are less than MANOVA and ANOVA. However, as the number of phenotypes increases, the differences among the methods are more and more obvious. The results of the three simulations illustrate that AGNEP effectively captures potential information and reduces the computing complexity. In particular, AGNEP is recommended for high-dimensional phenotypes and complex related structures.

Real Data Analysis

To further evaluate the performance of the different methods, we analyze an *Arabidopsis* real dataset with 19 quantitative phenotypes including LD, LDV, SD, SDV, FT10, FT16, FT22, 8WGHFT, 8WGHLN, FTF, FTD, LN10, LN16, LN22, Width10, Width16, Width22, Leafserr16, and Leafserr22. All phenotypes are related to flower, leaf, plant growth, and the presence of leaf serration. After filtering, the dataset consists of 137 samples and a total of 206,603 SNPs. The genetic correction of the phenotypic data is shown in **Figure 2**.

QTNs Detected

The numbers of putative QTNs for the six different methods are calculated by 10 permutations (**Figure 5**). Based on the maximum silhouette coefficient, AGNEP detects more putative QTNs than the other five methods, and the other multivariate algorithms and ANOVA have relatively poor detection ability. The results of the *Arabidopsis* real dataset show similar trends to simulation III. This may result from that the genetic structures are relatively complex, and the other methods cannot effectively capture this type of information, so their performances are not satisfactory.

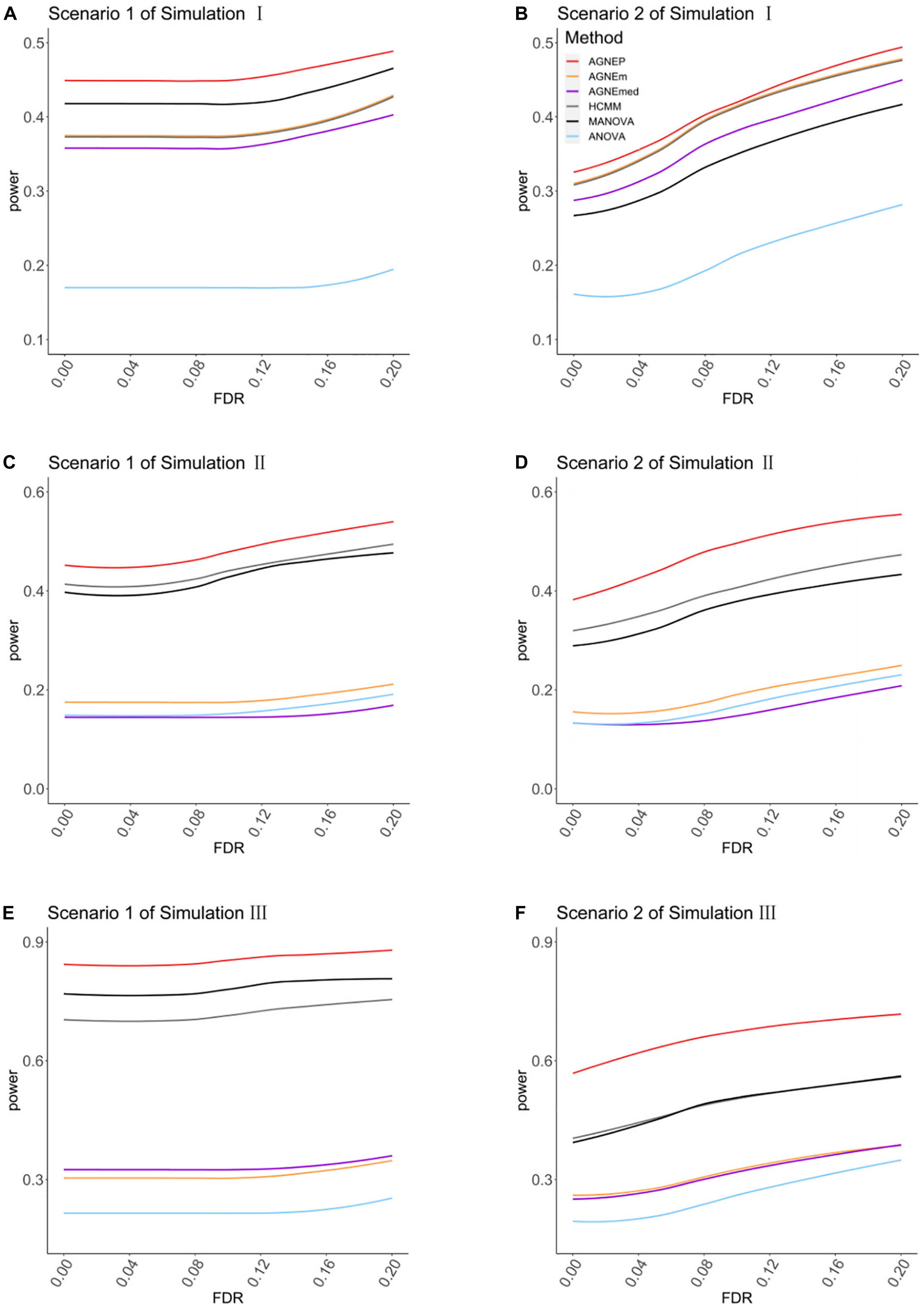

FIGURE 3 | The comparison of the power for AGNEP and established approaches. **(A,B)** The powers of simulation experiment I are presented. **(C,D)** The powers of simulation experiment II are presented. **(E,F)** The powers of simulation experiment III are presented. Scenario 1 and 2 indicate that 10 and 50 QTNs are simulated in the three simulations, respectively.

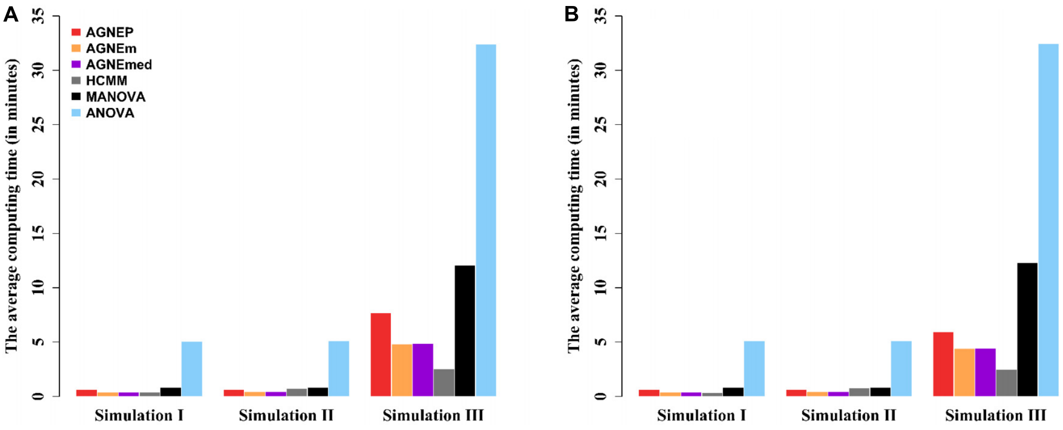

FIGURE 4 | The average computing time (in minutes) of the six methods in three simulations. Scenario 1 **(A)** and 2 **(B)** indicate that 10 and 50 QTNs are simulated, respectively.

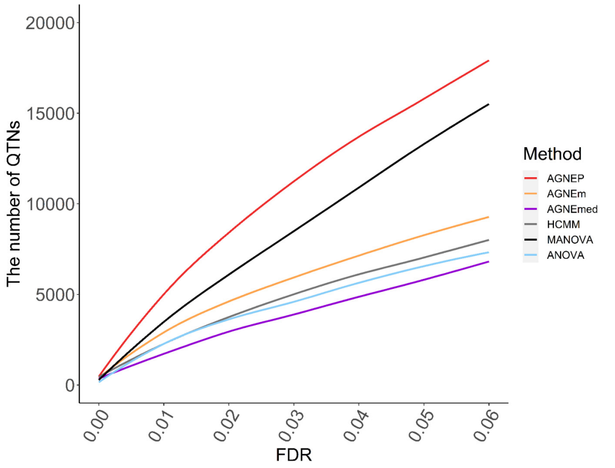

FIGURE 5 | Numbers of QTNs under various FDR detected by six methods for the *Arabidopsis* dataset.

Manhattan Plots

Manhattan plots of the *Arabidopsis* analysis are shown in **Supplementary Figures 2,3**. For ANOVA (**Supplementary Figure 2**), the QTNs related to phenotypes associated with flower and plant growth can be detected, whereas the QTNs related to other phenotypes have relatively low *P*-value. The results of statistical tests of AGNEP, AGNEm, AGNEmed, and HCMM (**Supplementary Figure 3**) show similar patterns, and several genomic regions reach the Bonferroni corrected threshold ($-\log_{10}(0.001/206603) = 8.3151$). According to the results for confirmed *Arabidopsis* genes, MANOVA detects more false associated SNPs. Therefore, compared to the univariate method, multivariate methods have the ability to increase statistical power. Moreover, multivariate methods based on the clustering algorithm further improve detection ability and accuracy by using information about complex genetic structure.

TABLE 2 | Average computing time (in minutes) and number of confirmed genes in analysis of the *Arabidopsis* dataset by six different methods.

Method	Number of confirmed genes	Computing time
AGNEP	453	91.33
AGNEm	386	113.49
AGNEmed	373	95.64
HCMM	321	105.05
MANOVA	315	110.72
ANOVA	159	788.15

Genomic Patterns

According to the results of the 19 traits of *Arabidopsis*, all significant QTNs are listed in **Figure 6** as hot spots, which illustrate information about the overall genomic patterns of significant SNPs (QTNs) on multiple traits. Almost all multivariate methods have the similar pattern. Compared to univariate method, multivariate methods easily identify associations between QTNs and phenotypes. This figure shows the genetic basis of functional relationships between phenotypes. These hot spots would be the primary targets for functional analysis and for genetic improvement by selection.

Confirmed Genes

To further validate the AGNEP method, we compare the number of candidate genes detected by six methods for the *Arabidopsis* dataset. All SNPs under 0 FDR within 20 kB of each putative QTN are used to mine the candidate genes by The *Arabidopsis* Information Resource[1]. **Table 2** shows the quantity of confirmed genes for all approaches (Hagemann and Gleissberg, 1996; Wang et al., 2003; Nikovics et al., 2006; Albayrak et al., 2012; Nakayama et al., 2012). AGNEP detects the largest number of confirmed genes, 453, followed by HCMM (439), AGNEm (386), AGNEmed (373), MANOVA (315), and ANOVA (159).

[1] https://www.arabidopsis.org/

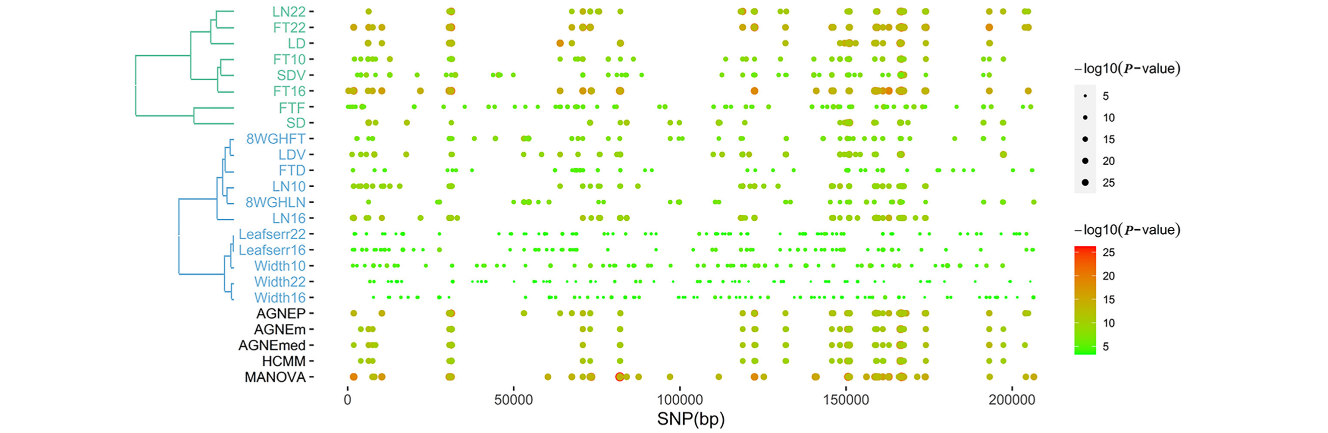

FIGURE 6 | The plot of top 100 significant SNPs of six methods in the analysis of 19 traits. (Left) The hierarchical cluster diagram of 19 traits. (Right) The bubble chart of *P*-value.

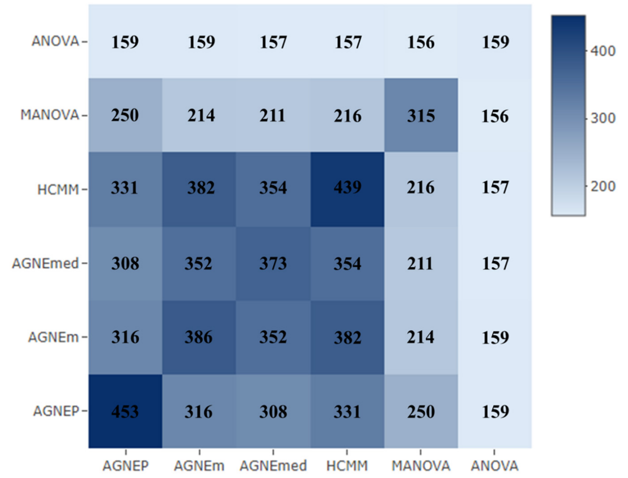

FIGURE 7 | The heat map of confirmed genes for the six methods in analysis of the *Arabidopsis* real dataset. The darker the square, the greater the number of confirmed genes detected two methods.

A heat map (**Figure 7**) illustrates the confirmed candidate genes simultaneously detected by two methods. It is obvious that the multivariate methods detect more identical confirmed genes than the univariate method (ANOVA). Furthermore, multivariate methods based on a clustering algorithm, say AGNEP, AGNEm, AGNEmed, and HCMM, detect more than 350 confirmed genes.

Computing Time

The computing time of each approach for the 19 *Arabidopsis* traits is listed in **Table 2**. Apparently, all the multivariate methods are faster than the univariate method, which consumes about seven to eight times longer than the multivariate methods. The multivariate analysis greatly reduce the calculating time and promotes analytical efficiency. AGNEP and AGNEmed have the shortest running time, less than 100 minutes; HCMM, AGNEm, and MANOVA have moderate computing times. All in all, AGNEP not only performs best in QTNs detection, but also has the fastest computing speed, which is validated by the analysis of the real dataset.

DISCUSSION

In this study, we propose a new method called AGNEP, which applies AGNES clustering algorithms and PCA to detect genetic associations between SNPs and multiple phenotypes in GWAS. The results of three simulations and a real data analysis indicate the merits of AGNEP. There are three main advantages. First, AGNEP easily captures the correlation of multiple phenotypes by clustering methods, which increases statistical power in analysis of simulations and *Arabidopsis* dataset (**Figures 3**, **5**). Second, the detection accuracy of AGNEP is significantly improved. From the *Arabidopsis* dataset, AGNEP detects the most confirmed genes, obviously more than the other established methods. Third, because of the decrease in phenotypic dimension and the optimization of representative phenotypes, AGNEP enjoys fast computing speed, even with high-dimensional phenotypes and complex genetic structures.

To further validate the new method, we incorporate representative phenotypes into seven different clustering methods, including K-means, PAM, CLARA, HCDS, HCM, FCM, and EM algorithms. All of these methods are used to reanalyze the simulated datasets and *Arabidopsis* real data. The PCA-based methods are more robust than the methods, MANOVA and ANOVA from the perspective of power (simulation results, **Supplementary Figure 4**; *Arabidopsis* results, **Supplementary Figure 5**), efficiency (**Supplementary Table 1**), and detection of confirmed genes (**Supplementary Table 2**). However, all of these methods perform slightly worse than AGNEP in the simulations and real data analysis. Furthermore, CLC is used to comparing, which appears a tremendous increase in computational burden along with permutation and the number of phenotypes, and thus the simulation I and II datasets are analyzed. Nevertheless, the performance of CLC is unsatisfactory in terms of statistical power and efficiency.

Essentially, the representative phenotypes of PCA are linear combinations of individual phenotypic data in the same cluster. When the cluster consists of highly positively correlated phenotypes, all the linear combinations can represent the cluster reasonably well (Bühlmann et al., 2013; Shah and Samworth, 2013). To further validate PCA combinations, the mixed (both positive and negative) correlations are induced to simulation II. The PCA-based methods are better than the mean and median, and ANOVA has the lowest power (**Supplementary Figure 7**). For mixed and complex correlated phenotypes, the results demonstrate the good performance of the PCA combinations as well (**Figure 3** and **Supplementary Figure 7**). This is because the PCA combinations consist of the most within-cluster information and reduce the phenotypic dimensions. It is necessary to further explore other representative phenotypes forms, such as quadratic and non-linear combinations.

With the development of life sciences and biotechnology, genetic data is becoming larger in scale and more complicated. How to cluster phenotypes efficiently and accurately is very important. In this study, the silhouette coefficient is a key index for evaluating the clustering model and determining the optimal number of clusters. In addition to the silhouette coefficient, many other criteria can be used to evaluate the model, such as Calinski-Harabaz, Dunn validity, and Davies-Bouldin. Silhouette coefficient is recommended according to empirical analysis.

DATA AVAILABILITY STATEMENT

The Arabidopsis data used for the analysis described in this manuscript was obtained from http://www.arabidopsis.usc.edu/.

AUTHOR CONTRIBUTIONS

JZ conceived and supervised the study and wrote and revised the manuscript. FL and ZZ performed all experiments, analyzed the data, and wrote the manuscript. MC and YW mined candidate genes from The *Arabidopsis* Information Resource in the *Arabidopsis* data analysis and created all figures and tables. All authors reviewed the manuscript.

FUNDING

This work was supported by grant 2020Z330 from the Postdoctoral Science Foundation of Jiang Su, grant 31301229 and 32070688 from the National Natural Science Foundation of China, and grant KJQN201414 from the Fundamental Research Funds for the Central Universities.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.648831/full#supplementary-material

REFERENCES

Albayrak, I., Nikora, V., Miler, O., and O'Hare, M. (2012). Flow-plant interactions at a leaf scale: effects of leaf shape, serration, roughness and flexural rigidity. *Aquatic Sci.* 74, 267–286. doi: 10.1007/s00027-011-0220-9

Allison, D. B., Thiel, B., St Jean, P., Elston, R. C., Infante, M. C., and Schork, N. J. (1998). Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* 63, 1190–1201. doi: 10.1086/302038

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631.

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2013). Correlated variables in regression: clustering and sparse estimation. *J. Stat. Plan. Inference* 143, 1835–1858. doi: 10.1016/j.jspi.2013.05.019

Cole, D. A., Maxwell, S. E., Arvey, R., and Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychol. Bull.* 115, 465–474. doi: 10.1037/0033-2909.115.3.465

Dahl, A., Iotchkova, V., Baud, A., Johansson, Å, Gyllensten, U., and Soranzo, N. (2016). A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* 48, 466–472. doi: 10.1038/ng.3513

Deng, L., Tan, T., Han, J., and Tian, T. (2018). IAGNES algorithm for protocol recognition. *High Technol. Lett.* 24, 408–416.

Ferreira, M. A., and Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics* 25, 132–133. doi: 10.1093/bioinformatics/btn563

Fung, G. (2001). *A Comprehensive Overview of Basic Clustering Algorithms, Technical Report*. Madison, WI: University of Winsconsin.

Galesloot, T. E., Kristel, V. S., Kiemeney, L. A. L. M., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLoS One* 9:e95923. doi: 10.1371/journal.pone.0095923

Hagemann, W., and Gleissberg, S. (1996). Organogenetic capacity of leaves: the significance of marginal blastozones in angiosperms. *Plant Syst. Evol.* 199, 121–152. doi: 10.1007/bf00984901

Huang, J., Johnson, A. D., and O'Donnell, C. J. (2011). PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics* 27, 1201–1206. doi: 10.1093/bioinformatics/btr116

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* 44, 1066–1071. doi: 10.1038/ng.2376

Liang, X., Sha, Q., Yeonwoo, R., and Zhang, S. (2018). A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genet. Epidemiol.* 42, 344–353. doi: 10.1002/gepi.22124

Murtagh, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* 31, 274–295. doi: 10.1007/s00357-014-9161-z

Nakayama, H., Yamaguchi, T., and Tsukaya, H. (2012). Acquisition and diversification of cladodes: leaf-like organs in the genus *Asparagus*. *Plant Cell* 24, 929–940. doi: 10.1105/tpc.111.092924

Nikovics, K., Blein, T., Peaucelle, A., Ishida, T., Morin, H., Aida, M., et al. (2006). The balance between the MIR164A and CUC2 genes controls leaf margin serration in *Arabidopsis*. *Plant Cell* 18, 2929–2945. doi: 10.1105/tpc.106.045617

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* 40, 1079–1087. doi: 10.2307/2531158

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M. R., et al. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* 7:e34861. doi: 10.1371/journal.pone.0034861

Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Sha, Q., Wang, Z., Zhang, X., and Zhang, S. (2019). A clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *Bioinformatics* 35, 1373–1379. doi: 10.1093/bioinformatics/bty810

Shah, R. D., and Samworth, R. J. (2013). Discussion of 'correlated variables in regression: clustering and sparse estimation' by Peter Bühlmann, Philipp Rütimann, Sara van de Geer and Cun-Hui Zhang. *J. Stat. Plann. Inference* 143, 1866–1868. doi: 10.1016/j.jspi.2013.05.022

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *PLoS Genetics* 14:483–495. doi: 10.1038/nrg3461

van der Sluis, S., Posthuma, D., and Dolan, C. V. (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* 9:e1003235. doi: 10.1371/journal.pgen.1003235

Wang, L., Li, J., Zhan, J., and Huang, W. (2003). Effects of salicylic acid on photosynthesis and assimilate distribution of grape seedlings under heat stress. *Plant Physiol. Commun.* 39, 215–216.

Xue, Y. (2007). *Statistical Modeling and R Software*. Beijing: Tsinghua University Press.

Zhou, X., and Stephens, M. (2014). Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nat. Methods* 11, 407–409. doi: 10.1038/nmeth.2848

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.

ORIGINAL RESEARCH
published: 20 July 2021
doi: 10.3389/fgene.2021.661109

Prediction of BRCA Gene Mutation in Breast Cancer Based on Deep Learning and Histopathology Images

Xiaoxiao Wang[1†], Chong Zou[1†], Yi Zhang[2,3†], Xiuqing Li[4], Chenxi Wang[4], Fei Ke[4], Jie Chen[4], Wei Wang[1], Dian Wang[1], Xinyu Xu[2,3], Ling Xie[4] and Yifen Zhang[4*]*

[1] Department of GCP Research Center, Jiangsu Province Hospital of Chinese Medicine, The Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China, [2] Department of Pathology, Jiangsu Cancer Hospital, Nanjing, China, [3] Jiangsu Institute of Cancer Research, The Affiliated Cancer Hospital of Nanjing Medical University, Nanjing, China, [4] Department of Pathology, Jiangsu Province Hospital of Chinese Medicine, The Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China

OPEN ACCESS

Edited by:
Shiquan Sun,
Xi'an Jiaotong University, China

Reviewed by:
Mantang Qiu,
Peking University People's Hospital, China
Rong Yin,
The Affiliated Cancer Hospital of Nanjing Medical University, China
Zaneta Swiderska-Chadaj,
Warsaw University of Technology, Poland

****Correspondence:***
Ling Xie
xl6250@163.com
Yifen Zhang
yifen2001@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 January 2021*
Accepted: *07 June 2021*
Published: *20 July 2021*

Citation:
Wang X, Zou C, Zhang Y, Li X, Wang C, Ke F, Chen J, Wang W, Wang D, Xu X, Xie L and Zhang Y (2021) Prediction of BRCA Gene Mutation in Breast Cancer Based on Deep Learning and Histopathology Images. Front. Genet. 12:661109. doi: 10.3389/fgene.2021.661109

Background: Breast cancer is one of the most common cancers and the leading cause of death from cancer among women worldwide. The genetic predisposition to breast cancer may be associated with a mutation in particular genes such as gene BRCA1/2. Patients who carry a germline pathogenic mutation in BRCA1/2 genes have a significantly increased risk of developing breast cancer and might benefit from targeted therapy. However, genetic testing is time consuming and costly. This study aims to predict the risk of gBRCA mutation by using the whole-slide pathology features of breast cancer H&E stains and the patients' gBRCA mutation status.

Methods: In this study, we trained a deep convolutional neural network (CNN) of ResNet on whole-slide images (WSIs) to predict the gBRCA mutation in breast cancer. Since the dimensions are too large for slide-based training, we divided WSI into smaller tiles with the original resolution. The tile-based classification was then combined by adding the positive classification result to generate the combined slide-based accuracy. Models were trained based on the annotated tumor location and gBRCA mutation status labeled by a designated breast cancer pathologist. Four models were trained on tiles cropped at 5×, 10×, 20×, and 40× magnification, assuming that low magnification and high magnification may provide different levels of information for classification.

Results: A trained model was validated through an external dataset that contains 17 mutants and 47 wilds. In the external validation dataset, AUCs (95% CI) of DL models that used 40×, 20×, 10×, and 5× magnification tiles among all cases were 0.766 (0.763–0.769), 0.763 (0.758–0.769), 0.750 (0.738–0.761), and 0.551 (0.526–0.575), respectively, while the corresponding magnification slides among all cases were 0.774 (0.642–0.905), 0.804 (0.676–0.931), 0.828 (0.691–0.966), and 0.635 (0.471–0.798), respectively. The study also identified the influence of histological grade to the accuracy of the prediction.

Conclusion: In this paper, the combination of pathology and molecular omics was used to establish the gBRCA mutation risk prediction model, revealing the correlation between the whole-slide histopathological images and gRCA mutation risk. The results indicated that the prediction accuracy is likely to improve as the training data expand. The findings demonstrated that deep CNNs could be used to assist pathologists in the detection of gene mutation in breast cancer.

Keywords: breast cancer, BRCA gene, deep learning, artificial intelligence, digital pathology

INTRODUCTION

Female breast cancer (BC) made up 11.7% of 19.3 million new cancer cases in 2020 and has overtaken lung cancer as the most diagnosed cancer globally, and ranks as the fourth leading cause of cancer-related mortality, according to a report from the International Agency for Research on Cancer (Sung et al., 2021). BC is a heterogeneous collection of diseases with various incidences, risk factors, genetic, prognosis, and treatment responses. Genetic susceptibilities to BC may be associated with mutations in a specific gene or a series of genes, including the key tumor suppressor gene BRCA (BRCA1 or BRCA2). BRCA1/2 mutation may be inherited (germline, gBRCA) or may arise *de novo* because of a combination of genetic and environmental factors (somatic) (Engel and Fischer, 2015). The frequency of these genetic mutations varies among different countries and ethnic groups. A study of a large cohort of a Chinese population shows that the BRCA mutation rate was 9.1% in BC patients with at least one risk factor, 3.5% in sporadic patients, and 0.38% in healthy controls (Lang et al., 2017). BRCA1/2 plays an essential role in DNA damage response, DNA double-strand break, repair, transcriptional regulation, etc. Loss of BRCA1/2 educes impairment of the homologous recombination DNA repair pathway, thereby leading to genomic instability which may ultimately contribute to cancer development. Patients who carry a germline pathogenic mutation in the BRCA1/2 gene have a significantly increased risk of developing BC and other cancers (e.g., ovarian, pancreatic, and prostate cancer) (Paul and Paul, 2014). Previous meta-analyses of published trials show that BRCA1 and BRCA2 carriers have a 57–65% and 45–49% probability of developing BC over lifetime, respectively. Furthermore, if there is a positive family history of BC, this risk increases to 85 and 84%, respectively (Antoniou et al., 2003; Chen and Parmigiani, 2007).

BCs with BRCA1/2 mutations are different from sporadic BC in clinical and pathological features. Patients with gBRCA1 mutations have a higher prevalence of triple-negative (absence of estrogen receptor, progesterone receptor, and HER-2 expression), invasive ductal carcinoma with medullary features (Sønderstrup et al., 2018). The multivariate analysis revealed that morphological features predictive of the BRCA1 phenotype include the presence of lymphocytic infiltrate, higher mitotic figures, and pushing margins compared with sporadic BC (Atchley et al., 2008). BRCA2 tumors are also more frequently higher histological grade compared with sporadic tumors. However, the unique characteristic that is significant for BRCA2-associated BC is lack of tubule formation and pushing margins (Atchley et al., 2008). The detection of a pathogenic gBRCA mutation in a woman diagnosed with BC may affect her current cancer treatment and prognosis, but it can also prevent future cancers and identify healthy mutation carriers in their family members (Metcalfe et al., 2014; Faraoni and Graziani, 2018; Torrisia et al., 2019). Knowing one's gBRCA status plays an important role for healthy women, because cancer can be prevented by risk-reducing mastectomy and salpingo-oophorectomy (Domchek et al., 2010). The latest recommendations in the guidelines for the treatment of gBRCA-mutated advanced BC highlight the promise of platinum-based chemotherapies and poly adenosine diphosphate–ribose polymerase inhibitors (PARPi) (National Comprehensive Cancer Network, 2020). Consequently, genetic testing becomes more and more important to identify patients with gBRCA-mutant tumors.

Although the methodology of detecting genetic variants has greatly improved, molecular testing is usually time-consuming and could be limited by availability of adequate samples. Moreover, the cost of genetic testing is still too high for most families. Therefore, BRCA detection has traditionally been limited to BC patients who have an *a priori* high risk of being a mutation carrier. These risk factors include triple-negative BC, young age at diagnosis (below 45 years), or a family history of breast and/or ovarian cancer (Wong-Brown et al., 2015; Grindedal et al., 2017). Although many guidelines in various countries focus on identifying such high-risk groups, the latest guidelines adopt broader criteria regardless of family history. This supports the increasing evidence in the literature that clinical criteria (e.g., family history) may omit individuals with BRCA1/2 mutations, some of which suggest that BRCA testing should be expanded to a wider population. Thus, the method to predict gene mutations quickly and inexpensively from histopathology images could be beneficial to the treatment of patients with BC given the importance and impact of these mutations.

The latest development in artificial intelligence (AI) provided a novel method to assist clinicians to classify medical information and images (Bera et al., 2019; Bi et al., 2019). The possibility of digitizing whole-slide images (WSIs) of pathology tissue has led to the emergence of AI and machine learning (ML) tools in digital pathology, which can mine the subvisual morphometric phenotypes and ultimately enhance patient management. Recently, pathologists and computer scientists have come together to apply the latest AI technology (e.g., deep learning) to the problem of analyzing pathology slides for assisting diagnosis, prediction, prognosis, and other clinically related purposes, as well as other applications such as improving the efficiency of the diagnostic workflow. In breast pathology, deep learning (DL) has already been applied in classifying the type and subtype of breast tumors, identifying metastasis in lymph nodes, detecting tubular formation and nuclear pleomorphism, tumor grading, counting mitotic figures, etc.

(Bejnordi et al., 2017; Sudharshan et al., 2019; Mahmood et al., 2020; Xu et al., 2020). Furthermore, researchers investigated whether the molecular characteristics of cancer are encoded in histomorphological structures that are beyond human apprehension (Xu et al., 2019; Schmauch et al., 2020; Bilal et al., 2021). As such, Shamai et al. (2019) applied an ML method, termed morphological-based molecular profiling (MBMP), on BC specimens to explore the associations between histomorphological characteristics and expression of multiple molecular biomarkers. For at least half of the patients in this study, MBMP seemed to predict the expression of biomarkers and is not inferior to immunohistochemistry (Shamai et al., 2019). Similarly, Narula et al. (2018) trained a deep convolutional neural image processing network to automatically classify histopathological subtypes from digital pathology slides of lung specimens and predict common mutant genes in lung adenocarcinoma.

These results suggest that DL models can be used to effectively assist pathologists in detecting gene mutations and tumor histological subtypes. However, it remains unclear whether DL can be applied to predict BRCA gene mutation status using BC digital pathology slides. Therefore, we focused on the BC specimens and tested whether DL can be trained to predict gBRCA1/2 mutations using images as the only input. In this study, we constructed DL models based on convolutional neural networks (CNN) using WSIs of hematoxylin and eosin (H&E)-stained digital pathology slides obtained from the Jiangsu Province Hospital of Chinese Medicine (JSPHCM) and Jiangsu Cancer Hospital (JSCH) to predict the gBRCA1/2 mutation status in BC.

MATERIALS AND METHODS

Study Cohort

All the cases were collected from two medical centers in China, which were Jiangsu Province Hospital of Chinese Medicine (JSPHCM) and Jiangsu Cancer Hospital (JSCH), Nanjing. A total of 22 BC patients were eventually enrolled in the BRCA-mutation group, and 40 patients were enrolled in the BRCA-wild group. We combined H&E-stained WSIs from two datasets: the JSPHCM dataset, which contains 60 H&E images from 12 BRCA-mutation patients and 50 H&E images from 10 BRCA-wild patients, and the JSCH dataset, which contains 25 H&E images from 10 BRCA-mutation patients and 87 H&E images from 30 BRCA-wild patients. Slides were digitized with a NanoZoomer Digital slide scanner (Hamamatsu Photonics Scientific Instrument Co., Ltd., Beijing, China) at a resolution of ×40. This study has been approved by the Institutional Ethical Review Boards of JSPHCM with patient consent.

The tumor pathology for all patients with BC was reviewed under the criteria of the World Health Organization Classification of Tumors: Breast Tumors (5th edition) (WHO, 2019) by one of our designated breast pathologists. All 22 patients with BRCA mutation have invasive breast carcinoma, not otherwise specified (invasive ductal carcinoma). Among the 40 patients with BRCA wild type, 36 cases were invasive ductal carcinoma, two cases were mucinous carcinoma, one case was invasive lobular carcinoma, and one case was a metaplastic carcinoma. Using Automated Slide Analysis Platform (ASAP 1.9), pathologists can navigate WSI images at a very high resolution and annotate the whole-tumor regions within slides for ease of adjudication. The DL model was trained based on the annotated tumor location.

Pathologists classified all the invasive BCs according to the Nottingham histological grading system (NGS). The NGS has three parameters, which are tubule formation, nuclear pleomorphism, and mitotic count. Each parameter has been divided into three categories, with the score from 1 to 3, assigned as follows: tubule formation (1: 75%, 2: 10–75%, 3: 10%); nuclear pleomorphism (1: none, 2: moderate, 3: pronounced); and the number of mitoses/10 high-power fields (HPF) (40 objective lens) (1: 0–9 mitoses; 2: 10–19 mitoses; and 3: > 19 mitoses). The final histological grade is based on a sum of the scores of the three parameters: 3, 4, or 5 = grade 1; 6 or 7 = grade 2; and 8 or 9 = grade 3 (WHO, 2019). In the cohort of 62 patients, grade 1 tumors have been observed in 1 patient with BRCA-wild (1/40, 2.5%), and none has been observed in patients with BRCA-mutation (0/22, 0%); grade 2 tumors have been observed in 14 patients with BRCA-wild (14/40, 35%) and 3 patients with BRCA-mutation (3/22, 13.6%); and grade 3 tumors have been observed in 25 patients with BRCA-wild (25/40, 62.5%) and 19 patients with BRCA-mutation (19/22, 86.4%).

Data on BRCA1/2 mutations were routinely collected and extracted in clinic from electronic medical records. BRCA testing has been done in a centralized clinical testing center (Nanjing Geneseeq Technology Inc., Nanjing, China), using germline DNA (from blood), according to protocols reviewed and approved by the ethical committee of each participating hospital, and the test results were categorized as either positive or negative of a deleterious mutation.

Method of DL With Convolution Neural Networks

The DL model we used in this study is a residual neural network (ResNet), which is a type of artificial neural network that builds based on pyramid cells in the cerebral cortex. The typical ResNet is built by having layer-skipping connections to avoid the problem of gradient vanishing. Thus, it allows to train on a deeper neural network (He et al., 2016). In this case, the network is ideal to be used to classify complex histomorphological structures. A ResNet with 18 layers has been used (**Figure 1**). At the end of the network, a fully connected layer is added for binary classification between BRCA-wild and BRCA-mutation. The model has been trained by using a dual GPU setup with 2 × 1,080 ti graphics card from Nvidia. The stochastic gradient descent method based on adaptive estimation of first-order and second-order moments has been used as the loss function in the training (Kingma and Ba, 2015).

From the JSPHCM dataset reviewed by our designated breast pathologists, 58 H&E images of BRCA-mutation and 44 H&E images of BRCA-wild have been selected for the training, among which 56 H&E images of BRCA-mutation, and 33 H&E images

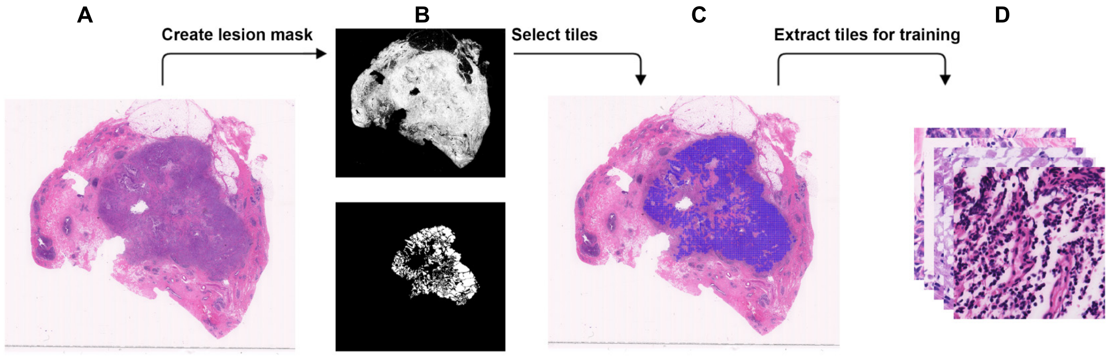

FIGURE 1 | Steps to extract small tiles from whole-slide images at ×40 magnification for training. **(A)** An example WSI with BRCA-mutation or BRCA-wild. **(B)** The binary mask (top) was created using color differences in RGB color space, and then element-wise matrix multiplication was done with the manually labeled lesion (bottom) to create a lesion mask without void spaces. **(C)** Tiles with resolution of 384 × 384 were extracted from the lesion mask. **(D)** The tiles that were extracted will be used in the training.

of BRCA-wild are categorized as histological grade 3. From the JSCH dataset, 17 H&E images of BRCA-mutation and 47 H&E images of BRCA-wild have been selected for external validation, among which 17 H&E images of BRCA-mutation and 27 H&E images of BRCA-wild are categorized as histological grade 3, In all the selected images, a brief location of the tumor is annotated and will be used as labels for supervised training. Training and internal testing datasets were created from JSPHCM dataset, and the external testing was created from the JSCH dataset.

Training Data Preparation

WSI has a large resolution which sometimes has a resolution larger than 50,000 × 50,000 pixels. It is not possible to process the entire image for DL due to the memory usage. Therefore, we chose to break down each image into tiles with a smaller resolution (Dimitriou et al., 2019). Using brief annotation of the tumor annotated by designated breast pathologists, tiles with tumor tissue were extracted using the labeled data. To avoid extracting tiles from the void area, a binary mask for cellular tissue was created by using the color spacing in the RGB space. An element-wise matrix multiplication between the binary mask and the labeled tumor area was performed to extract the tumor mask from the binary mask without the void area. The tumor mask was then divided into tiles. For each tile, at least 30% of the area is covered with tissue to make sure no void area is used in the DL computation. A detailed illustration is found in **Figure 2**. The input size of the DL model is 256 × 256, but during data preparation, we extracted tiles at a resolution of 384 × 384 to create enough resolution space for augmentation during training. To maintain an even number of tiles from each slide, the number of tiles to be extracted from each slide was determined by the minimum number of the tiles that could be extracted among all the slides.

WSI can be inspected at different magnifications. The pixel information varies at different magnifications within a fixed pixel area. Morphological structures of the cellular tissues were preserved at low magnification while better details of cellular structure were preserved at high magnification. Different morphological structures might contain different features that could contribute differently to the DL algorithm. In our study, we used four types of magnifications ranging between ×5, ×10, ×20, and ×40 to find the optimal range of magnification to achieve the best prediction. From 102 slides from the JSPHCM dataset, a total of 18,109 tiles were extracted with 10,140 BRCA-mutation tiles and 7,969 BRCA-wild tiles at ×5 magnification. At ×10 magnification, a total of 58,745 tiles were extracted with 32,344 BRCA-mutation tiles and 26,401 BRCA-wild tiles. At ×20 magnification, a total of 239,108 tiles were extracted with 131,467 BRCA-mutation tiles and 107,641 BRCA-wild tiles. At ×40 magnification, a total of 962,868 tiles were extracted, with 529,242 BRCA-mutation tiles and 433,626 BRCA-wild tiles. All the tiles were randomly and equally divided into 90 and 10% for the training and internal testing datasets, respectively, to be used to train the model. All the tiles with histological grade 3 were labeled.

Data Augmentation and Training

Due to limitations among the slides available, to prevent overfitting of the model, each tile in the training dataset undergoes multiple steps of augmentation before feeding into the model (**Figure 3**). Tiles at a resolution of 256 × 256 were extracted randomly from tiles created during data preparation that has a 384 × 384 resolution. The extracted tiles later went through random flips in left/right and up/down orientation to increase data complexity. Since each slide has variability in artifacts and staining, and H&E staining has high pixel intensity in the red and blue channels with RGB color space, a random intensity adjustment at both red and blue channels was also done in a range of –20 to + 20-pixel value in the 8-bit color channel to synthetically capture the variabilities. In the end, an overall brightness change was added in a range of –20 to + 20 to the image. Through these extensive augmentations, we tried to increase the data complexity to let the model focus on cellular morphology or other "unknown" features to differentiate BRCA-mutation and BRCA-wild, without being influenced by the color saturation or brightness, which is different case by case due

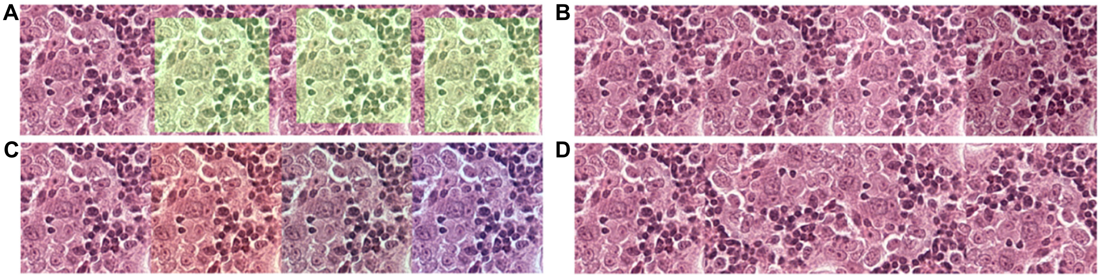

FIGURE 2 | Data augmentation to normalize the training data and simulate and prevent overfitting during training. **(A)** 256 × 256 tiles were randomly extracted from the 320 × 320 tiles. **(B)** The brightness of the tiles was randomly adjusted in a range of –20 to + 20-pixel value in an 8-bit space. **(C)** The red and blue channels of the tiles were randomly adjusted in a range of –20 to + 20-pixel value in an 8-bit space. **(D)** Tiles were randomly flipped up/down and left/right.

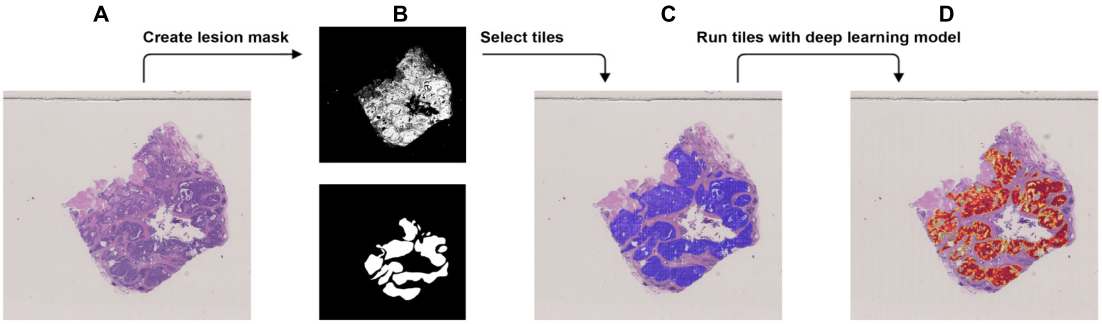

FIGURE 3 | Procedure to validate the trained model using the external dataset with BRCA-mutation. **(A)** The WSI with BRCA-mutation. **(B)** The binary mask (top) was created using color differences in the RGB color space and then element-wise matrix multiplication was done with the manually labeled lesion (bottom) to create a lesion mask without void spaces. **(C)** Tiles with a resolution of 256 × 256 were extracted from the lesion mask. **(D)** Heatmap which illustrates the probability of the region being classified as BRCA-mutation.

to scanning, staining method, etc. Models were trained for extensiveness among iterations until the minimum loss function gradient of the training validation dataset is reached. The same steps were repeated for tiles with histological grade 3.

External Validation

We used the JSCH dataset for external validation since it was not involved in any training. It served as a good validation dataset to evaluate the performance and robustness of our trained DL model. During external validation, each image was broken down into 256 × 256 tiles from the binary mask and the labeled tumor mask, which was the same as the data preparation for training using the range of ×5, ×10, ×20, and ×40 magnifications. The extracted tiles were fed into the model as the input and model output classification probability of BRCA-mutation. The outputs were illustrated as heatmap images; the higher the probability of BRCA-mutation, the higher the heatmap intensity (**Figures 4**, **5**). The average probability across all the slides was calculated from the probability of the tiles. Anything higher than the 0.5 probability was considered as BRCA-mutation. The receiver operating characteristic curve (ROC), area under the ROC (AUC), and confusion matrix were created for the testing results (**Figures 6**, **7** and **Table 1**). The same steps were repeated for WSI with histological grade 3.

Statistical Analysis

We performed the ROC and calculated the validity (true positive rate, false negative rate, false positive rate, true negative rate, likelihood ratio) and predictive value to demonstrate the classification ability of the DL model. Delong tests were then applied to compare the AUC of slides and tiles with different magnifications from all cases and grade 3 cases. A percentage bar plot was plotted to visualize the validity (true positive rate, false negative rate, true negative rate, and false positive rate) of the DL model. Box plot and Student's *t*-test were used to compare the predictive BRCA mutation probability of mutation and wild group by the DL model. A Bland–Altman plot was plotted to evaluate the agreement of predicted mutation probability for per-slide at different magnifications. All statistical analyses and figures were performed by using R version 4.0.3 (The R Foundation for Statistical Computing; Vienna, Austria) with packages "ggplot2" and "ggthemes." A *p*-value of less than 0.05 was considered as statistical significance.

RESULTS

In the external validation dataset, AUCs (95% CI) of DL models using ×40, ×20, ×10, and ×5 magnification tiles among all cases were 0.766 (0.763–0.769), 0.763 (0.758–0.769), 0.750

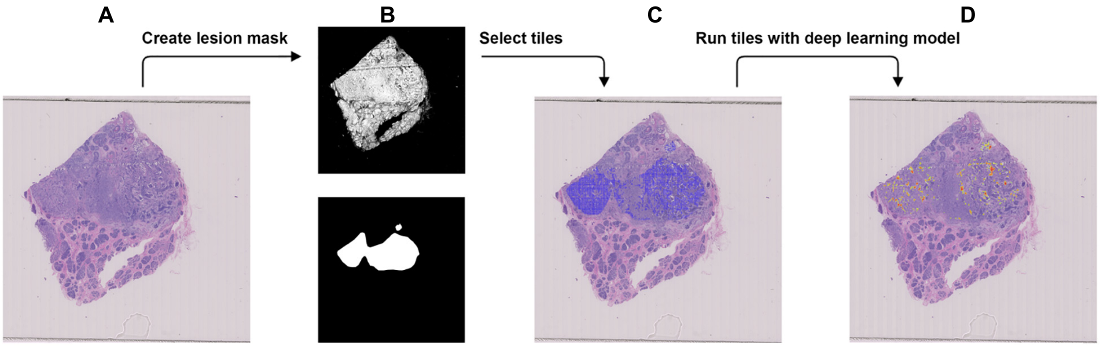

FIGURE 4 | Procedure to validate the trained model using the external dataset with BRCA-wild. **(A)** The WSI with BRCA-wild. **(B)** The binary mask (top) was created using color differences in the RGB color space and then element-wise matrix multiplication was done with the manually labeled lesion (bottom) to create a lesion mask without void spaces. **(C)** Tiles with a resolution of 256 × 256 were extracted from the lesion mask. **(D)** Heatmap which illustrates the probability of the region being classified as BRCA-mutation.

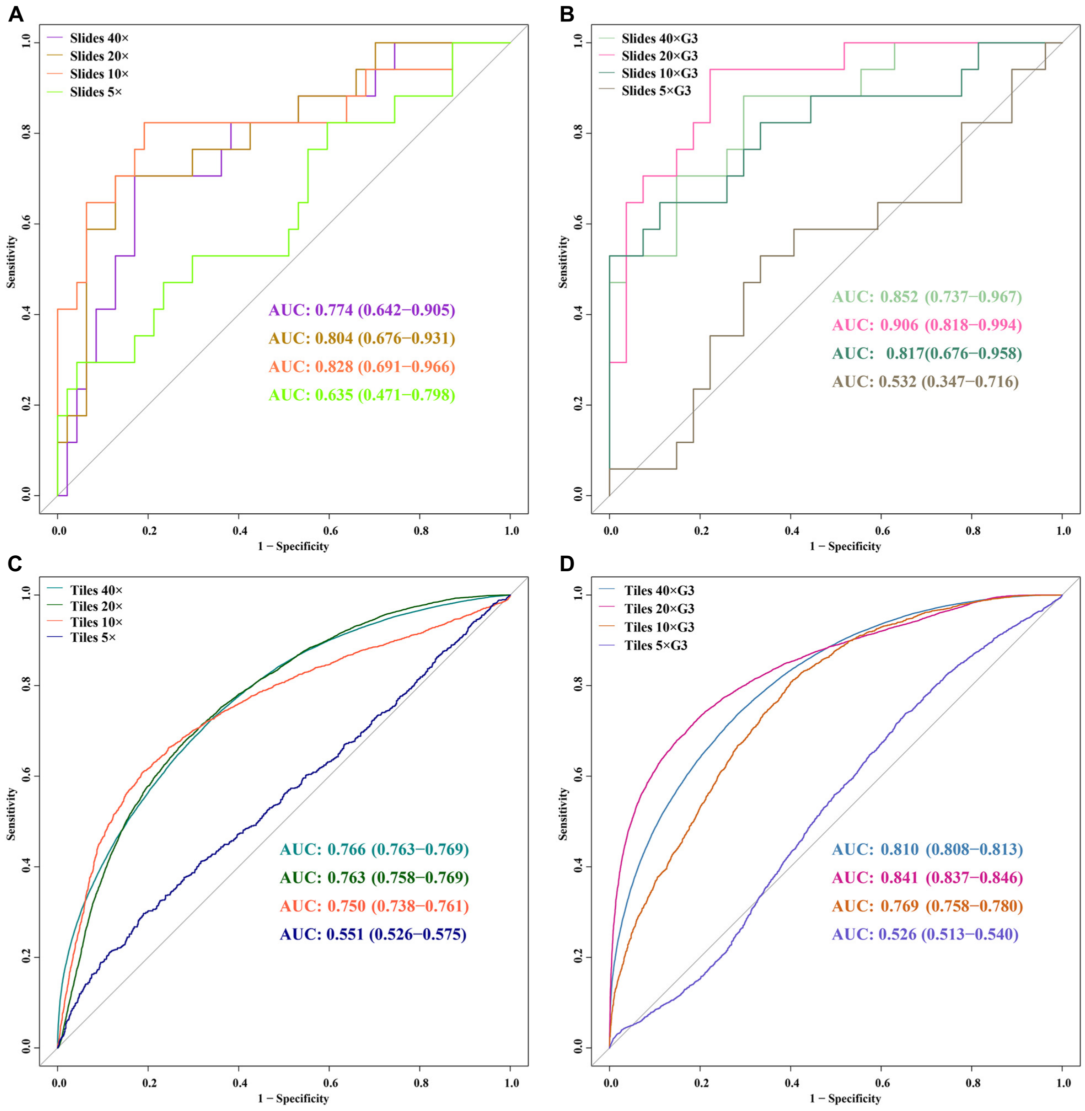

FIGURE 5 | ROC curves of 5 DL models for slides and tiles at ×40, ×20, ×10, and ×5 magnification. **(A)** slides, **(B)** slides G3, **(C)** tiles, **(D)** tiles G3.

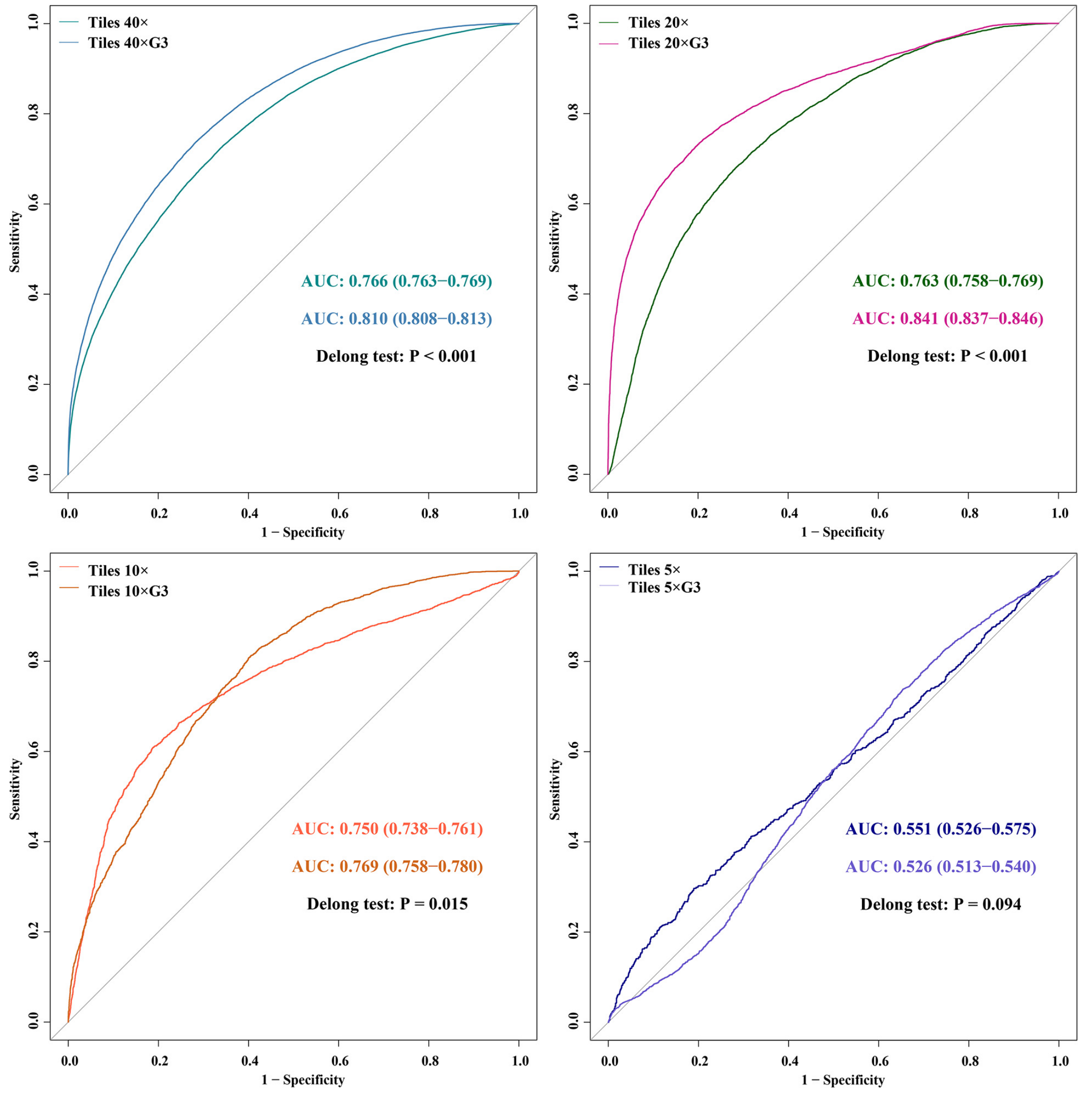

FIGURE 6 | The ROC and comparison of AUCs of DL models using tiles between all cases and G3 cases.

(0.738–0.761), and 0.551 (0.526–0.575), respectively; those using corresponding magnification tiles among grade 3 cases were 0.810 (0.808–0.813), 0.841 (0.837–0.846), 0.769 (0.758–0.780), and 0.526 (0.513–0.540), respectively, those using corresponding magnification slides among all cases were 0.774 (0.642–0.905), 0.804 (0.676–0.931), 0.828 (0.691–0.966), and 0.635 (0.471–0.798), respectively, and those using corresponding magnification slides among grade 3 cases were 0.852 (0.737–0.967), 0.906 (0.818–0.994), 0.817 (0.676–0.958), and 0.532 (0.347–0.716), respectively (**Figure 5**). Delong test demonstrated that AUCs (95% CI) of DL models using ×40 ($P < 0.001$), ×20 ($P < 0.001$), and ×10 ($P < 0.001$) magnification tiles among all cases were less than those among grade 3 cases, and that using ×5 magnification tiles among all cases and grade 3 cases was marginally significant (**Figure 6**). Additionally, the ROC and the comparison of AUCs among another magnification slides and tiles are listed in **Supplementary Figure 1**.

The validity and predictive value of DL models using different magnification tiles or slides are listed in **Figure 8** and **Table 1**. The positive likelihood ratios (+ LR) of DL models using ×10, ×20, and ×40 magnification slides and tiles among all cases and grade 3 cases were high, and the negative likelihood ratios (-LR) were low. Meanwhile, the almost negative predictive values of those were more than 0.700, except ×40 magnification tiles among grade 3 cases (0.681). The above results suggest that the validity of these models was high, and a higher proportion of patients with negative diagnoses (wild rather than mutation) were actually negative. Slides at ×10 magnification that had the best performance suggest that a bigger field of view contributes positively to the classification between BRCA-mutation and BRCA-wild. It corresponds to the features for the prediction of BRCA1 and BRCA2 mutation, such as the presence of lymphocytic infiltrate, pushing margin, and lack of tubule formation, which are mostly shown in ×10 slides

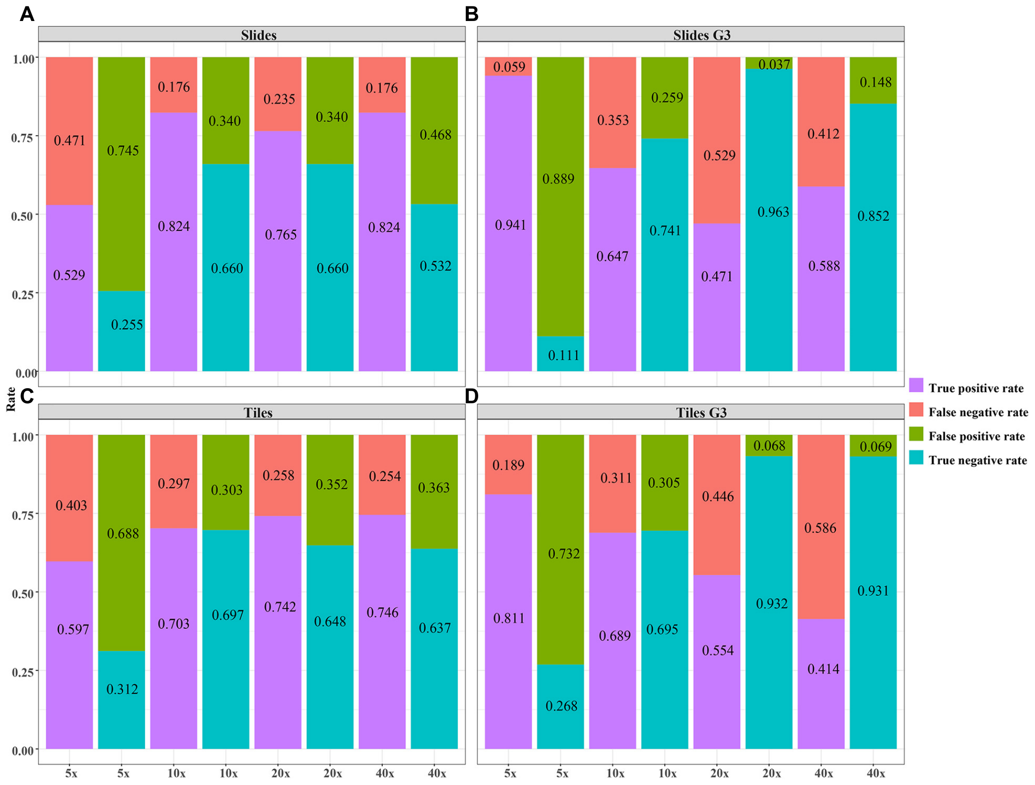

FIGURE 7 | The validity of the DL model using different magnification tiles and slides. **(A)** slides, **(B)** slides G3, **(C)** tiles, **(D)** tiles G3.

rather than ×20 and ×40 slides. Moreover, slides at ×20 magnification had the best performance among grade 3 cases, which suggests that a higher histological grade of BC has more complex and tiny features. Therefore, an appropriate higher magnification enables the DL model to have better ability of gBRCA-mutation classification.

In order to assess the classification accuracy on per-slide level and per-tile level, the results were aggregated using ×40, ×20, ×10, and ×5 magnifications. The ×10, ×20, and ×40 magnifications that were applied to the mutation groups show significantly higher probabilities than the wild group in both slide level and tile level (**Figure 8**). Next, according to the results of the box plot and the comparison, we plotted the Bland–Altman plot using the predictive probability obtained from the three magnifications (×10, ×20, and ×40) to investigate the agreement of per-slide on classification among all cases and grade 3 cases. **Figures 9-1, 9-2** show that most of the points were distributed within the range of 95% limits of agreement (LoA). In other words, all results were in a good agreement.

DISCUSSION

We have developed a computerized system (CNN-based DL) to predict molecular markers (gBRCA mutation) of BC by analysis of tumor histomorphology. Since BRCA1/2 gene mutations occur at a relatively high frequency in BC, to predispose an individual with developing BC and other cancers, PARP inhibitors are regarded as one of the potential targeted drugs for gBRCA mutant BC. Although it has been suggested that gene mutations could

TABLE 1 | The likelihood ratio (+LR and −LR) and predictive value (PPV and NPV) of DL models.

Classification	+LR	−LR	PPV	NPV
Slides × 40	1.761	0.331	0.389	0.893
Slides × 20	2.250	0.356	0.448	0.886
Slides × 10	2.424	0.267	0.467	0.912
Slides × 5	0.710	1.847	0.205	0.6
Slides × 40 G3	3.973	0.484	0.714	0.767
Slides × 20 G3	12.730	0.549	0.889	0.742
Slides × 10 G3	2.498	0.476	0.611	0.769
Slides × 5 G3	1.058	0.532	0.400	0.750
Tiles × 40	2.055	0.399	0.567	0.797
Tiles × 20	2.108	0.398	0.571	0.799
Tiles × 10	2.320	0.426	0.585	0.794
Tiles × 5	0.868	1.291	0.339	0.567
Tiles × 40 G3	6.000	0.629	0.817	0.681
Tiles × 20 G3	8.147	0.478	0.859	0.736
Tiles × 10 G3	2.259	0.447	0.624	0.752
Tiles × 5 G3	1.108	0.705	0.449	0.658

+LR, positive likelihood ratio; −LR, negative likelihood ratio; PPV, positive predictive value; NPV, negative predictive value; G3, grade 3.

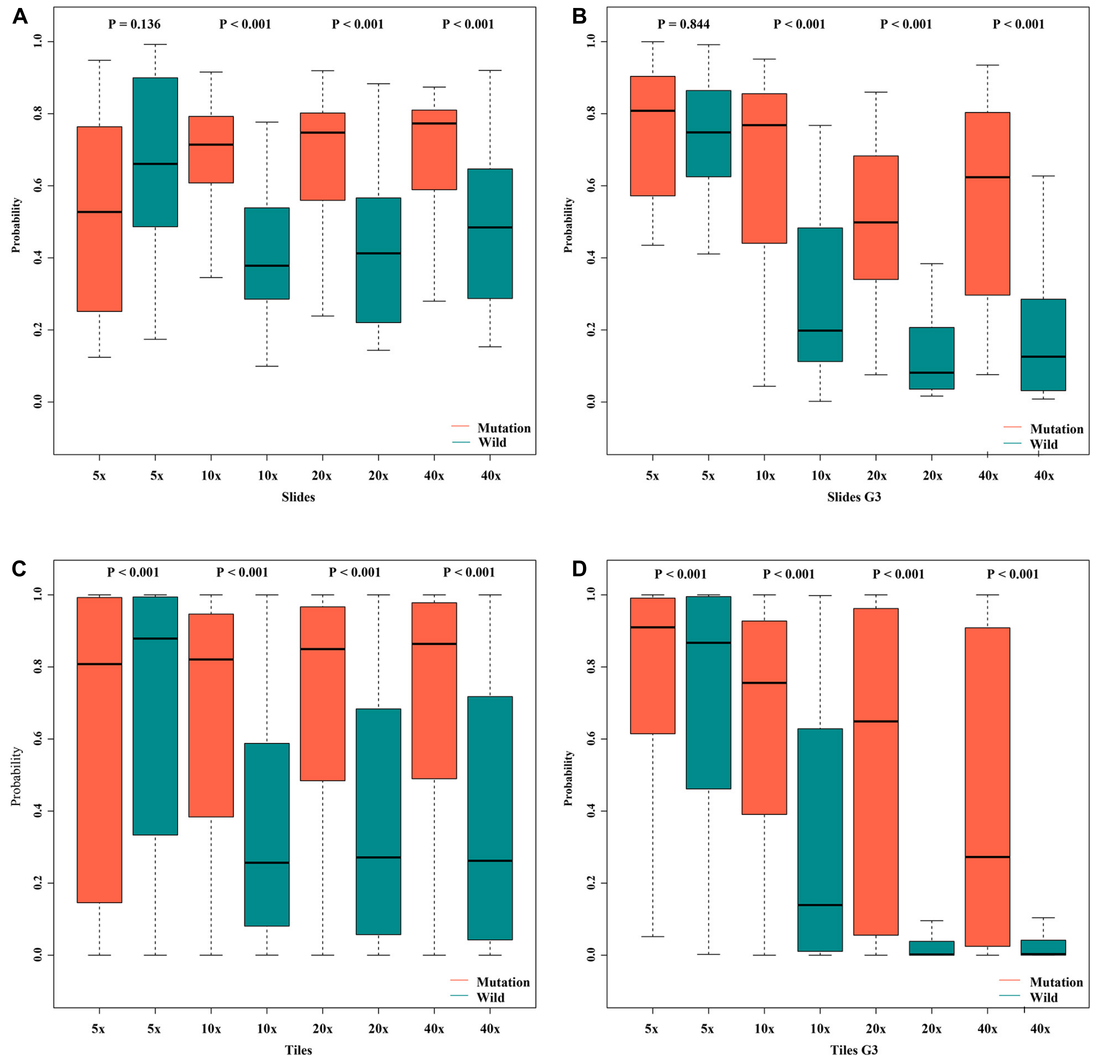

FIGURE 8 | Box plot of probabilities on slide and tile level with different magnifications. **(A)** slides, **(B)** slides G3, **(C)** tiles, **(D)** tiles G3.

be predicted from H&E-stained WSIs [AUC of ∼0.85 for the prediction of STK11 mutations from lung cancer H&E images (Narula et al., 2018); AUC of ∼0.71 for the prediction of SPOP mutations from prostate cancer H&E images (Schaumberg et al., 2018); AUC of 0.71∼0.89 for the prediction of CTNNB1, FMN2, TP53, and ZFX4 mutations from hepatocellular carcinoma H&E images (Chen et al., 2020)], prior to this study, it was unclear whether gBRCA mutations would affect the pattern of tumor cells on BC H&E-stained WSIs.

We have trained the DL model (ResNet) using the presence or absence of BRCA mutation as a label revealed that gBRCA mutational status can be predicted from image data alone (AUCs 0.55–0.91 in different DL models). Interestingly, BC cases with high-grade histology (Grade 3) achieved higher AUC (95% CI) using the DL model at ×40, ×20, and ×10 magnifications. It corresponds to that gBRCA-mutated BCs are more frequently of higher histological grade. Moreover, slides at ×20 magnification had the best performance among grade 3 cases (AUC up to 0.906), and slides at ×10 magnification had the highest negative predictive value (0.91), which suggest that histopathological images with different magnifications can represent different information. The images with low magnification (×5–×10) cover a larger field of view, while the images with high magnification (×20–×40) correspond to a relatively small area with more details. In the analysis of histopathological images, it is necessary to recognize complex morphological patterns of various sizes. AI can capture cellular level information by high-magnification images and tissue spatial structures by low-magnification images at the same time. We found that the tumor morphology captured in H&E-stained images contains signals that predict the status of tumor molecular markers. DL approaches can extract sub-visual morphological phenotypes from WSIs beyond that which a human is capable. We show that DL can recognize a group with morphological features within the tissue structure captured from WSIs and predict the gBRCA1/2 mutation status. The prediction of gBRCA mutation using DL will be of great significance for

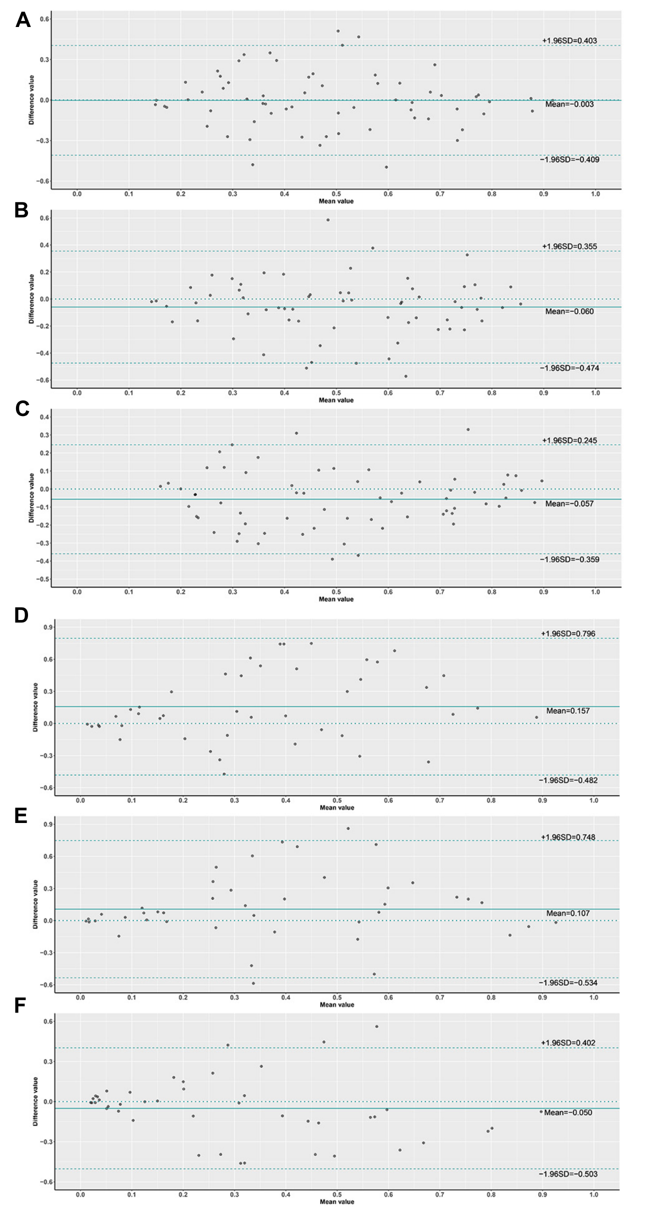

FIGURE 9 | Bland–Altman plot of probabilities on slide level. **(1) (A)** ×10 vs. ×20 slides, **(B)** ×10 vs. ×40 slides, **(C)** ×20 vs. ×40 slides, **(2) (D)** ×10 vs. ×20 slides G3, **(E)** ×10 vs. ×40 slides G3, **(F)** ×20 vs. ×40 slides G3.

select patients who are most likely to respond to PARP inhibitor-targeted therapy and identification of healthy mutation carriers within their families.

The model used in this study is ResNet with 18 layers. It is a simplified ResNet with a smaller number of layers, compared with the original ResNet (He et al., 2016) proposed by Kaiming He. The advantage of ResNet is that it is possible to go deeper without losing generalization capability. It is making sense that the deeper the network, the better the result for the convolutional network. However, we need to simplify the network to avoid overfitting due to the size of the dataset. The depth of the model can be adjusted as the size of datasets grows. Both models trained at ×10 and ×40 magnification show effective results. Each model operates at a different level of magnification, which the user could choose to use a single model for efficiency or a combined model for higher accuracy. With the updates in the computer systems and hardware, the resolution of the tiles and the number of models at multiple magnifications could also be improved to exploit for better accuracy.

This is the first study to predict the BRCA gene mutation in BC, while using an independent database from JSCH to externally validate the performance of the model. It has been proved that CNN-based DL can be used to assist gene mutation prediction based on histopathological slides in BC. However, the present study has several limitations. One limiting factor in achieving higher accuracy lies in the small number of slides containing BRCA mutation instances that can be used for training and validation. Furthermore, the ability of any such AI approaches to predict all targetable mutations is critical, as more and more molecular markers are expected to be quantified in each sample, and treatment decisions are usually delayed until information about all such driver mutations is obtained. Subsequently, further validation of our model is necessary in a larger dataset with multiple centers and other BC-related genes should be considered in further study.

In conclusion, the study demonstrates that CNN-based DL can predict the gBRCA mutation status from H&E-stained WSIs in BC, and DL potential to improve cancer prognosis and therapeutics by utilizing biological markers currently imperceptible to clinicians. Although AI cannot completely replace humans in practice nowadays, gene mutation prediction can be used as a prescreening to improve the cost efficiency before next-generation sequencing, thereby improving the performance of precision medicine.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee in Affiliated Hospital of Nanjing University of Chinese Medicine (Jiangsu Provincial Hospital of Chinese Medicine). The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

LX and YFZ: conception and design. XW, YZ, FK, JC, XX, and YFZ: provision of study materials or patients. XW, CZ, XL, CW, and LX: collection and assembly of data. XW, WW, and DW: data analysis and interpretation. All authors: writing and final approval of the manuscript.

FUNDING

This work was supported by the Six Talent Project of Jiangsu Province, China (Grant No. WSW-014), the Priority Academic Program Development of Jiangsu Higher Education Institutions (Grant No. PAPD2018-87), and the Innovative Development Foundation of Jiangsu Province Hospital of CM (Grant Nos. Y2019CX19 and Y2020CX43).

ACKNOWLEDGMENTS

We would like to express our appreciation and gratitude to the following persons: Fan Yang, Statistics and Mathematics Major, University of Toronto, St. George Campus, for the operation in statistical analysis and machine learning; and Ning Xuan, Radlink Inc. Senior Biomedical Software Engineer, for the operation in computer vision and deep learning neural network.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.661109/full#supplementary-material

Supplementary Figure 1 | The ROC and the comparison of AUCs among another magnification at tiles.

REFERENCES

Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., et al. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* 72, 1117–1130.

Atchley, D. P., Albarracin, C. T., Lopez, A., Valero, V., Amos, C. I., Gonzalez-Angulo, A. M., et al. (2008). Clinical and pathologic characteristics of patients with BRCA-positive and BRCA-negative breast cancer. *J. Clin. Oncol.* 26, 4282–4288. doi: 10.1200/jco.2008.16.6231

Bejnordi, B. E., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318, 2199–2210.

Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V., and Madabhushi, A. (2019). Artificial intelligence in digital pathology–new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* 16, 703–715. doi: 10.1038/s41571-019-0252-y

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., et al. (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. *CA. Cancer J. Clin.* 69, 127–157.

Bilal, M., Raza, S. E. A., Azam, A., Graham, S., Ilyas, M., Cree, I. A., et al. (2021). Novel deep learning algorithm predicts the status of molecular pathways and key mutations in colorectal cancer from routine histology images. *medRxiv* [Preprint]. doi: 10.1101/2021.01.19.21250122

Chen, M., Zhang, B., Topatana, W., Cao, J., Zhu, H., Juengpanich, S., et al. (2020). Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis. Oncol.* 4:14. doi: 10.1038/s41698-020-0120-3

Chen, S., and Parmigiani, G. (2007). Meta-analysis of BRCA1 and BRCA2 penetrance. *J. Clin. Oncol.* 25, 1329–1333.

Dimitriou, N., Arandjelović, O., and Caie, P. D. (2019). Deep learning for whole slide image analysis: an overview. *Front. Med.* 6:264. doi: 10.3389/fmed.2019.00264

Domchek, S. M., Friebel, T. M., Singer, C. F., Evans, D. G., Lynch, H. T., Isaacs, C., et al. (2010). Association of risk-reducing surgery in BRCA1 or BRCA2 mutation carriers with cancer risk and mortality. *JAMA* 304, 967–975. doi: 10.1001/jama.2010.1237

Engel, C., and Fischer, C. (2015). Breast cancer risks and risk prediction models. *Breast Care (Basel)* 10, 7–12. doi: 10.1159/000376600

Faraoni, I., and Graziani, G. (2018). Role of BRCA mutations in cancer treatment with poly (ADP-ribose) polymerase (PARP) inhibitors. *Cancers* 10:487. doi: 10.3390/cancers10120487

Grindedal, E. M., Heramb, C., Karsrud, I., Ariansen, S. L., Mæhle, L., Undlien, D. E., et al. (2017). Current guidelines for BRCA testing of breast cancer patients are insufficient to detect all mutation carriers. *BMC Cancer* 17:438. doi: 10.1186/s12885-017-3422-2

He, K., Zhang, X., Ren, S. and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 770-778. doi: 10.1109/CVPR.2016.90

Kingma, D., and Ba, J. (2015). Adam: a method for stochastic optimization. *Paper Presented at the 3rd International Conference for Learning Representations, Computer Science, Mathematics*, (San Diego, CA: CoRR).

Lang, G. T., Shi, J. X., Hu, X., Zhang, C. H., Shan, L., Song, C. G., et al. (2017). The spectrum of BRCA mutations and characteristics of BRCA-associated breast cancers in China: screening of 2,991 patients and 1,043 controls by next-generation sequencing. *Int. J. Cancer* 141, 129–142. doi: 10.1002/ijc.30692

Mahmood, T., Arsalan, M., Owais, M., Lee, M. B., and Park, K. R. (2020). Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. *J. Clin. Med.* 9:749. doi: 10.3390/jcm9030749

Metcalfe, K., Gershman, S., Ghadirian, P., Lynch, H. T., Snyder, C., Tung, N., et al. (2014). Contralateral mastectomy and survival after breast cancer in carriers of BRCA1 and BRCA2 mutations: retrospective analysis. *BMJ* 348:g226. doi: 10.1136/bmj.g226

Narula, N., Moreira, A. L., Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Fenyö, D., et al. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24, 1559–1567. doi: 10.1038/s41591-018-0177-5

National Comprehensive Cancer Network (2020). *NCCN Clinical Practice Guidelines in Oncology: Breast Cancer, Version 4*. Available online at: https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf.

Paul, A., and Paul, S. (2014). The breast cancer susceptibility genes (BRCA) in breast and ovarian cancers. *Front. Biosci.* 19, 605–618. doi: 10.2741/4230

Schaumberg, A. J., Rubin, M. A., and Fuchs, T. J. (2018). H&E-stained whole slide deep learning predicts SPOP mutation state in prostate cancer. *bioRxiv* [Preprint]. doi: 10.1101/064279

Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., et al. (2020). A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.* 11:3877. doi: 10.1038/s41467-020-17678-4

Shamai, G., Binenbaum, Y., Slossberg, R., Duek, I., Gil, Z., Kimmel, R., et al. (2019). Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. *JAMA Netw. Open* 2:e197700. doi: 10.1001/jamanetworkopen.2019.7700

Sønderstrup, I. M. H., Jensen, M. B., Ejlertsen, B., Eriksen, J. O., Gerdes, A. M., Kruse, T. A., et al. (2018). Subtypes in BRCA mutated breast cancer. *Hum. Pathol.* 84, 192–201.

Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L., Heutte, L., and Honeine, P. (2019). Multiple instances learning for histopathological breast cancer image classification. *Expert Syst.* 117, 103–111. doi: 10.1016/j.eswa.2018.09.049

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jema, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/caac.21660

Torrisia, R., Zuradellia, M., Agostinettoa, E., Masci, G., Losurdo, A., De Sanctis, R., et al. (2019). Platinum salts in the treatment of BRCA-associated breast cancer: a true targeted chemotherapy? *Crit. Rev. Oncol. Hematol.* 135, 66–75. doi: 10.1016/j.critrevonc.2019.01.016

WHO (2019). *Classification of Tumours Editorial Board. World Health Organisation Classification of Tumours: Breast Tumours*, 5th Edn. Lyon: International Agency for Research on Cancer (IARC).

Wong-Brown, M. W., Meldrum, C. J., Carpenter, J. E., Clarke, C. L., Narod, S. A., Jakubowska, A., et al. (2015). Prevalence of BRCA1 and BRCA2 germline mutations in patients with triple-negative breast cancer. *Breast Cancer Res. Treat.* 150, 71–80. doi: 10.1007/s10549-015-3293-7

Xu, H., Park, S., Lee, S. H., and Hwang, T. H. (2019). Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. *bioRxiv* [Preprint]. doi: 10.1101/554527

Xu, Z., Verma, A., Naveed, U., Bakhoum, S., Khosravi, P., and Elemento, O. (2020). Using histopathology images to predict chromosomal instability in breast cancer: a deep learning approach. *medRxiv* [Preprint]. doi: 10.1101/2020.09.23.20200139 doi: 10.1201/9780367854737-5

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.

The reviewer RY declared a shared affiliation with the authors YFZ and XW to the handling editor at the time of review.

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